Least squares and maximum likelihood

M.R.Osborne

Contents

Preface

Introduction

Least squares, associated with the names of Gauss and Legendre, and maximum likelihood, developed into a systematic technique by R.A. Fisher, are commonly used tools in the analysis of experimental data in cases where an underlying theory suggests that parametric modelling techniques are appropriate. The material presented here provides information concerning computational procedures available for the estimation of parameters by these techniques given appropriate experimental data . Basic to these applications is an assumed system model

$$\mathbf{F}\left(\mathbf{t}, \mathbf{x}, \boldsymbol{\beta}, \mathbf{u}\right) = 0,$$

where the exogenous variable $\mathbf{t}$ fixes the point at which the corresponding observation is made - it could, for example, be the scalar variable time, $\mathbf{x}$ gives the values of the variables describing the corresponding system state, $\boldsymbol{\beta}$ is a constant vector of parameters which serves to distinguish this system realisation from other members of the class of possible implementations of the underlying model, and $\mathbf{u}$ is the vector of outputs which provide functions of the state information that can be observed. The problem of interest is that of estimating the parameter values $\boldsymbol{\beta}$ and thus identifying the particular realisation under consideration given the results of a given set of n observations $\mathbf{u}\left(\mathbf{t}_i\right), i = 1, 2, \cdots, n$. In most cases considered the model will be explicit in the sense that the observed output is expressed as a function of the input data and the system state. That is $\mathbf{F}$ has the form

$$\mathbf{u}\left(\mathbf{t}\right) = \mathbf{G}\left(\mathbf{t}, \mathbf{x}, \boldsymbol{\beta}\right).$$

In the examples considered here F will be given by a functional expression (linear or nonlinear) of the system variables or will be determined by the numerical integration of a system of differential equations.

The aspect of the estimation problem which provides the particular emphasis in these notes is the further assumption that the experimental data

obtained by making measurements of the system outputs are contaminated by noise, typically additive in form and with known probability distribution. That is the observed quantities at $\mathbf{t} = \mathbf{t}_i$ have the representation

$$\mathbf{y}_i = \mathbf{u}(\mathbf{t}_i) + \boldsymbol{\varepsilon}_i, i = 1, 2 \cdots, n,$$

where the $\boldsymbol{\varepsilon}_i$ are random variables. For our purposes $\boldsymbol{\varepsilon}_i$ will be assumed to have bounded covariance matrix and $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\varepsilon}_j$ will be assumed independent if $i \neq j$. The form of distribution of $\boldsymbol{\varepsilon}_i$ proves to be important in the selection of the estimation procedure. The above assumptions are sufficient for successful application of least squares [52] while there are advantages in efficiency to be gained in maximising the likelihood when the form of distribution is known.

There is an intended method behind this presentation. Chapters 1 and 3 give an overview of the problem settings of the basic least squares and maximum likelihood procedures. The least squares problem in chapter 1 links Gauss-Markov theory, best linear prediction in stochastic models, and filtering and smoothing via the Kalman Filter. Here models are linear, and Hilbert space theory provides an elegant geometric picture. The maximum likelihood methods considered in Chapter 3 are going to lead to nonlinear estimation problems except in the case of linear models and Gaussian errors. Their justification is based on the concept of consistency which comes with the idea that the computed parameter estimates $\boldsymbol{\beta}_n \to \boldsymbol{\beta}^*$ in an appropriate probalistic sense where $\boldsymbol{\beta}_n$ is the estimate from the n'th set of experimental data, where the number of observations per set tends to ∞ as $n \to \infty$, and where $\boldsymbol{\beta}^*$ is the true parameter vector. This requires the following assumptions:

1. the model provides an exact description of the observed process;
2. the nature of the data collected is appropriate; and
3. the manner in which the data is collected can be described asymptotically as the number of observations tends to ∞.

The key tool used in developing the consistency results in this chapter and in Chapter 5 is the Kantorovich development of Newtons method [66].

Both chapters 1 and 3 are followed by chapters which describe appropriate classes of associated numerical algorithms in some detail. However, there is a presumption that the estimation problems lead to dense linear algebra formulations. Chapter 2 treats classic least squares methods such as the Choleski factorization of the normal matrix and orthogonal transformation of the design together with the related surgery to permit the adding and removing of observations in updating procedures. Methods for generalised

least squares include the generalisation of orthogonal techniques provided by V-invariant transformations. Methods are described for both information and covariance settings for the Kalman Filter and include a covariance implementation which provides a square root implementation for both filter and smoother. The key computational result developed in Chapter 4 is that the Gauss-Newton algorithm has seriously advantageous properties including an asymptotically second order rate of convergence as $n \to \infty$, strong transformation invariance, and genuinely powerful global convergence characteristics. All this is based on the framework used in the discussion of consistency which means that it is developed under the indicated assumptions. These refer to an ideal situation which can only be approximated in much modelling work, but it does serve to indicate what could be described as "best practice".

Chapter 5 deals with the estimation of differential equations and presents parallel development of both basic properties and algorithmic aspects for the two main types of estimation procedure – embedding and simultaneous. The numerical integration of the differential equations means that the likelihood, which is required explicitly in the embedding method, can only be evaluated approximately, and this provides a new feature. The integration is performed here courtesy of the trapezoidal rule which has well known stability advantages and proves more than accurate enough given non-trivial data errors. However, because the likelihood is now only evaluated approximately, it becomes necessary to show this approximation doesn't alter significantly the results obtained for exact likelihoods. Maximum likelihood can be used directly with the embedding methods for which Gauss-Newton provides a reliable workhorse, but it appears more indirectly in the simultaneous approach where the necessary conditions involve Lagrange multipliers as a result of the treatment of the differential equation as constraints on the estimation process. The unifying feature is a consequence of the identity of results produced by the two methods. In the simultaneous method there is a possible analog of the Gauss-Newton method in which second derivatives are deleted from the augmented matrix in the quadratic programming steps. This Bock iteration can be shown to be asymptotically second order only if the observational errors are strictly Gaussian, a weaker result than that for Gauss-Newton in the embedding method .

The final chapter discusses certain aspects of nonparametric approximating families developed from linear stochastic differential equations obtained by adding a random walk process with strength λ to a given linear system of differential equations. The best known examples include both splines [109]and g-splines [113]. The Kalman filter subject to a diffuse prior proves to be a suitable tool for fitting these families to observed data and estimating the strength parameter λ. This suggests a possible approach to model

selection from a given hierarchy of differential equation models in which the sequence of strength parameter estimates are compared with zero using an information criterion such as AIK or BIK.

Acknowledgement

Table of notation

Chapter 1

The Linear Least Squares Problem

1.1 Introduction

The linear least squares problem has become the starting point for most introductory discussions of system modelling in which the potential presence of data errors is admitted. It has the generic form

$$\min_{\mathbf{x}} \mathbf{r}^T\mathbf{r};\ \mathbf{r} = A\mathbf{x} - \mathbf{b}. \tag{1.1.1}$$

where the design matrix $A : R^p \to R^n$, the residual and observation vectors $\mathbf{r}$, $\mathbf{b} \in R^n$, and the vector of model parameters $\mathbf{x} \in R^p$. This is a simple optimization problem subject to equality constraints, and it will be convenient to refer to these constraints as the model equations . Typically p will be fixed corresponding to an assumed known model or at least a priori bounded in variable selection calculations, while n which controls the amount of data available will usually be assumed "large enough" to reflect the situation that "more" is "better" in most practical situations. Based on these assumptions, limiting processes will most often be concerned with p fixed and $n \to \infty$.

Remark 1.1.1 *Let* $\mathbf{x}' = T\mathbf{x}$ *be a linear (contravariant) transformation of* $\mathbf{x}$. *Then the equation defining the residual vector* $\mathbf{r}$ *in (1.1.1) transforms to*

$$\mathbf{r} = AT^{-1}\mathbf{x}' - \mathbf{b}.$$

That is the rows of A transform covariantly. In this sense equation (1.1.1) is invariant under diagonal rescaling of $\mathbf{x}$, *and only trivially modified by multiplication by a scalar. Thus it possible to impose scaling that makes quantities in (1.1.1) commensurate. For example, diagonal rescaling of* $\mathbf{x}$

allows the scaling of the columns of A and scalar multiplication scales the right hand side so that it is possible the satisfy the conditions

$$\frac{1}{\sqrt{n}}\|A_{*i}\| = 1,\ i = 1, 2, \cdots, p,\ \frac{1}{\sqrt{n}}\|\mathbf{b}\| = 1$$

where the norm is the usual euclidean norm.

The necessary conditions for a minimum for the sum of squares in (1.1.1) give

$$0 = \nabla_x \mathbf{r}^T\mathbf{r} = 2\mathbf{r}^T A. \tag{1.1.2}$$

Substituting for $\mathbf{r}$ from (1.1.1) gives the normal equations

$$A^T A\mathbf{x} = A^T\mathbf{b}. \tag{1.1.3}$$

This system defines the least squares estimator $\mathbf{x}^{(n)}$ and the corresponding residual vector $\mathbf{r}^{(n)}$ uniquely provided the design matrix A has full column rank p, and a strengthened form of this condition will usually be assumed.

An important modelling context in which the linear least squares problem arrises is the following. Assume noisy observations are made on a system at a sequence of configurations labelled by a reference variable $t_i, i = 1, 2, \cdots, n$ which could be time. Let the data collected from this investigation be summarised as

$$b_i = y(t_i) + \varepsilon_i,\ i = 1, 2, \cdots, n, \tag{1.1.4}$$

where $y(t)$ is the error free signal (true model) which is assumed to be expressible in parametric linear form as

$$y(t) = \sum_{i=1}^{p} x_i^* \phi_i(t), \tag{1.1.5}$$

the x_i^*, $i = 1, 2, \cdots, p$, are the (hypothesised) true parameter values, and the ε_i are random variables summarising the noise in the observations. A standard assumption would be that the ε_i are independent and have a normal distribution with mean 0, and standard deviation σ ($\varepsilon \backsim N(0, \sigma^2 I)$). In this case

$$A_{ij} = \phi_j(t_i),\ j = 1, 2, \cdots, p,\ i = 1, 2, \cdots, n.$$

It is important to know how the estimate $\mathbf{x}$ of $\mathbf{x}^*$ given by (1.1.3) behaves as the number of observations made in the investigation increases without limit because this permits statements to be made about rates of convergence of the parameter estimates to their true values. This is not just a theoretical point because it provides information on how much data needs to be collected in

order to be able to predict the model structure with confidence and so is directly relates to the practicality of the measurement exercise. It is also important to know how the computational algorithm chosen to solve (1.1.1) will behave on large data sets. In this connection, the first point to make is that a systematic process of designed experiments capable of automation is required to generate the values of the reference variable t_i and record the observations b_i associated with large data sets. In this sense there is a requirement for a *sequence of designed experiments* . The nature of this recording process must depend on the nature of the system being observed. Here two cases are distinguished:

1. The system has the property that after a finite horizon for the labelling variable t no further information on model structure is available. One case corresponds to signals decaying to zero. Such processes are called transient and can be expected to have relatively simple stability properties with imposed perturbations also dying away. However, the case of a finite observation window dictated by external factors is also included. Such systems may be required to be controlled, and may have much more complicated stability properties. The refinement process needed to increase the amount of information available is one in which independent trials are performed to obtain data for an increasing sequence of values of n. Usually it will be convenient to assume that transient processes correspond to measurements in the interval $[0, 1]$. Here the use of the descriptor systematic is interpreted to mean that the successive sampling regimes generate sequences of points $\left\{t_i^{(n)},\ i = 1, 2, \cdots n\right\}$ for increasing values of n associated with a limiting process such that

$$\frac{1}{n}\sum_{i=1}^{n} f(t_i^{(n)}) \to \int_0^1 f(t)dw(t),\ n \to \infty, \tag{1.1.6}$$

holds for all sufficiently smooth $f(t)$ ($f(t) \in C\,[0, 1]$ for example) where $w(t)$ is a weight function characteristic of the sampling regime. The left hand side in (1.1.6) can be interpreted as a simple quadrature formula . For example, $w(t) = t$ in the two cases:

   (a) The t_i are equispaced. The corresponding quadrature error for smooth enough $f(t)$ is strictly $O(1/n)$.

   (b) The t_i are uniformly distributed in $[0, 1]$. The corresponding quadrature error is asymptotically normally distributed with variance $O\,(1/n)$.

A sampling scheme for estimating a transient system satisfying this requirement for a designed experiment is called *regular* to stress that there is a sense in which the quadrature error is $o(1)$ asymptotically as $n \to \infty$. The above example shows that regular sampling may require different interpretations in different settings. It will be tacitly assumed that $w(t)$ has no intervals of constancy so that asymptotically the observation points fill out $[0, 1]$.

2. The system is persistent in the sense that observations made for arbitrary large values of t provide useful information on model structure. Often it is the case that system behaviour is largely independent of the time at which measurement commences. Here there may be no a priori reason for carrying out the graded sequence of experiments characterising the transient case, and the conceptually simplest sampling procedure keeps incrementing t by a constant amount so that $t_i = i\Delta$, $i = 1, 2, \cdots$. In this case difficulties may occur because the fixed sampling interval could have problems in resolving signals of too high a frequency. In nonlinear problems (for example, relaxation oscillations) the stability properties of the linearized model equations can be relatively complicated.

Now assume that the design matrix in (1.1.1) is constructed in association with a regular sampling scheme. Then the regularity condition gives

$$\begin{aligned} \frac{1}{n} A_{*i}^{(n)T} A_{*j}^{(n)} &= \frac{1}{n} \sum_{k=1}^{n} \phi_i(t_k^{(n)}) \phi_j(t_k^{(n)}) \\ &\to \int_0^1 \phi_i(t)\phi_j(t) dw(t) = G_{ij}. \end{aligned} \tag{1.1.7}$$

This shows that for large n the normal matrix is approximately proportional to the Gram matrix $G : R^p \to R^p$ determined by the weight function associated with the sampling procedure. The following result is immediate, and sets the paradigm for the form of the rank assumptions that will be made on the sequence of design matrices $A^{(n)}$.

Lemma 1.1 *Let $\sigma_i(A^{(n)})$, $i = 1, 2, \cdots, p$ be the singular values of $A^{(n)}$, and let the corresponding eigenvalues of G be λ_i, $i = 1, 2, \cdots, p$. Then*

$$\sigma_i(A^{(n)}) \to (n\lambda_i)^{1/2}, \; n \to \infty. \tag{1.1.8}$$

If G is positive definite, then $A^{(n)}$ has full rank for all n large enough.

p	$\kappa(G)$	$t_n(p)$
2	1.93 01	1
3	5.24 02	2
4	1.55 04	4
5	4.77 05	5
6	1.50 07	7
7	4.75 08	8
8	1.53 10	10
9	4.92 11	11
10	1.60 13	13

Table 1.1: Condition number for leading segments of the Hilbert matrix

Example 1.1.1 *Consider the case of polynomial regression :*

$$\phi_i(t) = t^{i-1},\ i = 1, 2, \cdots, p,$$

and the regular sampling scheme corresponding to choosing the values of the labelling variable to be equispaced:

$$t_i^{(n)} = (i-1)h^{(n)},\ h^{(n)} = \frac{1}{n-1},\ i = 1, 2, \cdots, n.$$

It is easy to verify this sampling scheme is regular by noting that (1.1.7) corresponds to a rectangular quadrature rule with weight function $w(t) = t$. This gives

$$G_{ij} = \frac{1}{i+j-1},\ 1 \le i, j \le p.$$

The Gram matrix can be identified with the $p \times p$ principal minor of the Hilbert matrix . While this is known to be positive definite for all finite p, it is also known to be notoriously ill conditioned. That is the condition number

$$\kappa(G) = \|G\| \left\|G^{-1}\right\|$$

is desperately large. This is illustrated in Table 1.1 where the entries are in floating point notation and have been rounded to three significant figures.

Remark 1.1.2 *The connection of regular sampling schemes with potentially ill conditioned limiting matrices raises interesting questions about our ability*

to obtain information on model parameters even assuming we start with a true model. To explore this further let $A^{(n)}$ *have the orthogonal factorization*

$$A^{(n)} = \begin{bmatrix} Q_1^{(n)} & Q_2^{(n)} \end{bmatrix} \begin{bmatrix} U^{(n)} \\ 0 \end{bmatrix} \tag{1.1.9}$$

where $Q^{(n)} : R^n \to R^n$ *is orthogonal, and* $U^{(n)} : R^p \to R^p$ *is upper triangular. Construction of such a factorization is described in the next chapter. Let* $\boldsymbol{\varepsilon} \backsim N(0, \sigma^2 I)$*. Then*

$$\begin{aligned} \mathbf{x}^{(n)} - \mathbf{x}^* &= U^{(n)-1} Q_1^{(n)T} \boldsymbol{\varepsilon}, \\ &= \left(\frac{1}{\sqrt{n}} U^{(n)} \right)^{-1} \frac{1}{\sqrt{n}} Q_1^{(n)T} \boldsymbol{\varepsilon}. \end{aligned}$$

It follows from the orthogonality of the rows of $Q_1^{(n)T}$ *that the components of* $Q_1^{(n)T}\boldsymbol{\varepsilon}$ *are also independent random variables with distribution* $N(0, \sigma^2)$*. If* G *is bounded, positive definite then*

$$\frac{1}{n} U^{(n)T} U^{(n)} \to G,$$

so that the term $\frac{1}{\sqrt{n}} U^{(n)}$ *tends to a Cholesky factor and hence is bounded in norm as* $n \to \infty$*. It follows that* $\mathbf{x}^{(n)} - \mathbf{x}^*$ *has a distribution which is multivariate normal with mean zero and variance*

$$\begin{aligned} \mathcal{V}\left\{ \mathbf{x}^{(n)} - \mathbf{x}^* \right\} &= \mathcal{E}\left\{ \left(\frac{1}{\sqrt{n}} U^{(n)} \right)^{-1} \frac{1}{\sqrt{n}} Q_1^{(n)T} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^T Q_1^{(n)} \frac{1}{\sqrt{n}} \left(\frac{1}{\sqrt{n}} U^{(n)} \right)^{-T} \right\}, \\ &= \frac{\sigma^2}{n} \left(\frac{1}{n} A^{(n)T} A^{(n)} \right)^{-1}, \\ &\to \frac{\sigma^2}{n} G^{-1}, \ n \to \infty. \end{aligned}$$

It is instructive to write this result in the approximate form, valid for n *large enough,*

$$\sqrt{n}\left(\mathbf{x}^{(n)} - \mathbf{x}^* \right) \backsim N\left(0, \sigma^2 G^{-1} \right).$$

This can be interpreted as showing a rate of convergence of $\mathbf{x}^{(n)}$ *to* $\mathbf{x}^*$ *of* $O\left(n^{-1/2} \right)$*. However, it also indicates that the discrepancy involves a random component, and that the spread of the distribution, and hence the size of confidence intervals for the components of* $\mathbf{x}^{(n)}$*, depends on the elements of the inverse of* G*. The implications of this is indicated in the third column of Table 1.1 which records* $t_n(p)$*, the power of* 10 *such that the order*

of magnitude of n gives $\|G^{-1}\|/n \approx 1$. This can be bounded easily for the range of values of p considered as $1 \leq \|G\| \leq 2$. The significance of this value of n is that the spread of the resulting random variable is determined by σ, the scale of the error in the observations. To gain any improvement in the parameter estimates obtained from the measurements made in a single experiment larger values of n are required.

Exercise 1.1.1 *Verify the form of $\omega(t)$ for the particular cases (a) equispaced observation points, (b) uniformly distributed observation points, and (c) observation points chosen as zeros of T_n, the Chebyshev polynomial of degree n shifted to the interval $[0, 1]$. Which set leads to the best conditioned Gram matrix?*

1.1.1 Under- and over-specified models

While the presumption that the posited model is correct is a valid working assumption in many circumstances there are also important situations in which an appropriate form of model is the question of major interest. Examples include:

1. Exploratory data analysis where the question asked is can the observed data of interest be explained adequately by a model based on a selection of terms from available covariate data.

2. Model economisation where a complex interacting system needs to be explained in terms of a dominant set of reactions for purposes of efficient control.

There are two basic approaches to developing suitable models in these cases:

1. To start with a minimal (typically underspecified) model and add design variables which meet a test of effective explanatory power until an adequate representation of the data is obtained.

2. To start with a maximum (typically overspecified) model and delete ineffective variables until an appropriate economy has been achieved.

The second approach is more likely to run into computational problems when the Gram matrix G corresponding to overspecified models is illconditioned. On the other hand, the first approach is forced to make decisions on information obtained from inadequate representations of the data.

To illustrate an underspecified model , let the true model be

$$\mathbf{b} = \begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} \mathbf{x}^* \\ \mathbf{y}^* \end{bmatrix} - \boldsymbol{\varepsilon}$$

where A constitutes the columns of the underspecified design and $\boldsymbol{\varepsilon} \backsim N(0, \sigma^2 I)$. Let an orthogonal factorization of the true design be

$$\begin{bmatrix} A & B \end{bmatrix} = \begin{bmatrix} Q_1 & Q_2 & Q_3 \end{bmatrix} \begin{bmatrix} U_1 & U_{12} \\ 0 & U_2 \\ 0 & 0 \end{bmatrix}.$$

The underspecified problem is

$$\min_{\mathbf{x}} \mathbf{s}^T\mathbf{s};\ \mathbf{s} = A\mathbf{x} - \mathbf{b},$$

and has solution

$$\begin{aligned} \widehat{\mathbf{x}} &= U_1^{-1}Q_1^T\mathbf{b}, \\ &= U_1^{-1}Q_1^T \left\{ \begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} \mathbf{x}^* \\ \mathbf{y}^* \end{bmatrix} - \boldsymbol{\varepsilon} \right\} \end{aligned}$$

so that

$$\widehat{\mathbf{x}} - \mathbf{x}^* = U_1^{-1}Q_1^T \left\{ B\mathbf{y}^* - \boldsymbol{\varepsilon} \right\}.$$

The estimate $\widehat{\mathbf{x}}$ is necessarily biassed unless $Q_1^T B = 0$ - that is the additional columns of the true design are orthogonal to the columns of the underspecified design.

To illustrate an overspecified model let the true model be

$$\mathbf{b} = A\mathbf{x}^* - \boldsymbol{\varepsilon},$$

and let the overspecified design matrix be $\begin{bmatrix} A & B \end{bmatrix}$. Then the overspecified problem is

$$\min_{\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}} \mathbf{s}^T\mathbf{s};\ \mathbf{s} = \begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} \mathbf{x} - \mathbf{x}^* \\ \mathbf{y} \end{bmatrix} + \boldsymbol{\varepsilon}$$

so that

$$\begin{bmatrix} \widehat{\mathbf{x}} - \mathbf{x}^* \\ \widehat{\mathbf{y}} \end{bmatrix} = - \begin{bmatrix} U_1^{-1} & -U_1^{-1}U_{12}U_2^{-1} \\ 0 & U_2^{-1} \end{bmatrix} \begin{bmatrix} Q_1^T\boldsymbol{\varepsilon} \\ Q_2^T\boldsymbol{\varepsilon} \end{bmatrix}.$$

Here $\begin{bmatrix} \widehat{\mathbf{x}} \\ \widehat{\mathbf{y}} \end{bmatrix}$ is an unbiassed estimator of $\begin{bmatrix} \mathbf{x}^* \\ 0 \end{bmatrix}$, but there is overfitting manifesting itself in the determination of $\widehat{\mathbf{y}}$ by the noise. The variance of the

estimator $\widehat{\mathbf{x}}$ is increased by comparison with $\mathbf{x}^{(n)}$:

$$\begin{aligned}\mathcal{V}\left\{\widehat{\mathbf{x}}\right\} &= U_1^{-1}\mathcal{E}\left\{\left\{Q_1^T\boldsymbol{\varepsilon}-U_{12}\widehat{\mathbf{y}}\right\}\left\{Q_1^T\boldsymbol{\varepsilon}-U_{12}\widehat{\mathbf{y}}\right\}^T\right\}U_1^{-T},\\ &= \sigma^2\left\{U_1^{-1}U_1^{-T}+Z_{12}Z_{12}^T\right\},\\ &= \mathcal{V}\left\{\mathbf{x}^{(n)}\right\}+\sigma^2 Z_{12}Z_{12}^T,\end{aligned}$$

where

$$Z_{12} = -U_1^{-1}U_{12}U_2^{-1}.$$

Again the exceptional case corresponds to

$$0 = U_{12} = U_1^T U_{12} = U_1^T Q_1^T \left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right]\left[\begin{array}{c} U_{12} \\ U_2 \end{array}\right] = A^T B.$$

That is to the case of mutually orthogonal sets of design variables.

Exercise 1.1.2 *Verify in detail the computation of the means and variances of the solution variables in both the under- and over-determined cases.*

1.1.2 Stepwise regression

Stepwise regression constitutes an important application of the above ideas to model exploration. Here a set of possible explanatory variables A_{*i}, $i = 1, 2, \cdots, p$, is given, and it is required to find an efficient subset for representing a given data vector. A tentative model is summarised by an index set σ with $|\sigma| = k$,

$$\sigma = \left\{\sigma(1), \sigma(2), \cdots \sigma(k)\right\},$$

which points to the columns of the current design A^σ in the sense that $\sigma(i)$ holds the column number in A of the column currently in the i'th position in A^σ. We let the solution of the corresponding least squares problem be $\mathbf{r}^\sigma, \mathbf{x}^\sigma$ and seek to test the effectiveness of the variable A_{*m} where $m = \sigma(j)$. It is convenient to swop the variable in question to the last position so that the design matrix becomes

$$A^\sigma P_m = \left[\begin{array}{cc} A^m & A_{*m} \end{array}\right]$$

where A^m is the partial design matrix which results when columns j and k of A^σ are exchanged and then $A^\sigma_{*k} = A_{*m}$ omitted. Here P_m is the permutation matrix which interchanges columns j, k. Let the "hat matrix"

$$H^m = A^m\left(A^{mT}A^m\right)^{-1}A^{mT} \tag{1.1.10}$$

be the projection onto the range of A^m. If $\mathbf{r}^m$ is the vector of residuals corresponding to the solution $\mathbf{x}^m$ of the least squares problem with design matrix A^m then straightforward calculations give

$$P_m \mathbf{x}^\sigma = \begin{bmatrix} \mathbf{x}^m - \frac{A_{*m}^T (I-H^m)\mathbf{b}}{A_{*m}^T (I-H^m) A_{*m}} \left(A^{mT} A^m\right)^{-1} A^{mT} A_{*m} \\ \frac{A_{*m}^T (I-H^m)\mathbf{b}}{A_{*m}^T (I-H^m) A_{*m}} \end{bmatrix}, \tag{1.1.11}$$

$$\left(A^{\sigma T} A^\sigma\right)_{jj}^{-1} = \frac{1}{A_{*m}^T \left(I - H^m\right) A_{*m}}, \tag{1.1.12}$$

$$\|\mathbf{r}^\sigma\|^2 = \|\mathbf{r}^m\|^2 - \frac{\left(A_{*m}^T \left(I - H^m\right) \mathbf{b}\right)^2}{A_{*m}^T \left(I - H^m\right) A_{*m}} = \frac{\left(x_j^\sigma\right)^2}{\left(A^\sigma\right)_{jj}^{-1}}. \tag{1.1.13}$$

These formulae provide a basis both for entering a single variable into the current selection and for deleting a single variable from the current selection. Typically this is done by testing the hypothesis $H_0 : \ x_j^\sigma = 0$ by computing the t statistic

$$t = \frac{\left|x_j^\sigma\right|}{\left(s_\sigma \left(\left(A^{\sigma T} A^\sigma\right)_{jj}^{-1}\right)\right)^{1/2}}.$$

where s_σ is the current estimate of variance

$$s_\sigma = \frac{\|\mathbf{r}^\sigma\|^2}{ndf},$$

where *ndf* is the effective degrees of freedom. Equivalently the corresponding F statistic could be used.

There are two important shortcomings with this procedure [70]:

1. This procedure cannot be guaranteed to give best subsets in any global sense. In particular, it is easy to construct examples in which the best subset of k variables does not contain the best subset of $k-1$ variables.

2. The procedure is open to the criticism of selection bias if the same data is used both in the selection of the model and in subsequent data analysis.

Exercise 1.1.3 *1. Verify the update equations (1.1.11 - 1.1.13).*
2. Construct an example in which the best set of two variables does not contain the best single variable set.

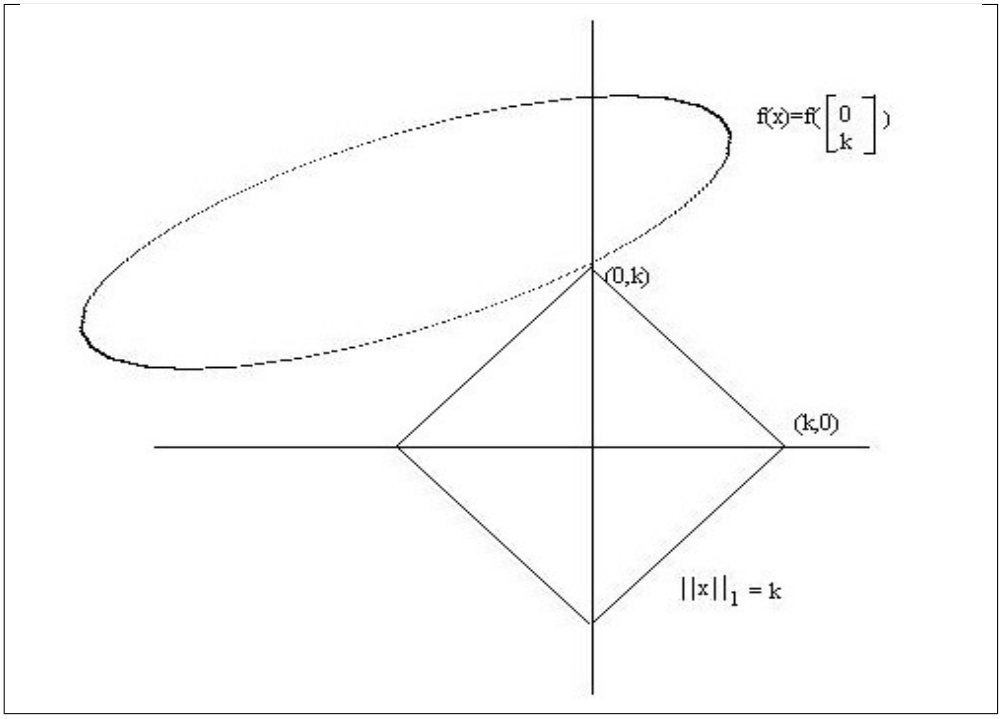

Table 1.2: The l_1 norm provides a mechanism for variable selection

1.1.3 The lasso

The lasso provides an interesting alternative to stepwise regression [104], [79]. Here the basic problem solved is

$$\min_{\|\mathbf{x}\|_1 \leq \kappa} \frac{1}{2} \|\mathbf{r}\|^2 .$$

The method uses properties of the l_1 norm illustrated for a two variable problem in the following figure. The constraint region centred on the origin has the characteristic form appropriate to the l_1 norm and is scaled by κ. It just touches the critical ellipsoidal contour of $\frac{1}{2} \|\mathbf{r}\|^2$ where it intersects the x_2 axis. It follows that $x_1 = 0$ so that only the second function is selected in this example.

This pattern of selection is followed in general. The analogue of the stepwise regression procedure proves to be a piecewise linear homotopy which steps from $\kappa = 0$ to $\kappa = \|\widehat{\mathbf{x}}\|_1$ where $\widehat{\mathbf{x}}$ is the unconstrained minimizer of $\frac{1}{2} \|\mathbf{r}\|^2$. Each slope discontinuity that occurs as κ is increased corresponds to addition or deletion of (generically) just one component variable to the selected set. As κ is further increased this new selection moves away from zero initially. In contrast to stepwise regression, the lasso has the advantage

of optimality for each point of the homotopy trajectory. It has the current disadvantage that statistical testing of the estimates is not well developed.

1.1.4 The case of an intercept term

In the case that the design matrix can be partitioned in the form

$$A = \left[\begin{array}{cc} \mathbf{e} & A_1 \end{array} \right] \tag{1.1.14}$$

then the intercept variable corresponding to the column of 1's can be removed from the computation and its value then determined by solving a reduced problem involving one fewer variable. Here it is said that the model contains an *intercept* term. Models of this form are common in statistical computation. The reduction argument is as follows. Let $\mathbf{r}^{(n)}$ be the optimal residual vector. Then it is a consequence of the necessary conditions (1.1.1) that

$$\begin{aligned} A^T\mathbf{r}^{(n)} = 0 &\Rightarrow \mathbf{e}^T\mathbf{r}^{(n)} = 0. \\ &\Rightarrow (I-P)\,\mathbf{r}^{(n)} = \mathbf{r}^{(n)} \end{aligned}$$

where $P = \mathbf{e}\mathbf{e}^T/n$ is a projection matrix. Now consider the linear least squares problem based on the linear model

$$\mathbf{s} = \overline{A}\mathbf{y} - \overline{\mathbf{b}}$$

where

$$\begin{aligned} \overline{A} &= (I-P)\,A_1 = \left[\begin{array}{ccc} \cdots & A_{*i} - \frac{\mathbf{e}^T A_{*i}}{n}\mathbf{e} & \cdots \end{array} \right],\ i = 2, \cdots, p, \\ \overline{\mathbf{b}} &= (I-P)\,\mathbf{b} \ = \mathbf{b} - \frac{\mathbf{e}^T\mathbf{b}}{n}\mathbf{e}. \end{aligned}$$

The design matrix for this reduced problem has full rank if and only if A has full rank. Now $(I-P)\,\mathbf{s} = \mathbf{s}$ by construction so that, if $\mathbf{z}^{(n)}, \mathbf{s}^{(n)}$ is the optimal solution, then

$$\overline{A}^T\mathbf{s}^{(n)} = A_1^T\mathbf{s}^{(n)} = \left[\begin{array}{c} \mathbf{e}^T \\ A_1^T \end{array} \right] \mathbf{s}^{(n)} = 0.$$

Also

$$\begin{aligned} \mathbf{s}^{(n)} &= \overline{A}\mathbf{z}^{(n)} - \overline{\mathbf{b}}, \\ &= \left(\frac{\mathbf{e}^T}{n} \left(\mathbf{b} - \sum_{i=2}^{p} A_{*i} z_i^{(n)} \right) \right) \mathbf{e} + A_1\mathbf{z}^{(n)} - \mathbf{b} \\ &= A \left[\begin{array}{c} \frac{\mathbf{e}^T}{n} \left(\mathbf{b} - \sum_{i=2}^{p} A_{*i} z_i^{(n)} \right) \\ \mathbf{z}^{(n)} \end{array} \right] - \mathbf{b} \end{aligned}$$

It follows that

$$\mathbf{r}^{(n)} = \mathbf{s}^{(n)},\ \mathbf{x}^{(n)} = \begin{bmatrix} \frac{\mathbf{e}^T}{n}\left(\mathbf{b} - \sum_{i=2}^{p} A_{*i} z_{i-1}^{(n)}\right) \\ \mathbf{z}^{(n)} \end{bmatrix}$$

solves (1.1.1). The operation of removing the mean values from the components of the columns of A_1 and of the right hand side $\mathbf{b}$ is called *centring* .

1.2 The generalised least squares problem

The simplest step beyond the assumption that the errors have a standard normal distribution corresponds to the assumption that the ε_i have the multivariate normal distribution

$$\boldsymbol{\varepsilon} \backsim N(0, V), \tag{1.2.1}$$

where $V = \mathcal{E}\left\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right\}$ is the variance covariance matrix (assumed for the moment to be nonsingular and therefore positive definite). It follows that

$$\widetilde{\boldsymbol{\varepsilon}} = V^{-1/2}\boldsymbol{\varepsilon} \backsim N(0, I),$$

and this permits the transformed linear model

$$\widetilde{\mathbf{r}} = V^{-1/2} A\left(\mathbf{x} - \mathbf{x}^*\right) - \widetilde{\boldsymbol{\varepsilon}},$$

to be identified with the previous discussion of the linear least squares problem. Substitution leads to the generalised least squares problem

$$\min \mathbf{r}^T V^{-1} \mathbf{r};\ \mathbf{r} = A\mathbf{x} - \mathbf{b}. \tag{1.2.2}$$

The necessary conditions for this problem are

$$\mathbf{r}^T V^{-1} A = 0,$$

giving the linear system determining $\mathbf{x}^{(n)}$ in the form

$$A^T V^{-1} A\mathbf{x} = A^T V^{-1}\mathbf{b}. \tag{1.2.3}$$

The corresponding distributional results are that $\mathbf{x}^{(n)} - \mathbf{x}^*$ has a multivariate normal distribution with

$$\mathcal{E}\left\{\mathbf{x}^{(n)} - \mathbf{x}^*\right\} = 0,$$
$$\mathcal{V}\left\{\mathbf{x}^{(n)} - \mathbf{x}^*\right\} = \frac{1}{n}\left\{\frac{1}{n} A^T V^{-1} A\right\}^{-1}.$$

In practice the above requirement that V is invertible is often associated with the implication that it makes good computational sense to invert the corresponding Choleski factors $V = LL^T$ where L is lower triangular. The computation of this factorization is discussed in Chapter 2. However, this strategy does not always make good sense, and an alternative approach which weakens the condition that V is stably invertible can be based on the constrained formulation

$$\min_{\mathbf{s},\mathbf{x}} \mathbf{s}^T\mathbf{s};\ L\mathbf{s} = A\mathbf{x} - \mathbf{b}, \tag{1.2.4}$$

where $LL^T = V$ and $\mathbf{r} = L\mathbf{s}$. Note that this provides an alternative interpretation of the transformation of random variables

$$\boldsymbol{\varepsilon} = L\widetilde{\boldsymbol{\varepsilon}},\ \widetilde{\boldsymbol{\varepsilon}} \backsim N(0, I)$$

considered above. The two formulations are clearly equivalent when L is nonsingular and the revised formulation has the potential to make sense in at least some cases of singular V. Note that a sufficient condition for the constraint equation to be consistent is range$\left\{ \left[\begin{array}{cc} L & A \end{array} \right] \right\} = R^n$ so that (1.2.4) makes sense for arbitrary right hand side $\mathbf{b}$. The necessary conditions for (1.2.4) are

$$\left[\begin{array}{cc} \mathbf{s}^T & 0 \end{array} \right] = \boldsymbol{\lambda}^T \left[\begin{array}{cc} L & -A \end{array} \right],$$

where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers . Eliminating $\mathbf{s}$ between the necessary conditions and the constraint equations gives

$$\left[\begin{array}{cc} LL^T & -A \\ -A^T & 0 \end{array} \right] \left[\begin{array}{c} \boldsymbol{\lambda} \\ \mathbf{x} \end{array} \right] = \left[\begin{array}{c} -\mathbf{b} \\ 0 \end{array} \right].$$

This system can be well determined even when V is singular. Transforming this system using the orthogonal factorization of A (1.1.9) gives

$$\left[\begin{array}{cc} Q^TVQ & -\left[\begin{array}{c} U \\ 0 \end{array} \right] \\ -\left[\begin{array}{cc} U^T & 0 \end{array} \right] & 0 \end{array} \right] \left[\begin{array}{c} Q^T\boldsymbol{\lambda} \\ \mathbf{x} \end{array} \right] = \left[\begin{array}{c} -Q^T\mathbf{b} \\ 0 \end{array} \right]. \tag{1.2.5}$$

In certain circumstances this factorization provides an effective way of solving the augmented equations . It is called the null-space method in [73]. Setting

$$Q^TVQ = \left[\begin{array}{cc} V_{11} & V_{12} \\ V_{21} & V_{22} \end{array} \right],\ Q^T\boldsymbol{\lambda} = \left[\begin{array}{c} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \end{array} \right],$$

then the solution is given by:

$$\begin{aligned} \boldsymbol{\lambda}_1 &= 0, \\ \boldsymbol{\lambda}_2 &= -V_{22}^{-1}Q_2^T\mathbf{b}, \\ \mathbf{x}^{(n)} &= U^{-1}\left\{Q_1^T\mathbf{b} - V_{12}V_{22}^{-1}Q_2^T\mathbf{b}\right\}. \end{aligned} \tag{1.2.6}$$

Condition 1.1 *This shows that sufficient conditions for (1.2.5) to have a well determined solution are*

1. U is stably invertible . This would appear to be a natural condition to want to satisfy in any controlled modelling situation.
2. $Q_2^T V Q_2$ is well conditioned.

The corresponding form for the covariance matrix is

$$\mathcal{V}\left\{\mathbf{x}^{(n)}\right\} = U^{-1}\left\{V_{11} - V_{12}V_{22}^{-1}V_{21}\right\}U^{-T}.$$

One application of this result is to equality constrained least squares . Consider the problem

$$\min_{\mathbf{s},\mathbf{x}} \mathbf{s}^T\mathbf{s};\ \begin{bmatrix} 0_k & 0 \\ 0 & I_{n-k} \end{bmatrix}\mathbf{s} = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}\mathbf{x} - \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}, \tag{1.2.7}$$

where 0_k is a $k \times k$ zero matrix. This corresponds to a least squares problem with k equality constraints where these consist of the first k equations with constraint matrix A_1. Here the augmented matrix in (1.2.5) has the form

$$\begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & I_{n-k} \end{bmatrix} & \begin{bmatrix} -A_1 \\ -A_2 \end{bmatrix} \\ \begin{bmatrix} -A_1^T & -A_2^T \end{bmatrix} & 0 \end{bmatrix}$$

when $k \leq p$, and the requirement that this is nonsingular is the condition for a well determined problem. Note that the placement of the constraint equations is significant in ensuring V_{22} is nonsingular [41]. This connects with the argument for row pivoting in solving penalised least squares problems given in [90]. This approach considers the problem

$$\min_{\mathbf{x}} \mathbf{r}^T \begin{bmatrix} \gamma^{-2}I_k & \\ & I_{n-k} \end{bmatrix}\mathbf{r};\ \mathbf{r} = A\mathbf{x} - \mathbf{b},$$

and

$$L_\gamma = \begin{bmatrix} \gamma I_k & \\ & I_{n-k} \end{bmatrix} \to \begin{bmatrix} 0 & \\ & I_{n-k} \end{bmatrix},\ \gamma \to 0.$$

Exercise 1.2.1 *Show that the appropriate generalisation of (1.2.2) to the case of singular V is*

$$\min_x \mathbf{r}^T V^+ \mathbf{r};\ \mathbf{r} = A\mathbf{x} - \mathbf{b}, \tag{1.2.8}$$

where V^+ is the pseudo-inverse of V. Use this result to interpret the penalised least squares problem discussed above.

1.3 Minimum variance estimates

1.3.1 Gauss-Markov theory

The least squares estimate $\mathbf{x}^{(n)}$ has further interesting properties. The use of the covariance matrix V of the error term $\boldsymbol{\varepsilon}$ to scale the least squares problem makes good sense, but it is reasonable to ask if more of the structure of the problem could be exploited. For example, the estimation problem could be approached by asking is it possible to select an estimator $\mathbf{x}$ to minimize the expected mean square error. This requires

$$\min \mathcal{E}\left\{\|\mathbf{x}-\mathbf{x}^*\|_2^2\right\} \tag{1.3.1}$$

where the expectation is taken with respect to the density (1.2.1). Expanding the expectation term gives

$$\begin{aligned}\mathcal{E}\left\{\|\mathbf{x}-\mathbf{x}^*\|_2^2\right\} &= \mathcal{E}\left\{\begin{array}{c}\|\mathbf{x}-\mathcal{E}\{\mathbf{x}\}\|_2^2+2\left(\mathbf{x}-\mathcal{E}\{\mathbf{x}\}\right)^T\left(\mathcal{E}\{\mathbf{x}\}-\mathbf{x}^*\right)\\ +\|\mathcal{E}\{\mathbf{x}\}-\mathbf{x}^*\|_2^2\end{array}\right\},\\ &= V\{\mathbf{x}\}+\|\mathbf{x}^*-\mathcal{E}\{\mathbf{x}\}\|_2^2. \end{aligned}\tag{1.3.2}$$

Thus the expected mean square error can be decomposed into terms representing variance and squared bias respectively.

The estimator is constructed as a function of the data so it is natural to represent it as a mapping of the data vector $\mathbf{b}$. *a priori* this does not have to be a linear mapping. However, this choice of form of mapping families proves tractible. Let

$$\mathbf{x}^{(n)} = T\mathbf{b},\ T: R^n \to R^p, \tag{1.3.3}$$

and set

$$\begin{aligned}\mathbf{w} &= T\mathbf{b}-\mathbf{x}^* = T\left(A\mathbf{x}^*-\boldsymbol{\varepsilon}\right)-\mathbf{x}^*,\\ &= \left(TA-I\right)\mathbf{x}^*-T\boldsymbol{\varepsilon}.\end{aligned}$$

Then, as $\mathcal{E}\{\boldsymbol{\varepsilon}\}=0$, the objective function becomes

$$\begin{aligned}\mathcal{E}\left\{\|T\mathbf{b}-\mathbf{x}^*\|_2^2\right\} &= \mathcal{E}\left\{\|\mathbf{w}\|_2^2\right\},\\ &= \mathcal{E}\left\{\operatorname{trace}\left(\mathbf{w}\mathbf{w}^T\right)\right\},\\ &= \operatorname{trace}\left\{TVT^T\right\}+\|\left(TA-I\right)\mathbf{x}^*\|_2^2.\end{aligned}$$

This depends on the unobservable vector $\mathbf{x}^*$unless $TA-I=0$.

Remark 1.3.1 *This condition ensures that the linear estimator is* unbiassed*:*

$$\mathcal{E}\{\mathbf{x}\} = T\mathcal{E}\{\mathbf{b}\} = TA\mathbf{x}^* = \mathbf{x}^*.$$

It requires that A have full column rank, and no weaker form of this condition is possible without structural assumptions being made on $\mathbf{x}^*$.

Thus the construction of the minimum variance, linear, unbiassed estimator comes down to solving the problem

$$\min_T \text{trace}\left\{TVT^T\right\};\ TA = I. \tag{1.3.4}$$

This decomposes into the sequence of problems for the rows of the matrix T:

$$\min \mathbf{t}_i^T V \mathbf{t}_i;\ \mathbf{t}_i^T A = \mathbf{e}_i^T,\ \mathbf{t}_i = T_{i*},\ i = 1, 2, \cdots, p.$$

The necessary conditions for the i'th problem give

$$\mathbf{t}_i^T V = \boldsymbol{\lambda}_i^T A^T,$$

where $\boldsymbol{\lambda}_i \in R^p$ is the vector of Lagrange multipliers. Eliminating $\mathbf{t}_i$ using the constraint equation gives

$$\begin{aligned} \mathbf{e}_i^T &= \boldsymbol{\lambda}_i^T A^T V^{-1} A, \\ T &= \left(A^T V^{-1} A\right)^{-1} A^T V^{-1}. \end{aligned} \tag{1.3.5}$$

The resulting estimator is best minimum variance, linear, unbiassed. From (1.3.3), (1.3.5) it is also the generalised least squares estimator. Let Λ be the matrix whose rows are the Lagrange multiplier vectors. Then $\Lambda = \left(A^T V^{-1} A\right)^{-1}$. This result is known as the Gauss-Markov Theorem . To calculate the variance we have

$$\begin{aligned} \mathcal{E}\left\{\left(\mathbf{x}^* - \mathbf{x}^{(n)}\right)\left(\mathbf{x}^* - \mathbf{x}^{(n)}\right)^T\right\} &= \mathcal{E}\left\{\left(\left(TA - I\right)\mathbf{x}^* - T\boldsymbol{\varepsilon}\right)\left(\left(TA - I\right)\mathbf{x}^* - T\boldsymbol{\varepsilon}\right)^T\right\}, \\ &= \mathcal{E}\left\{T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T T^T\right\} = TVT^T, \\ &= \left(A^T V^{-1} A\right)^{-1} = \Lambda. \end{aligned} \tag{1.3.6}$$

Remark 1.3.2 *It is interesting that the (implicit) assumption that* V *is invertible is not used until the elimination leading to (1.3.5). Written out in full, this system of equations is*

$$\begin{bmatrix} T & \Lambda \end{bmatrix} \begin{bmatrix} V & -A \\ -A^T & 0 \end{bmatrix} = \begin{bmatrix} 0 & -I \end{bmatrix}. \tag{1.3.7}$$

To compute the resulting minimum variance, note that from (1.3.7)

$$TV - \Lambda A^T = 0$$

so that

$$TVT^T = \Lambda A^T T^T = \Lambda \tag{1.3.8}$$

by the condition that the estimator be unbiassed. Thus the matrix of Lagrange multipliers coincides with the covariance matrix .

Remark 1.3.3 *This analysis extends to cases in which the components of the design A are known random variables, independent of the data* $\mathbf{b}$*, with distributions independent of* $\mathbf{x}^*$*. It is necessary only to condition on the realization of A in the above calculations. If the design components are observed with error, and hence are unknown, then the problem becomes more difficult. This is the error in variables problem. Typically it requires additional information to resolve it.*

Exercise 1.3.1 *Complete the extension indicated in Remark 1.3.3.*

1.3.2 Prediction of random effects

What happens if $\mathbf{x}^*$ in the linear model is a random vector with known statistical properties? The problem now is to predict the particular realization of $\mathbf{x}^*$ given the observations, and is to be distinguished from the problem of estimating a vector of constants. In this case there is a sense that more information is available, and it is possible to characterize the predictor giving the realization of $\mathbf{x}^*$ which minimizes the expected mean square error if it is possible to compute the conditional expectation $\mathcal{E}\{\mathbf{x}^*|y(\cdot)\}$ of $\mathbf{x}^*$ given data $y(\cdot)$ [55]. Here the notation permits the possibility that the data is derived from more complicated objects than the finite collections of random variables which is the main type considered here. The identification of the best predictor requires the following lemma which is quoted without proof (see [8]).

Lemma 1.2 *The conditional expectation satisfies the condition*

$$\mathcal{E}\{(x - \mathcal{E}\{x|y(\cdot)\})\, g(y(\cdot))\} = 0$$

for all functionals $g(y(\cdot))$ *for which the expectation has meaning.*

Theorem 1.2 *The best predictor of* $\mathbf{x}^*$ *in the sense that it minimizes the expected mean square error of prediction is given by*

$$\mathbf{h}(y(.)) = \mathcal{E}\{\mathbf{x}^*|y(.)\}. \tag{1.3.9}$$

Proof. Let $\mathbf{h}(y(\cdot))$ be a predictor for the realization of $\mathbf{x}^*$. The calculation of the mean square error gives

$$\begin{aligned}
\mathcal{E}\left\{\|\mathbf{x}^* - \mathbf{h}(y(\cdot))\|_2^2\right\} &= \mathcal{E}\left\{\|\mathbf{x}^* - \mathcal{E}\{\mathbf{x}^*|y(\cdot)\} + \mathcal{E}\{\mathbf{x}^*|y(\cdot)\} - \mathbf{h}(y(\cdot))\|_2^2\right\}, \\
&= \mathcal{E}\left\{\|\mathbf{x}^* - \mathcal{E}\{\mathbf{x}^*|y(\cdot)\}\|_2^2\right\} + \\
&\quad \mathcal{E}\left\{\|\mathcal{E}\{\mathbf{x}^*|y(\cdot)\} - \mathbf{h}(y(\cdot))\|_2^2\right\} + \\
&\quad 2\mathcal{E}\left\{(\mathbf{x}^* - \mathcal{E}\{\mathbf{x}^*|y(\cdot)\})^T (\mathcal{E}\{\mathbf{x}^*|y(\cdot)\} - \mathbf{h}(y(\cdot)))\right\}.
\end{aligned}$$

The last term in this expression vanishes as a consequence of the lemma, while the second term is nonnegarive and vanishes provided (1.3.9) holds . The remaining term is independent of $\mathbf{h}(y(.))$ so the result follows. ■

The predictor (1.3.9) is linear if the error distributions are normal, but not otherwise. In the normal case the assumptions amount to the assumption that the mean and variance are known:

$$\mathcal{E}\{\mathbf{x}^*\} = \overline{\mathbf{x}},\ \mathcal{V}\{\mathbf{x}^*\} = R_{11}. \tag{1.3.10}$$

The problem is now specialised to that of predicting the realisation of $\mathbf{x}^*$ given the vector of observations $\mathbf{b}$ with

$$\mathcal{E}\{\mathbf{b}\} = \overline{\mathbf{b}},\ \mathcal{V}\{\mathbf{b}\} = R_{22},\ \mathcal{C}\{\mathbf{x}^*, \mathbf{b}\} = R_{12}. \tag{1.3.11}$$

Proceeding much as in the development of the Gauss-Markov theory, we seek the linear predictor

$$\mathbf{x} = \overline{\mathbf{x}} + T\left(\mathbf{b} - \overline{\mathbf{b}}\right) \tag{1.3.12}$$

to minimize the expected mean square error

$$\min_T \mathcal{E}\left\{\|(\mathbf{x}^* - \mathbf{x})\|_2^2\right\} = \min_T \mathcal{E}\left\{\left\|T\left(\mathbf{b} - \overline{\mathbf{b}}\right) - (\mathbf{x}^* - \overline{\mathbf{x}})\right\|_2^2\right\},$$

where the expectation is taken with respect to the joint distribution of $\mathbf{b}$, $\mathbf{x}^*$. It follows from (1.3.12) that this predictor is unbiassed in the sense that

$$\mathcal{E}\{\mathbf{x}^*\} = \mathcal{E}\{\mathbf{x}\}.$$

Expanding this gives (with $T_{i*} = \mathbf{t}_i^T$)

$$\min_T \mathcal{E}\left\{\sum_{i=1}^{p}\left[\left(\left(\mathbf{b} - \overline{\mathbf{b}}\right)^T \mathbf{t}_i\right)^2 - 2\left(x_i^* - \overline{x}_i\right)\left(\mathbf{b} - \overline{\mathbf{b}}\right)^T \mathbf{t}_i\right] + \|\mathbf{x}^* - \overline{\mathbf{x}}\|_2^2\right\}.$$

This leads to a series of minimization problems for each of the $\mathbf{t}_i$ separately. The necessary conditions for the i'th problem are

$$0 = \mathcal{E}\left\{\left(\mathbf{t}_i^T\left(\mathbf{b} - \overline{\mathbf{b}}\right)\left(\mathbf{b} - \overline{\mathbf{b}}\right)^T - \left(x_i^* - \overline{x}_i\right)\left(\mathbf{b} - \overline{\mathbf{b}}\right)^T\right)\right\},\ i = 1, 2, \cdots, p, \tag{1.3.13}$$

and these can be summarised as

$$0 = \mathcal{E}\left\{\left(\overline{\mathbf{x}} - \mathbf{x}^* + T\left(\mathbf{b} - \overline{\mathbf{b}}\right)\right)\left(\mathbf{b} - \overline{\mathbf{b}}\right)^T\right\} \tag{1.3.14}$$

$$= \mathcal{C}\left\{\mathbf{x}^*, \mathbf{b}^T\right\} - T\mathcal{V}\{\mathbf{b}\}. \tag{1.3.15}$$

Equations (1.3.12), (1.3.14) show that the error in the best prediction is uncorrelated with the data vector $\mathbf{b}$.

$$\begin{aligned}\mathcal{E}\left\{(\mathbf{x}-\mathbf{x}^*)\left(\mathbf{b}-\overline{\mathbf{b}}\right)^T\right\} &= \mathcal{E}\left\{\left((\mathbf{x}-\overline{\mathbf{x}})-(\mathbf{x}^*-\overline{\mathbf{x}})\right)\left(\mathbf{b}-\overline{\mathbf{b}}\right)^T\right\},\\ &= \mathcal{E}\left\{\left(T\left(\mathbf{b}-\overline{\mathbf{b}}\right)-(\mathbf{x}^*-\overline{\mathbf{x}})\right)\left(\mathbf{b}-\overline{\mathbf{b}}\right)^T\right\},\\ &= 0.\end{aligned}$$

Thus, in an important sense, all the information in the data has been used in constructing the prediction $\mathbf{x}$. The necessary conditions can be solved for the prediction matrix using (1.3.11) to give

$$T = \mathcal{C}\left\{\mathbf{x}^*,\mathbf{b}\right\}\mathcal{V}\left\{\mathbf{b}\right\}^{-1} = R_{12}R_{22}^{-1}. \tag{1.3.16}$$

It should be noted that the operation expressed by T is a projection. Assume for simplicity that means are zero, then

$$\begin{aligned}T^2\mathbf{b} &= \mathcal{C}\left\{\mathbf{x}^*,T\mathbf{b}\right\}\mathcal{V}\left\{T\mathbf{b}\right\}^{-1}T\mathbf{b},\\ &= \mathcal{C}\left\{\mathbf{x}^*,\mathbf{b}\right\}T^T\left\{T\mathcal{V}\left\{\mathbf{b}\right\}T^T\right\}^{-1}T\mathbf{b},\\ &= \mathcal{C}\left\{\mathbf{x}^*,\mathbf{b}\right\}T^T\left\{\mathcal{C}\left\{\mathbf{x}^*,\mathbf{b}\right\}\mathcal{V}\left\{\mathbf{b}\right\}^{-1}\mathcal{C}\left\{\mathbf{x}^*,\mathbf{b}\right\}^T\right\}^{-1}T\mathbf{b},\\ &= \mathcal{C}\left\{\mathbf{x}^*,\mathbf{b}\right\}\mathcal{V}\left\{\mathbf{b}\right\}^{-1}\mathcal{C}\left\{\mathbf{x}^*,\mathbf{b}\right\}^T\left\{\mathcal{C}\left\{\mathbf{x}^*,\mathbf{b}\right\}\mathcal{V}\left\{\mathbf{b}\right\}^{-1}\mathcal{C}\left\{\mathbf{x}^*,\mathbf{b}\right\}^T\right\}^{-1}T\mathbf{b},\\ &= T\mathbf{b}.\end{aligned}$$

Because T is a projection it follows that the necessary conditions (1.3.14) have the character of orthogonality conditions. The target space will be formlated as a Hilbert space of random variables.

Remark 1.3.4 *The best prediction result generalises in a couple of directions.*

1. *The best linear expected mean square predictor of $Z\mathbf{x}^*$ is $Z\left(\overline{\mathbf{x}}+T\left(\mathbf{b}-\overline{\mathbf{b}}\right)\right)$. It is only necessary to replace $\mathbf{x}^*$ by $Z\mathbf{x}^*$ in the above development.*

2. *If the expectation in mean square is replaced by a weighted least squares norm with weight matrix W then the best predictor is unchanged. Here we seek a predictor $\mathbf{x}=\overline{\mathbf{x}}+S\left(\mathbf{b}-\overline{\mathbf{b}}\right)$ by solving*

$$\begin{aligned}&\min_S \mathcal{E}\left\{\left(S\left(\mathbf{b}-\overline{\mathbf{b}}\right)-(\mathbf{x}^*-\overline{\mathbf{x}})\right)^T W^{-1}\left(S\left(\mathbf{b}-\overline{\mathbf{b}}\right)-(\mathbf{x}^*-\overline{\mathbf{x}})\right)\right\}\\ &= \min_{\widetilde{S}} \mathcal{E}\left\{\left\|\widetilde{S}\left(\mathbf{b}-\overline{\mathbf{b}}\right)-W^{-1/2}(\mathbf{x}^*-\overline{\mathbf{x}})\right\|_2^2\right\},\end{aligned}$$

where $\widetilde{S}=W^{-1/2}S$. It follows from the previous result that $\widetilde{S}=W^{-1/2}T$, so that $S=T$.

Remark 1.3.5 *The connection between best mean square prediction and conditional expectation permit convenient evaluation of the latter when the underlying distributions are multivariate normal. Let*

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \backsim N\left(\begin{bmatrix} \overline{\mathbf{x}} \\ \overline{\mathbf{y}} \end{bmatrix}, \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \right),$$

where $\mathbf{y} \in \mathbf{R}^m$ *and the lack of independence between* $\mathbf{x}$ *and* $\mathbf{y}$ *is indicated by nontrivial offdiagonal blocks in the covariance matrix. Equation (1.3.14) suggests seeking a representation*

$$\mathbf{x} = \overline{\mathbf{x}} + T\left(\mathbf{y} - \overline{\mathbf{y}}\right) + \mathbf{w} \tag{1.3.17}$$

such that $\mathbf{w}$ *is uncorrelated with* $\mathbf{y}$ *and has mean zero. Note that (1.3.17) can be interpreted as expressing an orthogonal decomposition in which covariance is used as an analogue of scalar product. Here the assumption that the variables are normally distributed implies that they are independent so that the joint density of* $\begin{bmatrix} \mathbf{w} \\ \mathbf{y} - \overline{\mathbf{y}} \end{bmatrix}$ *factorizes into the product of the respective marginal densities . The condition that* $\mathcal{C}\left\{\mathbf{w}, \mathbf{y}\right\} = 0$ *gives*

$$\mathcal{C}\left\{\mathbf{x}, \mathbf{y}\right\} = R_{12} = TR_{22} = T\mathcal{V}\left\{\mathbf{y}\right\},$$

whence (compare with (1.3.16))

$$T = \mathcal{C}\left\{\mathbf{x}, \mathbf{y}\right\} \mathcal{V}\left\{\mathbf{y}\right\}^{-1}. \tag{1.3.18}$$

The mean of $\mathbf{w}$ *is zero by assumption. The variance is given by*

$$\mathcal{V}\left\{\mathbf{w}\right\} = R_{11} - R_{12}R_{22}^{-1}R_{21}, \tag{1.3.19}$$

The best prediction of $\mathbf{x}$ *based on the information contained in* $\mathbf{y}$ *is*

$$\mathcal{E}\left\{\mathbf{x}|\mathbf{y}\right\} = \overline{\mathbf{x}} + T\left(\mathbf{y} - \overline{\mathbf{y}}\right) \tag{1.3.20}$$

and is equal to the corresponding conditional expectation by Theorem 1.2. Note that it is linear in $\mathbf{y}$*. The conditional distribution of* $\mathbf{x}$ *given* $\mathbf{y}$*,* $p\left(\mathbf{x}|\mathbf{y}\right)$*, in which* $\mathbf{y}$ *is treated as a constant vector, is*

$$p\left(\mathbf{x}|\mathbf{y}\right) = N\left(\overline{\mathbf{x}} + T\left(\mathbf{y} - \overline{\mathbf{y}}\right), R_{11} - R_{12}R_{22}^{-1}R_{21}\right). \tag{1.3.21}$$

Remark 1.3.6 *The analogy used above can be taken much further as it turns out this result has an important interpretation in terms of Hilbert spaces of random variables . Here this space* $\mathcal{H}\left\{\mathbf{x} - \overline{\mathbf{x}}, \mathbf{y} - \overline{\mathbf{y}}\right\}$ *is generated by the*

components of $\mathbf{x}-\overline{\mathbf{x}}$ *and* $\mathbf{y}-\overline{\mathbf{y}}$ *- that is each component of an element in* $\mathcal{H}$ *is a linear combination of all the components of both* $\mathbf{x}-\overline{\mathbf{x}}$ *and* $\mathbf{y}-\overline{\mathbf{y}}$*. Let* $\mathbf{w}\in\mathcal{H}$ *be defined by*

$$\mathbf{w}=W\left[\begin{array}{c}\mathbf{x}-\overline{\mathbf{x}}\\ \mathbf{y}-\overline{\mathbf{y}}\end{array}\right]$$

where the linear combination is defined by the matrix $W:R^{2n}\rightarrow R^{n}$*. The variance of an element provides the norm ,*

$$\left\|\mathbf{w}-\overline{\mathbf{w}}\right\|_{r}^{2}=\mathcal{E}\left\{\left\|\mathbf{w}-\overline{\mathbf{w}}\right\|^{2}\right\}=\operatorname{trace}\mathcal{V}\left\{\mathbf{w}\right\}, \tag{1.3.22}$$

provided $\mathcal{V}\left\{\left[\begin{array}{c}\mathbf{x}-\overline{\mathbf{x}}\\ \mathbf{y}-\overline{\mathbf{y}}\end{array}\right]\right\}$ *is nonsingular, and the corresponding scalar product of elements* $\mathbf{u},\ \mathbf{v}\in\mathcal{H}$ *is given by* $\operatorname{trace}\mathcal{C}\left\{\mathbf{u},\mathbf{v}\right\}$*, with expectations being taken with respect to the joint density. In this context the conditional expectation* $\mathcal{E}\left\{\mathbf{x}|\mathbf{y}\right\}$ *of* $\mathbf{x}$ *given* $\mathbf{y}$ *is computed using the orthogonal projection of* $\mathbf{x}-\overline{\mathbf{x}}$ *onto the subspace* $\mathcal{H}\left\{\mathbf{y}-\overline{\mathbf{y}}\right\}$ *generated by the components of* $\mathbf{y}-\overline{\mathbf{y}}$*. This orthogonal projection operation defined on* $\mathcal{H}\left\{\mathbf{x}-\overline{\mathbf{x}},\mathbf{y}-\overline{\mathbf{y}}\right\}$ *is just that given by the operator* T*. In typical applications* $\mathbf{w}$ *could be an error vector in which case* $\left\|\mathbf{w}-\overline{\mathbf{w}}\right\|_{r}^{2}$ *would be the expected mean square error.*

Exercise 1.3.2 *Sketch the geometrical picture showing the orthogonal projection and use the Hilbert space setting to derive the expression (1.3.20) for the conditional expectation.*

1.3.3 Mixed models

The following application establishes a connection between the the solution of the generalisd least squares problem (1.3.16) and a limiting case of best linear prediction

$$\mathbf{b}=A\mathbf{x}+\boldsymbol{\varepsilon},\ \mathbf{x}\backsim N\left(0,R\right),\ \boldsymbol{\varepsilon}\backsim N\left(0,V\right),\ \mathcal{C}\left(\mathbf{x},\boldsymbol{\varepsilon}\right)=0. \tag{1.3.23}$$

Then

$$\mathcal{E}\left\{\mathbf{b}\right\}=0,\ \mathcal{E}\left\{\mathbf{b}\mathbf{b}^{T}\right\}=ARA^{T}+V,$$

and

$$\mathcal{E}\left\{\mathbf{b}\mathbf{x}^{T}\right\}=A\mathcal{E}\left\{\mathbf{x}\mathbf{x}^{T}\right\}=AR.$$

This gives the prediction (1.3.16)

$$\widehat{\mathbf{x}}=RA^{T}\left(ARA^{T}+V\right)^{-1}\mathbf{b} \tag{1.3.24}$$

which is available provided $\mathcal{E}\left\{\mathbf{b}\mathbf{b}^T\right\}$ is invertible. By (1.6.7) in the chapter appendix this is the same as

$$\widehat{\mathbf{x}} = \left(A^T V^{-1} A + R^{-1}\right)^{-1} A^T V^{-1} \mathbf{b} \tag{1.3.25}$$

assuming that V and R have full rank. The variance calculation can make use of the same device. From (1.3.19) it follows that

$$\begin{aligned}\mathcal{V}\left\{\widehat{\mathbf{x}} - \mathbf{x}\right\} &= R - RA^T\left(ARA^T + V\right)^{-1} AR, \\ &= \left(A^T V^{-1} A + R^{-1}\right)^{-1}.\end{aligned}$$

Letting $R^{-1} \to 0$ gives the solution of the generalised least squares problem (1.2.3). This limiting process expresses a form of the assumption that there is no prior information on the values of the parameter vector $\mathbf{x}$ (assumption of a diffuse prior).

Also (1.3.25) can be expressed as the solution of the least squares problem

$$\min \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix}^T \begin{bmatrix} V^{-1} & \\ & R^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} \tag{1.3.26}$$

subject to the constraints

$$\begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} = \begin{bmatrix} A \\ I \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix}. \tag{1.3.27}$$

It is easy to verify that the variance calculation is also consistent. The Gauss-Markov correspondence gives

$$\begin{aligned} V_G &= \begin{bmatrix} V & \\ & R \end{bmatrix}, \\ T_G &= \left(A^T V^{-1} A + R^{-1}\right)^{-1} \begin{bmatrix} A^T & I \end{bmatrix} \begin{bmatrix} V^{-1} & \\ & R^{-1} \end{bmatrix}, \end{aligned}$$

so that (1.3.8) gives

$$\Lambda = T_G V_G T_G^T = \left(A^T V^{-1} A + R^{-1}\right)^{-1} = \mathcal{V}\left\{\widehat{\mathbf{x}} - \mathbf{x}\right\}. \tag{1.3.28}$$

This form proves convenient for the development of computational algorithms.

This argument can be extended to the case of mixed models . These have the form

$$\mathbf{b} = A\mathbf{x} + B\mathbf{z} + \boldsymbol{\varepsilon} \tag{1.3.29}$$

where now $\mathbf{x}$ is a vector of parameters (the fixed effects), $\mathbf{z}$ are the random effects uncorrelated with $\boldsymbol{\varepsilon}$ with known probability distribution $N(0, R_{22})$, and $\boldsymbol{\varepsilon} \sim N(0, V)$. The idea is to turn the mixed model problem into a prediction problem by associating the vector of parameters $\mathbf{x}$ with a random vector uncorrelated with both $\mathbf{z}$ and $\boldsymbol{\varepsilon}$ and having covariance matrix R_{11} such that R_{11}^{-1} is small and then taking the limiting solution of the resultant prediction problem by letting $R_{11}^{-1} \to 0$. The prediction problem sets

$$\mathbf{b} = \widetilde{A}\widetilde{\mathbf{x}} + \boldsymbol{\varepsilon},\ \mathcal{C}(\widetilde{\mathbf{x}}, \boldsymbol{\varepsilon}) = 0,$$

where

$$\widetilde{A} = \begin{bmatrix} A & B \end{bmatrix},\ \widetilde{\mathbf{x}} = \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix} \backsim N\left(0, \begin{bmatrix} R_{11} & 0 \\ 0 & R_{22} \end{bmatrix}\right).$$

An application of (1.3.25) now gives the prediction

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix} = \left[\widetilde{A}^T V^{-1} \widetilde{A} + R^{-1}\right]^{-1} \widetilde{A}^T V^{-1} \mathbf{b}.$$

The limiting process appropriate here is $R_{11}^{-1} \to 0$. This gives the system of equations

$$A^T V^{-1} A\mathbf{x} + A^T V^{-1} B\mathbf{z} = A^T V^{-1}\mathbf{b}, \tag{1.3.30}$$

$$B^T V^{-1} A\mathbf{x} + \left(B^T V^{-1} B + R_{22}^{-1}\right)\mathbf{z} = B^T V^{-1}\mathbf{b}. \tag{1.3.31}$$

The corresponding least squares formulation is

$$\min \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix}^T \begin{bmatrix} V^{-1} & 0 \\ 0 & R_{22}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix}$$

subject to the constraints

$$\begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & I \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix} - \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix}.$$

Exercise 1.3.3 *Formalise the estimation of* $\mathbf{x}$ *given the distribution of* $\mathbf{z}$ *as a generalised least squares problem with covariance* $V + BR_{22}B^T$*. Show that the solution of this problem is identical with that obtained by solving for* $\mathbf{z}$ *from (1.3.31), substituting the result in (1.3.30) to obtain an equation for* $\mathbf{x}$ *alone, and then simplifying the result using (1.6.7) in the chapter appendix.*

1.3.4 The projection theorem

What happens when additional information becomes available? Let

$$\mathbf{x}_1 = T_1 \left(\mathbf{y}_1 - \overline{\mathbf{y}}_1 \right), \; T_1 = \mathcal{C} \left\{ \mathbf{x}^*, \mathbf{y}_1 \right\} \mathcal{V} \left\{ \mathbf{y}_1 \right\}^{-1},$$

be the best (linear, minimum variance) prediction of $\mathbf{x}^* - \overline{\mathbf{x}}$ based on data $\mathbf{y}_1$. Then

$$\mathbf{x}^* = \overline{\mathbf{x}} + \mathbf{x}_1 + \mathbf{w}$$

where $\mathcal{C} \left\{ \mathbf{y}_1, \mathbf{w} \right\} = 0$ by (1.3.14), and $\mathbf{x}_1$ is the best approximation to $\mathbf{x}^* - \mathbf{x}$ from the Hilbert space of random variables $\mathcal{H} \left\{ \mathbf{y}_1 - \mathbf{y}_1 \right\}$ in the $\| \cdot \|_r$ norm (1.3.22). Let fresh data $\mathbf{y}_2$ become available. Then

$$\widetilde{\mathbf{y}}_2 = \mathbf{y}_2 - \mathcal{C} \left\{ \mathbf{y}_2, \mathbf{y}_1 \right\} \mathcal{V} \left\{ \mathbf{y}_1 \right\}^{-1} \left(\mathbf{y}_1 - \overline{\mathbf{y}}_1 \right)$$

is uncorrelated with $\mathbf{y}_1$ by (1.3.18)and so corresponds to the new information available. Setting

$$\mathbf{w} = T_2 \left(\widetilde{\mathbf{y}}_2 - \overline{\mathbf{y}}_2 \right) + \mathbf{w}_2,$$

where $\mathbf{w}_2$ is uncorrelated with $\widetilde{\mathbf{y}}_2$ (compare (1.3.17)), then

$$T_2 = \mathcal{C} \left(\mathbf{x}^*, \widetilde{\mathbf{y}}_2 \right) \mathcal{V} \left(\widetilde{\mathbf{y}}_2 \right)^{-1}.$$

Thus the best prediction of $\mathbf{x}^* - \overline{\mathbf{x}}$ based on the augmented data is given by

$$\mathbf{x}_2 = \mathbf{x}_1 + T_2 \left(\widetilde{\mathbf{y}}_2 - \overline{\mathbf{y}}_2 \right),$$

where the representation is based on decomposing the estimate into its projections onto the orthogonal spaces $\mathcal{H} \left\{ \mathbf{y}_1 - \overline{\mathbf{y}}_1 \right\}$ and $\mathcal{H} \left\{ \widetilde{\mathbf{y}}_2 - \overline{\mathbf{y}}_2 \right\}$.

Example 1.3.1 *A typical application of this projection result is the following. Assume* $\overline{\mathbf{x}} = 0$ *and let* $\mathbf{x}_1$ *be the best estimate of* $\mathbf{x}^*$ *where*

$$\begin{aligned} &\mathbf{b}_1 = A_1 \mathbf{x}^* + \boldsymbol{\varepsilon}_1, \; \mathcal{E} \left\{ \mathbf{b}_1 \right\} = 0, \\ &\boldsymbol{\varepsilon}_1 \backsim N \left(0, V_1 \right), \; \mathcal{C} \left\{ \mathbf{x}^*, \boldsymbol{\varepsilon}_1 \right\} = 0. \end{aligned}$$

Then

$$\mathbf{x}_1 = T_1 \mathbf{b}_1$$

where

$$T_1 = \mathcal{C} \left\{ \mathbf{x}^*, \mathbf{b}_1 \right\} \mathcal{V} \left\{ \mathbf{b}_1 \right\}^{-1} = R A_1^T \left\{ A_1 R A_1^T + V_1 \right\}^{-1}$$

by (1.3.16). The best approximation property is expressed in the orthogonality condition

$$\mathcal{C} \left\{ \mathbf{x}^* - \mathbf{x}_1, \mathbf{x}_1 \right\} = 0. \tag{1.3.32}$$

This follows from the necessary conditions (1.3.14) which give

$$0 = \mathcal{E}\left\{(\mathbf{x}^* - \mathbf{x}_1)\,\mathbf{b}_1^T\right\} \Rightarrow 0 = \mathcal{E}\left\{(\mathbf{x}^* - \mathbf{x}_1)\,\mathbf{b}_1^T T_1^T\right\}$$

in this case. If the new data is

$$\begin{aligned}&\mathbf{b}_2 = A_2\mathbf{x}^* + \boldsymbol{\varepsilon}_2,\\ &\boldsymbol{\varepsilon}_2 \backsim N(0, V_2),\ \mathcal{C}\{\mathbf{x}^*, \boldsymbol{\varepsilon}_2\} = 0,\ \mathcal{C}\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2\} = 0,\end{aligned}$$

then the updated estimate is

$$\mathbf{x}_2 = \mathbf{x}_1 + \mathcal{C}\left\{\mathbf{x}^*, \widetilde{\mathbf{b}}_2\right\}\mathcal{V}\left\{\widetilde{\mathbf{b}}_2\right\}^{-1}\widetilde{\mathbf{b}}_2 \tag{1.3.33}$$

where

$$\widetilde{\mathbf{b}}_2 = \mathbf{b}_2 - \mathcal{C}\{\mathbf{b}_2, \mathbf{b}_1\}\,\mathcal{V}\{\mathbf{b}_1\}^{-1}\mathbf{b}_1$$

is uncorrelated (here orthogonal as expectations are zero) with $\mathbf{b}_1$*.*

This result, which decomposes the best approximation into the sum of orthogonal components obtained by projecting into orthogonal subspaces, is known as the *projection theorem* . Note that if the condition $\mathcal{C}\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2\} = 0$ is inserted in the previous equation (this is the only point at which it is used) then this gives

$$\begin{aligned}\widetilde{\mathbf{b}}_2 &= \mathbf{b}_2 - A_2\mathcal{C}\{\mathbf{x}^*, \mathbf{b}_1\}\,\mathcal{V}\{\mathbf{b}_1\}^{-1}\mathbf{b}_1,\\ &= \mathbf{b}_2 - A_2\mathbf{x}_1.\end{aligned} \tag{1.3.34}$$

It also illustrates Remark 1.3.4 which emphasises that $A_2\mathbf{x}_1$ is the best approximation to $\mathbf{b}_2$ from $\mathcal{H}\{\mathbf{b}_1\}$ in this case. Let

$$\mathcal{V}\{\mathbf{x}_1 - \mathbf{x}^*\} = R_1,$$

then

$$\begin{aligned}\mathcal{C}\left\{\mathbf{x}^*, \widetilde{\mathbf{b}}_2\right\} &= \mathcal{C}\{\mathbf{x}^*, A_2(\mathbf{x}^* - \mathbf{x}_1)\},\\ &= \mathcal{C}\{\mathbf{x}^* - \mathbf{x}_1, \mathbf{x}^* - \mathbf{x}_1\}A_2^T,\\ &= R_1 A_2^T,\end{aligned}$$

using the best approximation property of $\mathbf{x}_1$ to $\mathbf{x}^*$ (1.3.32), and

$$\begin{aligned}\mathcal{V}\left\{\widetilde{\mathbf{b}}_2\right\} &= \mathcal{E}\left\{(\mathbf{b}_2 - A_2\mathbf{x}_1)(\mathbf{b}_2 - A_2\mathbf{x}_1)^T\right\},\\ &= A_2 R_1 A_2^T + V_2.\end{aligned}$$

This gives the revised estimate in the form

$$\mathbf{x}_2 = \mathbf{x}_1 + R_1 A_2^T\left(A_2 R_1 A_2^T + V_2\right)^{-1}(\mathbf{b}_2 - A_2\mathbf{x}_1).$$

Exercise 1.3.4 *Diagrams to illustrate the geometrical situations are very helpful here.*

1.4 Estimation of dynamic models

An important application of the projection theorem is the development via the discrete Kalman filter of a description of the evolution of an important class of discrete linear dynamical systems . Here the starting point is a state variable $\mathbf{x}_k = \mathbf{x}(t_k) \in R^p$ which describes the state of the system under consideration at time $t_k, k = 1, 2, \cdots, n$ and which satisfies the evolution (dynamics) equation or state equation

$$\mathbf{x}_{k+1} = X_k \mathbf{x}_k + \mathbf{u}_k,\ k = 1, 2, \cdots, n-1. \tag{1.4.1}$$

Information on the unobserved state variables is available through an observation equation ,

$$\mathbf{y}_k = H_k \mathbf{x}_k + \boldsymbol{\varepsilon}_k,\ k = 1, 2, \cdots, n, \tag{1.4.2}$$

and the requirement is to predict the realisation of the $\mathbf{x}_{k+1}$ given past values $\mathbf{x}_j, j = 1, 2, \cdots, k$. Here $X_k : R^p \to R^p$, while $H_k : R^p \to R^m,\ m < p$ is assumed to have full rank, and the random effects $\mathbf{u}_i \in R^p, i = 1, 2, \cdots, k,\ \boldsymbol{\varepsilon}_j \in R^m, j = 1, 2, \cdots, k+1$ are mutually independent, normally distributed, random vectors for all i, j with covariance matrices

$$\mathcal{V}\{\mathbf{u}_k\} = R_k,\ \mathcal{V}\{\boldsymbol{\varepsilon}_k\} = V_k.$$

It follows from the form of the state equations that past state variables are uncorrelated with the random effects in both state and observation equations,

$$\mathcal{C}\{\mathbf{x}_j, \mathbf{u}_k\} = 0,\ \mathcal{C}\{\mathbf{x}_j, \boldsymbol{\varepsilon}_k\} = 0,\ j \leq k.$$

Remark 1.4.1 *Typically conditions of observability and reachability are imposed on (1.4.1) and (1.4.2) [42]. The simplest case corresponds to $m = 1$ and $X_t = X$ constant. The resulting system is observable if the initial value ($\mathbf{x}_t$ say) of any consecutive sequence $\{t, t+1, \cdots, t+p-1\}$ can be recovered from the corresponding observations $\{y_t, y_{t+1}, \cdots, y_{t+p-1}\}$ - here $H = \mathbf{h}^T$. This requires that the matrix with columns $\left[\mathbf{h}, X^T\mathbf{h}, \cdots, \left(X^{p-1}\right)^T \mathbf{h}\right]$ has full rank. Reachability is the condition that a minimum length sequence $\mathbf{u}_t, \mathbf{u}_{t+1}, \cdots, \mathbf{u}_s$ can be imposed to drive $\mathbf{x}_{s+1}$ to any desired point. This can be expressed in similar algebraic form when appropriate assumptions are made about the covariance structure of $\mathbf{u}$.*

To start the recurrence requires information on an initial state. This could be either constant or random, but the random case, in which the initial state is assumed to be given by $\mathbf{x}_{1|0}$ with $\mathcal{E}\left\{\mathbf{x}_{1|0}\right\} = 0$, and covariance

$$S_{1|0} = \mathcal{E}\left\{\left(\mathbf{x}_1 - \mathbf{x}_{1|0}\right)\left(\mathbf{x}_1 - \mathbf{x}_{1|0}\right)^T\right\},$$

is the one considered here. Frequently $S_{1|0}$ is assumed to be large corresponding to the assumption of a diffuse prior , and it will be recalled that there is algebraic identity between the system determining constant $\mathbf{x}$ by generalised least squares, and that predicting the realisation of random $\mathbf{x}$ under the assumption of a diffuse prior. It is further assumed that

$$\begin{aligned}\mathcal{C}\left\{\mathbf{x}_{1|0}, \mathbf{u}_k\right\} &= 0,\ k = 1, 2, \cdots, n-1,\\ \mathcal{C}\left\{\mathbf{x}_{1|0}, \boldsymbol{\varepsilon}_k\right\} &= 0,\ k = 1, 2, \cdots, n.\end{aligned}$$

Here the problem data is of two kinds. The initial vector $\mathbf{x}_{1|0}$ has something of an existential role with the entire computation being conditioned on it. The data of principal interest are the observations $\mathbf{y}_k$, $k = 1, 2, \cdots, n$, in the sense that replications of the system observations are considered to generate new sets of the $\mathbf{y}_k$ for given system initialisation $\mathbf{x}_{1|0}$.

Let the input data be written $\mathcal{Y}_k = \left\{\mathbf{x}_{1|0}, \mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_k\right\}$, $k = 1, 2, \cdots, n$. Then the linear, minimum variance prediction of $\mathbf{x}_i$ conditional on $\mathcal{Y}_k$ is given by $\mathcal{E}\left\{\mathbf{x}_i | \mathcal{Y}_k\right\} = \mathbf{x}_{i|k}$ with corresponding covariance $\mathcal{V}\left\{\mathbf{x}_i - \mathbf{x}_{i|k}\right\} = S_{i|k}$. The prediction problem is typically formulated in two parts:

1. predict $\mathbf{x}_{k|k}$, $k = 1, 2, \cdots, n$ recursively (*the filtering problem*); and

2. predict $\mathbf{x}_{k|n}$, $k = 1, 2, \cdots, n$, the dependence of the state predictions on all of the data (*the smoothing problem*).

The filtering problem can be considered in the context of revising a system state estimate as additional information becomes available, and here it has been spectacularly successful in applications [2]. An analysis is given in the next two subsections. The smoothing problem can be formulated as a generalised least squares problem. This follows on noting that the system dynamics (1.4.1) can be written - here quantities with subscript c are formed by amalgamating the component elements -

$$L_c\mathbf{x} - \begin{bmatrix} \mathbf{x}_{1|0} \\ 0_c \end{bmatrix} = \mathbf{u}_c,$$

where

$$L_c = \begin{bmatrix} I & & & \\ -X_1 & I & & \\ & \cdots & \cdots & \\ & & -X_{n-1} & I \end{bmatrix}, \ \mathbf{u}_c = \begin{bmatrix} \mathbf{u}_0 \\ \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_{n-1} \end{bmatrix}.$$

In this case the minimum variance estimation problem corresponds to the data

$$\mathbf{y}_c - H_c\mathbf{x}_c = \boldsymbol{\varepsilon}_c \backsim N(0_c, V_c), \ L_c\mathbf{x}_c - \begin{bmatrix} \mathbf{x}_{1|0} \\ 0_c \end{bmatrix} = \mathbf{u}_c \backsim N(0, R_c),$$

where the block diagonal matrices H_c, R_c, V_c are given by

$$H_c = \begin{bmatrix} H_1 & & \\ & \cdots & \\ & & H_n \end{bmatrix}, \ V_c = \begin{bmatrix} V_1 & & \\ & \cdots & \\ & & V_n \end{bmatrix},$$

$$R_c = \begin{bmatrix} S_{1|0} & & & \\ & R_1 & & \\ & & \cdots & \\ & & & R_{n-1} \end{bmatrix}.$$

This can be written in the form of (1.3.23) by introducing the new variable $\mathbf{t}_c = L_c\mathbf{x}_c - \begin{bmatrix} \mathbf{x}_{1|0} \\ 0_c \end{bmatrix} \backsim N(0, R_c)$ satisfying $C\{\mathbf{t}_c, \boldsymbol{\varepsilon}_c\} = 0$. This leads to the least squares problem (compare (1.3.26) and (1.3.27))

$$\min_{\mathbf{t}} \left\{ \mathbf{r}_1^T V_c^{-1} \mathbf{r}_1 + \mathbf{r}_2^T R_c^{-1} \mathbf{r}_2 \right\}$$

where

$$\begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} = \begin{bmatrix} \widetilde{H}_c \\ I \end{bmatrix} \mathbf{t}_c - \begin{bmatrix} \mathbf{y}_c - \widetilde{H}_c \begin{bmatrix} \mathbf{x}_{1|0} \\ 0_c \end{bmatrix} \\ 0_c \end{bmatrix}, \ \widetilde{H}_c = H_c L_c^{-1}.$$

Remark 1.4.2 *Note that the predicted realisation is conditional on the value selected for* $\mathbf{x}_{1|0}$*. An important special case corresponds to the case of no prior information on* $\mathbf{x}_1$*. This can be treated by letting* $S_{1|0} \to \infty I$ *in similar fashion to the treatment of mixed models. This serves to remove the equations corresponding to the initial conditions from the objective function.*

In terms of the untransformed variables this becomes

$$\min_{\mathbf{x}} \left\{ \mathbf{r}_1^T V_c^{-1} \mathbf{r}_1 + \mathbf{r}_2^T R_c^{-1} \mathbf{r}_2 \right\} \tag{1.4.3}$$

where

$$\begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} = \begin{bmatrix} H_c \\ L_c \end{bmatrix} \mathbf{x}_c - \begin{bmatrix} \mathbf{y}_c \\ \begin{bmatrix} \mathbf{x}_{1|0} \\ 0_c \end{bmatrix} \end{bmatrix}. \tag{1.4.4}$$

This is a result originally presented in the dynamical systems context by [26].

1.4.1 The filtering problem

The recursive calculation of $\mathbf{x}_{k+1|k+1}$ given $\mathbf{x}_{k|k}$ and the new data from the observation equation is an exercise in the use of the projection theorem . First the best estimate of $\mathbf{x}_{k+1}$ given the past is obtained from the dynamics equation:

$$\mathbf{x}_{k+1|k} = X_k \mathbf{x}_{k|k}.$$

The interesting component in the new observation, the *innovation* , is obtained using (1.3.34) which gives

$$\begin{aligned} \widetilde{\mathbf{y}}_{k+1} &= \mathbf{y}_{k+1} - H_{k+1}\mathbf{x}_{k+1|k}, \\ &= H_{k+1}\left(\mathbf{x}_{k+1} - \mathbf{x}_{k+1|k}\right) + \boldsymbol{\varepsilon}_{k+1}. \end{aligned} \tag{1.4.5}$$

If this is plugged into (1.3.33) the result is

$$\mathbf{x}_{k+1|k+1} = \mathbf{x}_{k+1|k} + \mathcal{C}\left\{\mathbf{x}_{k+1}, \widetilde{\mathbf{y}}_{k+1}\right\} \mathcal{V}\left\{\widetilde{\mathbf{y}}_{k+1}\right\}^{-1} \widetilde{\mathbf{y}}_{k+1}, \tag{1.4.6}$$

where, using the orthogonality result (1.3.32),

$$\begin{aligned} \mathcal{C}\left\{\mathbf{x}_{k+1}, \widetilde{\mathbf{y}}_{k+1}\right\} &= \mathcal{C}\left\{\mathbf{x}_{k+1}, \mathbf{x}_{k+1} - \mathbf{x}_{k+1|k}\right\} H_{k+1}^T, \\ &= \mathcal{C}\left\{\mathbf{x}_{k+1} - \mathbf{x}_{k+1|k}, \mathbf{x}_{k+1} - \mathbf{x}_{k+1|k}\right\} H_{k+1}^T, \\ &= S_{k+1|k} H_{k+1}^T, \end{aligned} \tag{1.4.7}$$

and

$$\mathcal{V}\left\{\widetilde{\mathbf{y}}_{k+1}\right\} = H_{k+1} S_{k+1|k} H_{k+1}^T + V_{k+1}. \tag{1.4.8}$$

These calculations require a knowledge of $S_{k+1|k}$ but this also can be recurred using

$$\begin{aligned} S_{k+1|k} &= \mathcal{V}\left\{\mathbf{x}_{k+1} - \mathbf{x}_{k+1|k}\right\}, \\ &= \mathcal{V}\left\{X_k\left(\mathbf{x}_k - \mathbf{x}_{k|k}\right) + \mathbf{u}_k\right\}, \\ &= X_k S_{k|k} X_k^T + R_k. \end{aligned} \tag{1.4.9}$$

This, in turn, requires $S_{k|k}$ which is given by

$$\begin{aligned} S_{k|k} &= \mathcal{V}\left\{\mathbf{x}_k - \mathbf{x}_{k|k-1} - \mathcal{C}\left\{\mathbf{x}_k, \widetilde{\mathbf{y}}_k\right\}\mathcal{V}\left\{\widetilde{\mathbf{y}}_k\right\}^{-1}\widetilde{\mathbf{y}}_k\right\}, \\ &= S_{k|k-1} + \mathcal{C}\left\{\mathbf{x}_k, \widetilde{\mathbf{y}}_k\right\}\mathcal{V}\left\{\mathbf{y}_k\right\}^{-1}\mathcal{C}\left\{\widetilde{\mathbf{y}}_k, \mathbf{x}_k\right\} \\ &\quad - \mathcal{C}\left\{\mathbf{x}_k - \mathbf{x}_{k|k-1}, \mathbf{y}_k\right\}\mathcal{V}\left\{\widetilde{\mathbf{y}}_k\right\}^{-1}\mathcal{C}\left\{\widetilde{\mathbf{y}}_k, \mathbf{x}_k\right\} \\ &\quad - \mathcal{C}\left\{\mathbf{x}_k, \widetilde{\mathbf{y}}_k\right\}\mathcal{V}\left\{\mathbf{y}_k\right\}^{-1}\mathcal{C}\left\{\widetilde{\mathbf{y}}_k, \mathbf{x}_k - \mathbf{x}_{k|k-1}\right\}, \\ &= S_{k|k-1} - \mathcal{C}\left\{\mathbf{x}_k, \widetilde{\mathbf{y}}_k\right\}\mathcal{V}\left\{\widetilde{\mathbf{y}}_k\right\}^{-1}\mathcal{C}\left\{\widetilde{\mathbf{y}}_k, \mathbf{x}_k\right\}. \end{aligned}$$

This is computed readily by expanding out the component terms and then applying the above results. We obtain

$$S_{k|k} = S_{k|k-1} - S_{k|k-1}H_k^T\left\{H_k S_{k|k-1}H_k^T + V_k\right\}^{-1}H_k S_{k|k-1}. \tag{1.4.10}$$

The minus sign in this equation emphasises that $S_{k|k}$ is less positive definite (smaller!) than $S_{k|k-1}$ showing that the information on $\mathbf{x}_k$ has increased.

The filter equations are readily extended to certain cases of correlated data . The simplest one corresponds to additional correlation between $\mathbf{u}_k$ and $\boldsymbol{\varepsilon}_{k+1}$ only. Here the innovation $\widetilde{\mathbf{y}}_{k+1}$ is orthogonal to the Hilbert space $\mathcal{H}\{\mathcal{Y}_k\}$ containing $\mathbf{x}_{k+1|k}$. Thus (1.4.6) remains valid. Evaluating the terms now gives

$$\begin{aligned} \mathcal{C}\left\{\mathbf{x}_{k+1}, \widetilde{\mathbf{y}}_{k+1}\right\} &= \mathcal{C}\left\{\mathbf{x}_{k+1} - \mathbf{x}_{k+1|k}, \widetilde{\mathbf{y}}_{k+1}\right\}, \\ &= \mathcal{C}\left\{\mathbf{x}_{k+1} - \mathbf{x}_{k+1|k}, H_{k+1}\left(\mathbf{x}_{k+1} - \mathbf{x}_{k+1|k}\right) + \boldsymbol{\varepsilon}_{k+1}\right\}, \\ &= S_{k+1|k}H_{k+1}^T + \mathcal{C}\left\{\mathbf{x}_{k+1}, \boldsymbol{\varepsilon}_{k+1}\right\}, \\ &= S_{k+1|k}H_{k+1}^T + \mathcal{C}\left\{\mathbf{u}_k, \boldsymbol{\varepsilon}_{k+1}\right\}; \\ \mathcal{V}\left\{\widetilde{\mathbf{y}}_{k+1}\right\} &= H_{k+1}S_{k+1|k}H_{k+1}^T + H_{k+1}\mathcal{C}\left\{\mathbf{u}_k, \boldsymbol{\varepsilon}_{k+1}\right\} + \\ &\quad \mathcal{C}\left\{\mathbf{u}_k, \boldsymbol{\varepsilon}_{k+1}\right\}^T H_{k+1}^T + V_{k+1}. \end{aligned}$$

Also, while the basic form of the recurrence is preserved, the computation of the covariance term $S_{k|k}$ must be modified to take account of the change in the correlation calculations which follow from the weakened assumptions. Discussion of the case when the only new correlation is between $\mathbf{u}_{k-1}$ and $\boldsymbol{\varepsilon}_{k+1}$ is given in [2] and [19]. Here the new information in $\mathbf{y}_{k+1}$ is used to update estimates of both $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$.

Exercise 1.4.1 *Derive the filter equations in the case when there is correlation is between* $\mathbf{u}_{k-1}$ *and* $\boldsymbol{\varepsilon}_{k+1}$.

1.4.2 The smoothing problem

Here the question considered is given the output of the filter $\mathbf{x}_{i|i}$, $i = 1, 2, \cdots, n$, find the values $\mathbf{x}_{i|n}$ showing the dependence on all the data corresponding to the solution of the generalised least squares problem (1.4.3), (1.4.4). This can be done using a neat argument due to [3]. Note that $\mathcal{C}\{\mathbf{x}_{i+1}, \boldsymbol{\varepsilon}_{i+1}\} = X_i\mathcal{C}\{\mathbf{x}_i, \boldsymbol{\varepsilon}_{i+1}\} = 0$, and that, as $\mathbf{x}_{i+1|i} \in \mathcal{H}\{\mathcal{Y}_i\}$,

$$\begin{aligned} \mathbf{y}_{i+1} &= H_{i+1}\left(\mathbf{x}_{i+1} - \mathbf{x}_{i+1|i}\right) + H_{i+1}\mathbf{x}_{i+1|i} + \boldsymbol{\varepsilon}_{i+1}, \\ \Rightarrow \mathcal{Y}_{i+1} &\subset \mathcal{H}\left\{\mathcal{Y}_i, \mathbf{x}_{i+1} - \mathbf{x}_{i+1|i}, \boldsymbol{\varepsilon}_{i+1}\right\}. \end{aligned}$$

It then follows from the dynamics and observation equations that

$$\mathcal{Y}_n \subset \mathcal{H}\{\mathcal{U}_i\}$$

where

$$\mathcal{U}_i = \mathcal{Y}_i \cup \left\{\mathbf{x}_{i+1} - \mathbf{x}_{i+1|i}\right\} \cup \{\boldsymbol{\varepsilon}_{i+1}, \cdots \boldsymbol{\varepsilon}_n, \mathbf{u}_{i+1}, \cdots, \mathbf{u}_{n-1}\},$$

and the component sets of vectors are independent. As $\mathbf{x}_i$ is independent of $\{\boldsymbol{\varepsilon}_{i+1}, \cdots \boldsymbol{\varepsilon}_n, \mathbf{u}_{i+1}, \cdots, \mathbf{u}_{n-1}\}$, it follows that the projections of $\mathbf{x}_i$ into the corresponding orthogonal subspaces are related by

$$\begin{aligned} \mathcal{E}\{\mathbf{x}_i|\mathcal{U}_i\} &= \mathcal{E}\{\mathbf{x}_i|\mathcal{Y}_i\} + \mathcal{E}\left\{\mathbf{x}_i|\mathbf{x}_{i+1} - \mathbf{x}_{i+1|i}\right\}, \\ &= \mathbf{x}_{i|i} + \mathcal{C}\left\{\mathbf{x}_i, \mathbf{x}_{i+1} - \mathbf{x}_{i+1|i}\right\} S_{i+1|i}^{-1}\left(\mathbf{x}_{i+1} - \mathbf{x}_{i+1|i}\right), \\ &= \mathbf{x}_{i|i} + \mathcal{C}\left\{\mathbf{x}_i, X_i\left(\mathbf{x}_i - \mathbf{x}_{i|i}\right) + \mathbf{u}_i\right\} S_{i+1|i}^{-1}\left(\mathbf{x}_{i+1} - \mathbf{x}_{i+1|i}\right), \\ &= \mathbf{x}_{i|i} + S_{i|i}X_i^T S_{i+1|i}^{-1}\left(\mathbf{x}_{i+1} - \mathbf{x}_{i+1|i}\right) \end{aligned}$$

as $\mathcal{C}\left\{\mathbf{x}_{i|i}, \mathbf{x}_i - \mathbf{x}_{i|i}\right\} = 0$ by the best approximation property (1.3.32), and $\mathcal{C}\{\mathbf{x}_i, \mathbf{u}_i\} = 0$ by independence. The only element on the right hand side that is not already in the subspace $\mathcal{H}\{\mathcal{Y}_n\}$ is that involving $\mathbf{x}_{i+1}$, but this is contained in $\mathcal{H}\{\mathcal{U}_i\}$ as a consequence of the above identity. Projecting both sides into $\mathcal{H}\{\mathcal{Y}_n\}$ gives

$$\mathbf{x}_{i|n} = \mathbf{x}_{i|i} + A_i\left(\mathbf{x}_{i+1|n} - \mathbf{x}_{i+1|i}\right) \tag{1.4.11}$$

where the *interpolation gain* A_i is given by

$$A_i = S_{i|i}X_i^T S_{i+1|i}^{-1}. \tag{1.4.12}$$

The variance $S_{i|n}$ of the smoothed predictor is given by

$$\mathcal{V}\left\{\mathbf{x}_i - \mathbf{x}_{i|n}\right\} = \mathcal{C}\left\{\mathbf{x}_{i|i} - \mathbf{x}_i, \mathbf{x}_{i|n} - \mathbf{x}_i\right\} + A_i\mathcal{C}\left\{\mathbf{x}_{i+1|n} - \mathbf{x}_{i+1|i}, \mathbf{x}_{i|n} - \mathbf{x}_i\right\}.$$

The first term can be written

$$\mathcal{C}\left\{\mathbf{x}_{i|i}-\mathbf{x}_i,\mathbf{x}_{i|n}-\mathbf{x}_i\right\}=S_{i|i}+\mathcal{C}\left\{\mathbf{x}_{i|i}-\mathbf{x}_i,\mathbf{x}_{i|n}-\mathbf{x}_{i|i}\right\}.$$

The second gives

$$\begin{aligned}&A_i\mathcal{C}\left\{\mathbf{x}_{i+1|n}-\mathbf{x}_{i+1|i},\mathbf{x}_{i|n}-\mathbf{x}_i\right\}=\\&A_i\mathcal{C}\left\{\mathbf{x}_{i+1|n}-\mathbf{x}_{i+1|i},\mathbf{x}_{i+1|n}-\mathbf{x}_{i+1|i}\right\}A_i^T+\mathcal{C}\left\{\mathbf{x}_{i|n}-\mathbf{x}_{i|i},\mathbf{x}_{i|i}-\mathbf{x}_i\right\}.\end{aligned}$$

We have

$$\begin{aligned}&A_i\mathcal{C}\left\{\mathbf{x}_{i+1|n}-\mathbf{x}_{i+1|i},\mathbf{x}_{i+1|n}-\mathbf{x}_{i+1|i}\right\}A_i^T\\&=A_i\left\{\begin{array}{c}S_{i+1|n}+S_{i+1|i}-\mathcal{C}\left\{\mathbf{x}_{i+1|n}-\mathbf{x}_{i+1},\mathbf{x}_{i+1|i}-\mathbf{x}_{i+1}\right\}-\\\mathcal{C}\left\{\mathbf{x}_{i+1|i}-\mathbf{x}_{i+1},\mathbf{x}_{i+1|n}-\mathbf{x}_{i+1}\right\}\end{array}\right\}A_i^T\\&=A_i\left\{S_{i+1|n}-S_{i+1|i}\right\}A_i^T-\\&\mathcal{C}\left\{\mathbf{x}_{i|n}-\mathbf{x}_{i|i},\mathbf{x}_{i+1|i}-\mathbf{x}_{i+1}\right\}A_i^T-A_i\mathcal{C}\left\{\mathbf{x}_{i+1|i}-\mathbf{x}_{i+1},\mathbf{x}_{i|n}-\mathbf{x}_{i|i}\right\}.\end{aligned}$$

At this point we have that

$$S_{i|n}=S_{i|i}+A_i\left\{S_{i+1|n}-S_{i+1|i}\right\}A_i^T+W$$

where

$$W=U+U^T,$$

and

$$\begin{aligned}U&=\mathcal{C}\left\{\mathbf{x}_{i|n}-\mathbf{x}_{i|i},\mathbf{x}_{i|i}-\mathbf{x}_i\right\}-\mathcal{C}\left\{\mathbf{x}_{i|n}-\mathbf{x}_{i|i},\mathbf{x}_{i+1|i}-\mathbf{x}_{i+1}\right\}A_i^T,\\&=\mathcal{C}\left\{\mathbf{x}_{i|n}-\mathbf{x}_{i|i},\mathbf{x}_{i|i}-\mathbf{x}_i-A_i\left\{\mathbf{x}_{i+1|i}-\mathbf{x}_{i+1}\right\}\right\},\\&=0.\end{aligned}$$

This follows because the first term in the covariance is $\mathbf{x}_{i|n}-\mathbf{x}_{i|i}\subset\mathcal{H}\left\{\mathcal{U}_i\right\}$, while the second is orthogonal to $\mathcal{H}\left\{\mathcal{U}_i\right\}$ as $A_i\left\{\mathbf{x}_{i+1}-\mathbf{x}_{i+1|i}\right\}$ is just the projection of $\mathbf{x}_i-\mathbf{x}_{i|i}$ onto $\mathcal{H}\left\{\mathbf{x}_{i+1}-\mathbf{x}_{i+1|i}\right\}$. Thus the final result is

$$S_{i|n}=S_{i|i}+A_i\left\{S_{i+1|n}-S_{i+1|i}\right\}A_i^T. \tag{1.4.13}$$

In certain circumstances there can be a need to interpolate the smoother output at intermediate time points t, $t_i<t<t_{i+1}$. Here it is assumed that the intermediate dynamics is given by

$$\mathbf{x}(t)=X(t,t_i)\,\mathbf{x}_i+\mathbf{u}(t,t_i),$$

where the matrix of the dynamics equation satisfies

$$X_i=X(t_{i+1},t)\,X(t,t_i),$$

and the stochastic term possesses the covariance property

$$\mathcal{C}\left\{\mathbf{u}\left(t,t_i\right),\mathbf{u}_i\right\}=\mathcal{V}\left\{\mathbf{u}\left(t,t_i\right)\right\}X\left(t_{i+1},t\right)^T,$$

where $\mathcal{V}\left\{\mathbf{u}\left(t,t_i\right)\right\}\rightarrow R_i$, $t\rightarrow t_{i+1}$, $\mathcal{V}\left\{\mathbf{u}\left(t,t_i\right)\right\}\rightarrow 0$, $t\rightarrow t_i$. It will turn out that this property is appropriate for an important class of random walk processes. The argument is similar to that employed above. We have

$$\begin{aligned}\mathcal{E}\left\{\mathbf{x}\left(t\right)|\mathcal{U}_i\right\}&=\mathcal{E}\left\{X\left(t,t_i\right)\mathbf{x}_i+\mathbf{u}\left(t,t_i\right)|\mathcal{U}_i\right\},\\&=\mathcal{E}\left\{X\left(t,t_i\right)\mathbf{x}_i+\mathbf{u}\left(t,t_i\right)|\mathcal{Y}_i\right\}\\&+\mathcal{E}\left\{X\left(t,t_i\right)\mathbf{x}_i+\mathbf{u}\left(t,t_i\right)|\mathbf{x}_{i+1}-\mathbf{x}_{i+1|i}\right\},\\&=X\left(t,t_i\right)\mathbf{x}_{i|i}+\left(\begin{array}{c}X\left(t,t_i\right)S_{i|i}X_i^T+\\\mathcal{C}\left\{\mathbf{u}\left(t,t_i\right),\mathbf{x}_{i+1}-\mathbf{x}_{i+1|i}\right\}\end{array}\right)S_{i+1|i}^{-1}\left(\mathbf{x}_{i+1}-\mathbf{x}_{i+1|i}\right),\\&=X\left(t,t_i\right)\mathbf{x}_{i|i}+A\left(t,t_i\right)\left(\mathbf{x}_{i+1}-\mathbf{x}_{i+1|i}\right),\end{aligned}$$

where

$$A\left(t,t_i\right)=\left(X\left(t,t_i\right)S_{i|i}X_i^T+\mathcal{V}\left\{\mathbf{u}\left(t,t_i\right)\right\}X\left(t_{i+1},t\right)^T\right)S_{i+1|i}^{-1}. \tag{1.4.14}$$

Projecting into $\mathcal{Y}_n$ gives

$$\mathcal{E}\left\{\mathbf{x}\left(t\right)|\mathcal{Y}_n\right\}=X\left(t,t_i\right)\mathbf{x}_{i|i}+A\left(t,t_i\right)\left(\mathbf{x}_{i+1|n}-\mathbf{x}_{i+1|i}\right). \tag{1.4.15}$$

The corresponding formula for the variance is

$$\begin{aligned}\mathcal{V}\left\{\mathbf{x}\left(t\right)-\mathcal{E}\left\{\mathbf{x}\left(t\right)|\mathcal{Y}_n\right\}\right\}=\mathcal{V}\left\{\mathbf{u}\left(t,t_i\right)\right\}+X\left(t,t_i\right)S_{i|i}X\left(t,t_i\right)^T+\\A\left(t,t_i\right)\left(S_{i+1|n}-S_{i+1|i}\right)A\left(t,t_i\right)^T.\end{aligned}$$

The argument used to develop the smoothing algorithm extends to the case of correlation between $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_{i+1}$, but at a cost of additional complexity. Again it follows that $\mathcal{Y}_n\subset\mathcal{H}\left\{\mathcal{U}_i\right\}$ where it is convenient to write the basis set as

$$\mathcal{U}_i=\mathcal{Y}_i\cup\left\{\mathbf{x}_{i+1}-\mathbf{x}_{i+1|i}\right\}\cup\left\{\boldsymbol{\varepsilon}_{i+1}\right\}\cup\left\{\boldsymbol{\varepsilon}_{i+2},\boldsymbol{\varepsilon}_{i+3},\cdots,\boldsymbol{\varepsilon}_n,\mathbf{u}_{i+1},\mathbf{u}_{i+2},\cdots,\mathbf{u}_{n-1}\right\}$$

and to express the contribution of the interesting part as

$$\mathcal{H}\left\{\left\{\mathbf{x}_{i+1}-\mathbf{x}_{i+1|i}\right\}\cup\left\{\boldsymbol{\varepsilon}_{i+1}\right\}\right\}=\mathcal{H}\left\{\left\{\mathbf{x}_{i+1}-\mathbf{x}_{i+1|i}\right\}\cup\left\{\widetilde{\boldsymbol{\varepsilon}}_{i+1}\right\}\right\},$$

where

$$\widetilde{\boldsymbol{\varepsilon}}_{i+1}=\boldsymbol{\varepsilon}_{i+1}-P_i\left(\mathbf{x}_{i+1}-\mathbf{x}_{i+1|i}\right),$$

and

$$P_i=\mathcal{C}\left\{\boldsymbol{\varepsilon}_{i+1},\mathbf{x}_{i+1}-\mathbf{x}_{i+1|i}\right\}S_{i+1|i}^{-1}=\mathcal{C}\left\{\boldsymbol{\varepsilon}_{i+1},\mathbf{u}_i\right\}S_{i+1|i}^{-1},$$

in order to provide a decomposition of $\mathcal{H}\{\mathcal{U}_i\}$ into orthogonal subspaces. Now

$$\mathcal{E}\{\mathbf{x}_i|\mathcal{U}_i\} = \mathbf{x}_{i|i} + \mathcal{E}\left\{\mathbf{x}_i|\mathbf{x}_{i+1} - \mathbf{x}_{i+1|i}\right\} + \mathcal{E}\{\mathbf{x}_i|\widetilde{\boldsymbol{\varepsilon}}_{i+1}\}.$$

Evaluating the new term gives

$$\begin{aligned}\mathcal{E}\{\mathbf{x}_i|\widetilde{\boldsymbol{\varepsilon}}_{i+1}\} &= -\mathcal{C}\left\{\mathbf{x}_i, \mathbf{x}_{i+1} - \mathbf{x}_{i+1|i}\right\} P_i^T \mathcal{V}\{\widetilde{\boldsymbol{\varepsilon}}_{i+1}\}^{-1} \widetilde{\boldsymbol{\varepsilon}}_{i+1}, \\ &= -S_{i|i} X_i^T P_i^T \mathcal{V}\{\widetilde{\boldsymbol{\varepsilon}}_{i+1}\}^{-1} \widetilde{\boldsymbol{\varepsilon}}_{i+1}, \\ &= -A_i \mathcal{C}\{\mathbf{u}_i, \boldsymbol{\varepsilon}_{i+1}\} \mathcal{V}\{\widetilde{\boldsymbol{\varepsilon}}_{i+1}\}^{-1} \left(\boldsymbol{\varepsilon}_{i+1} - P_i\left(\mathbf{x}_{i+1} - \mathbf{x}_{i+1|i}\right)\right),\end{aligned}$$

where, as in the development of the filter equations, covariances must be modified to take account of the correlations and

$$\mathcal{V}\{\widetilde{\boldsymbol{\varepsilon}}_{i+1}\} = V_{i+1} - \mathcal{C}\{\boldsymbol{\varepsilon}_{i+1}, \mathbf{u}_i\} S_{i+1|i}^{-1} \mathcal{C}\{\mathbf{u}_i, \boldsymbol{\varepsilon}_{i+1}\}.$$

Collecting terms gives

$$\begin{aligned}\mathcal{E}\{\mathbf{x}_i|\mathcal{U}_i\} = \mathbf{x}_{i|i} &+ A_i\left(I + \mathcal{C}\{\mathbf{u}_i, \boldsymbol{\varepsilon}_{i+1}\} \mathcal{V}\{\widetilde{\boldsymbol{\varepsilon}}_{i+1}\}^{-1} \mathcal{C}\{\boldsymbol{\varepsilon}_{i+1}, \mathbf{u}_i\} S_{i+1|i}^{-1}\right)\begin{pmatrix}\mathbf{x}_{i+1} \\ -\mathbf{x}_{i+1|i}\end{pmatrix} \\ &- A_i \mathcal{C}\{\mathbf{u}_i, \boldsymbol{\varepsilon}_{i+1}\} \mathcal{V}\{\widetilde{\boldsymbol{\varepsilon}}_{i+1}\}^{-1} (\mathbf{y}_{i+1} - H_{i+1}\mathbf{x}_{i+1}).\end{aligned}$$

Projecting from $\mathcal{H}\{\mathcal{U}_i\}$ into $\mathcal{H}\{\mathcal{Y}_n\}$ gives

$$\begin{aligned}\mathbf{x}_{i|n} = \mathbf{x}_{i|i} &+ A_i\left(I + \mathcal{C}\{\mathbf{u}_i, \boldsymbol{\varepsilon}_{i+1}\} \mathcal{V}\{\widetilde{\boldsymbol{\varepsilon}}_{i+1}\}^{-1} \mathcal{C}\{\boldsymbol{\varepsilon}_{i+1}, \mathbf{u}_i\} S_{i+1|i}^{-1}\right)\begin{pmatrix}\mathbf{x}_{i+1|n} \\ -\mathbf{x}_{i+1|i}\end{pmatrix} \\ &- A_i \mathcal{C}\{\mathbf{u}_i, \boldsymbol{\varepsilon}_{i+1}\} \mathcal{V}\{\widetilde{\boldsymbol{\varepsilon}}_{i+1}\}^{-1} \left(\mathbf{y}_{i+1} - H_{i+1}\mathbf{x}_{i+1|n}\right).\end{aligned}$$

Exercise 1.4.2 *Formulate the orthogonal subspaces needed to develop the smoothing recurrence in the case that $\mathbf{u}_{k-1}$ is orthogonal to $\boldsymbol{\varepsilon}_{k+1}$. What is the form of this smoothing recurrence.*

1.5 Mean models

A standard example leading to a constrained least squares problem is the mean model arising in experimental design . The problem situation here is expressed by a multiway table , and the data to be modelled are the sets of outcomes recorded in the cells of the table. For simplicity consider a two-way table consisting of n_p rows and n_q columns. Let the number of outcomes

recorded in cell ij be n_{ij}, and assume a description of these outcomes is provided by the mean model

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk},\ i = 1, 2, \cdots, n_p,\ j = 1, 2, \cdots, n_q,\ k = 1, 2, \cdots, n_{ij}, \quad (1.5.1)$$

where $\varepsilon_{ijk} \backsim N(0, \sigma^2)$, and $\mathcal{E}\{\varepsilon_{ijk}\varepsilon_{pqr}\} = \delta_{ip}\delta_{jq}\delta_{kr}$. The object is to fit the cell means by a linear model consisting of a mean term μ plus contrasts α, β:

$$\mu_{ij} = \mu + \alpha_i + \beta_j, \quad (1.5.2)$$

to the observed outcomes y_{ijk} by least squares. The linear model is given by a two dimensional array which can be written in matrix form

$$\mu_{**} = \mu \mathbf{e}^{(p)}\mathbf{e}^{(q)T} + \boldsymbol{\alpha}\mathbf{e}^{(q)T} + \mathbf{e}^{(p)}\boldsymbol{\beta}^T.$$

Note that the problem of minimising

$$\sum_{i=1}^{n_p}\sum_{j=1}^{n_q}\sum_{k=1}^{n_{ij}}(y_{ijk} - \mu_{ij})^2$$

is equivalent to the problem of minimizing the fit to the cell means

$$\sum_{i=1}^{n_p}\sum_{j=1}^{n_q} n_{ij}\left(\overline{y}_{ij} - \mu_{ij}\right)^2$$

where the cell means are

$$\overline{y}_{ij} = \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} y_{ijk},$$

as the two sums differ by a term depending only on the data. The matrix representation of the array of values $\overline{y}_{ij}$ will be written $\overline{y}_{**}$. Also there is inherent indeterminacy in the formulation (1.5.2) of the linear model as

$$\mu_{ij} = \mu + \alpha_i + \beta_j = \widetilde{\mu} + \widetilde{\alpha}_i + \widetilde{\beta}_j,$$

where $\widetilde{\alpha}_i = \alpha_i + \Delta_1,\ i = 1, 2, \cdots, n_p,\ \widetilde{\beta}_j = \beta_j + \Delta_2,\ j = 1, 2, \cdots, n_q,\ \widetilde{\mu} = \mu - \Delta_1 - \Delta_2$. It follows that the problem is not well posed without the addition of extra constraints. Typically these are taken to have the form

$$\mathbf{e}^{(p)T}\boldsymbol{\alpha} = \sum_{i=1}^{n_p}\alpha_i = 0,\ \mathbf{e}^{(q)T}\boldsymbol{\beta} = \sum_{j=1}^{n_q}\beta_j = 0.$$

The estimation problem is said to be *balanced* if n_{ij} is independent of i, j, otherwise *unbalanced* . The key feature of the balanced case is that the

solution of the least squares problem is a straight forward computation. The first step is to use each of the equality constraints to remove a degree of freedom from the solution. Let orthogonal matrices be defined by

$$Q_p^T \mathbf{e}^{(p)} = \sqrt{n_p}\mathbf{e}_1^{(p)},\ Q_q^T \mathbf{e}^{(q)} = \sqrt{n_q}\mathbf{e}_1^{(q)},$$

where

$$Q_p = \left[\begin{array}{cc} \frac{1}{\sqrt{n_p}}\mathbf{e}^{(p)} & Q_p^2 \end{array} \right],\ Q_q = \left[\begin{array}{cc} \frac{1}{\sqrt{n_q}}\mathbf{e}^{(q)} & Q_q^2 \end{array} \right].$$

For example, a suitable Aitken–Householder form for Q_p is

$$Q_p = -\left(I - 2\mathbf{w}\mathbf{w}^T\right),\ \mathbf{w} = \frac{\mathbf{e}^{(p)} + \sqrt{n_p}\mathbf{e}_1^{(p)}}{\sqrt{2\left(n_p + \sqrt{n_p}\right)}}.$$

Thus the imposed constraints are satisfied by setting

$$\boldsymbol{\alpha} = Q_p^2 \mathbf{x}_p,\ \mathbf{x}_p \in R^{n_p - 1},\ \boldsymbol{\beta} = Q_q^2 \mathbf{x}_q,\ \mathbf{x}_q \in R^{n_q - 1}.$$

The indeterminancy in the governing equations can now be removed by writing these in the block form

$$\mu_{*j} = \mu \mathbf{e}^{(p)} + Q_p^2 \mathbf{x}_p + \mathbf{e}^{(p)} \mathbf{e}_j^T Q_q^2 \mathbf{x}_q,\ j = 1, 2, \cdots, n_q.$$

or

$$\boldsymbol{\mu} = X \left[\begin{array}{c} \mu \\ \mathbf{x}_p \\ \mathbf{x}_q \end{array} \right],$$

where X is the corresponding design matrix,

$$X = \left[\begin{array}{ccc} \mathbf{e}^{(p)} & Q_p^2 & \mathbf{e}^{(p)} \mathbf{e}_1^T Q_q^2 \\ \vdots & \vdots & \vdots \\ \mathbf{e}^{(p)} & Q_p^2 & \mathbf{e}^{(p)} \mathbf{e}_{n_q}^T Q_q^2 \end{array} \right] \tag{1.5.3}$$

and $\boldsymbol{\mu}$, $\overline{\mathbf{y}}$ have the block components $\mu_{*j}, \overline{y}_{*j}$ $j = 1, 2, \cdots, n_q$. The normal equations are

$$X^T D X \left[\begin{array}{c} \mu \\ \mathbf{x}_p \\ \mathbf{x}_q \end{array} \right] = X^T D \overline{\mathbf{y}}$$

where $D = \text{diag}\{n_{ij}, i = 1, 2, \cdots, n_p, j = 1, 2, \cdots, n_q\}$. In the balanced case the relevant quantities are

$$X^T X = \begin{bmatrix} n_q \mathbf{e}^{(p)T}\mathbf{e}^{(p)} & 0 & 0 \\ 0 & n_q \left(Q_p^2\right)^T Q_p^2 & 0 \\ 0 & 0 & \sum_j \left(Q_q^2\right)_{j*}^T \mathbf{e}^{(p)T}\mathbf{e}^{(p)} \left(Q_q^2\right)_{j*} \end{bmatrix},$$
$$= \begin{bmatrix} n_p n_q & 0 & 0 \\ 0 & n_q I_{p-1} & 0 \\ 0 & 0 & n_p I_{q-1} \end{bmatrix},$$

and

$$X^T \overline{\mathbf{y}} = \begin{bmatrix} \sum_j \mathbf{e}^{(p)T}\overline{y}_{*j} \\ \left(Q_p^2\right)^T \sum_j \overline{y}_{*j} \\ \left(Q_q^2\right)^T \sum_j \mathbf{e}_j \mathbf{e}^{(p)T}\overline{y}_{*j} \end{bmatrix} = \begin{bmatrix} n_p n_q \overline{y}_{\bullet\bullet} \\ n_q \left(Q_p^2\right)^T \overline{y}_{*\bullet} \\ n_p \left(Q_q^2\right)^T \sum_j \mathbf{e}_j \overline{y}_{\bullet j} \end{bmatrix},$$

where $\overline{y}_{\bullet\bullet} = \frac{1}{n_p n_q}\sum_{i,j}\overline{y}_{ij}$, $\overline{y}_{*\bullet} = \frac{1}{n_q}\sum_j \overline{y}_{*j}$, $\overline{y}_{\bullet j} = \mathbf{e}^{(p)T}\overline{y}_{*j}$. The solution is

$$\begin{bmatrix} \mu \\ \mathbf{x}_p \\ \mathbf{x}_q \end{bmatrix} = \begin{bmatrix} \overline{y}_{\bullet\bullet} \\ \left(Q_p^2\right)^T \overline{y}_{*\bullet} \\ \left(Q_q^2\right)^T \sum_j \overline{y}_{\bullet j} \end{bmatrix}. \tag{1.5.4}$$

1.6 Appendix: Matrix identities and projections

The starting point for the discussion of the linear least squares problem included the assumption that the design matrix A had its full column rank p. This assumption is a reasonable reflection of the main application priorities but is by no means the full story. Following this assumption leads to the solution operator

$$A^+ = \left(A^T A\right)^{-1} A^T,$$

with the corresponding solution given by $\mathbf{x} = A^+\mathbf{b}$. This operator has the projection properties that $A^+A = I$ projects onto the domain of A, and $AA^+ = A\left(A^T A\right)^{-1} A^T$ projects onto the range of A. This provides the clue for extending the definition of A^+ to the case when the rank assumption is weakened. Assume now that $\text{rank}\, A = k < p$, and let $\sigma_1, \sigma_2, \cdots, \sigma_k$ be the nonzero eigenvalues of $A^T A$. Then

$$A^T A\mathbf{v}_j = \sigma_j \mathbf{v}_j \Rightarrow AA^T\left(A\mathbf{v}_j\right) = \sigma_j\left(A\mathbf{v}_j\right), j = 1, 2, \cdots, k,$$

so the σ_j are also the nonzero eigenvalues of AA^T with corresponding eigenvectors $\mathbf{u}_j$, where $\sigma_j \mathbf{u}_j = A\mathbf{v}_j, j = 1, 2, \cdots, k$. Then the singular value decomposition of A is

$$A = U_k \Sigma_k V_k^T \tag{1.6.1}$$

where $U_k \in R^k \to R^n$, $V_k \in R^k \to R^p$ are matrices whose columns are the normalised eigenvectors of AA^T and $A^T A$ respectively, and $\Sigma_k \in R^k \to R^k$ is the diagonal matrix of the nonzero eigenvalues. Conveniently defined using the singular value decomposition is the pseudo or generalised inverse of A:

$$A^+ = V_k \Sigma^{-1} U_k^T. \tag{1.6.2}$$

This definition extends that of the least squares solution operator given above and is justified by the associated projection properties

$$AA^+ = U_k U_k^T \tag{1.6.3}$$

giving an orthogonal projection onto the range of A, and

$$A^+ A = V_k V_k^T \tag{1.6.4}$$

giving an orthogonal projection onto that part of the domain of A which is not the pre-image of 0 under A. These results are usually expressed as the *Moore-Penrose* conditions

$$\begin{aligned} AA^+A = A, \quad \left(AA^+\right)^T = AA^+, \\ A^+AA^+ = A^+, \quad \left(A^+A\right)^T = A^+A. \end{aligned}$$

The generalised inverse provides the equipment needed to solve the least squares problem when A does not have full rank. The unique solution of minimum norm is given by

$$\mathbf{x} = A^+ \mathbf{b}. \tag{1.6.5}$$

Substituting in (1.1.1)gives

$$\mathbf{r} = -\left(I - AA^+\right)\mathbf{b}.$$

The necessary conditions requires $A^T\mathbf{r} = -\left(A^T - A^T AA^+\right)\mathbf{b} = 0$ and this follows directly from (1.6.3)

$$A^T = A^T AA^+.$$

To show that the generalised inverse solution is the solution of minimum norm consider $\mathbf{r} = A(\mathbf{x} + \mathbf{dx}) - \mathbf{b}$ where $\mathbf{x} + \mathbf{dx}$ is also a solution. The

necessary conditions give

$$\begin{aligned} A^T\mathbf{r} &= A^TA\mathbf{x} + A^TA\mathbf{dx} - A^T\mathbf{b}, \\ &\Rightarrow A^TA\mathbf{dx} = 0, \\ &\Rightarrow V_k^T\mathbf{dx} = 0. \end{aligned}$$

This shows that $\mathbf{x}$ is the orthogonal projection onto the set of solutions and hence has minimum norm.

The generalised least squares problem leads to more complicated algebra largely due to the critical scaling role of the inverse of the covariance matrix which provides the natural metric for projection operations. A good example is provided by the following derivation of (1.6.7) which was used in showing the equivalence of best linear predictor forms (1.3.24) and (1.3.25). The starting point is a formula for the inverse of the sum of a nonsingular matrix plus a rank one term and is easily found by a direct calculation. For example:

$$\left(V + \mathbf{x}\mathbf{x}^T\right)^{-1} = V^{-1} - \frac{V^{-1}\mathbf{x}\mathbf{x}^TV^{-1}}{1 + \mathbf{x}^TV^{-1}\mathbf{x}}.$$

This result is capable of significant generalisation.

Lemma 1.3 *Let $X : R^p \to R^q$, $V : R^q \to R^q$, and $W : R^p \to R^p$ have their maximum ranks, with V, W symmetric positive definite. Then the following identity is valid:*

$$\left(V + XWX^T\right)^{-1} = V^{-1} - V^{-1}X\left(X^TV^{-1}X + W^{-1}\right)^{-1}X^TV^{-1}.. \quad (1.6.6)$$

Proof. A straightforward calculation gives

$$\left(I + XWX^T\right)^{-1} = I - X\left(W^{-1} + XX^T\right)^{-1}X^T.$$

The desired result now follows by using this result to compute the inverse of

$$\left(V + XWX^T\right) = V^{1/2}\left(I + V^{-1/2}XWX^TV^{-1/2}\right)V^{1/2}.$$

■ ■ A corollary gives the desired equivalence:

Corollary 1.1 *following identity holds.*

$$WX^T\left(V + XWX^T\right)^{-1} = \left(W^{-1} + X^TV^{-1}X\right)^{-1}X^TV^{-1}. \quad (1.6.7)$$

Proof. This follows by expanding the lefthand side using (1.6.6). ■

Now let $\mathcal{X} \subseteq R^q$ be defined by

$$\mathcal{X} = \{\mathbf{x};\ \mathbf{x} = X\mathbf{z} \forall\, \mathbf{z} \in R^p\},$$

where $X : R^p \to R^q$ is assumed to have full rank $p \leq q$, and let $\mathcal{X}$ be endowed with the scalar product (and associated V^{-1} metric $\|\bullet\|_V$)

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_V = \mathbf{x}_1^T V^{-1} \mathbf{x}_2$$

where V is positive definite. The notation here reflects the role of the inverse of the covariance matrix in the normal distribution.

The projection of a point $\mathbf{a} \in R^q$ onto X in the V^{-1} metric is defined by

$$\min_{\mathbf{z}} \|\mathbf{a} - X\mathbf{z}\|_V^2$$

This corresponds to the problem (1.2.2) with the solution $\mathbf{z}$ given by (1.2.3)

$$\mathbf{z} = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} \mathbf{a}.$$

The corresponding projection matrix that realises the point $X\mathbf{z}$ is given by

$$X\mathbf{z} = P_{\mathcal{X}}^V \mathbf{a} = X \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} \mathbf{a}. \tag{1.6.8}$$

Let $Q_2 : R^{n-p} \to R^n$ provide a basis for the null space of X^T. Then

$$I - P_Q^V = I - VQ_2 \left(Q_2^T V Q_2\right)^{-1} Q_2^T \tag{1.6.9}$$

also gives a projection onto the range of X, and $I - P_Q^V = P_{\mathcal{X}}^V$. This is a consequence of the equalities

$$\left(I - P_Q^V\right) X = P_{\mathcal{X}}^V X,\ Q_2^T \left(I - P_Q^V\right) = Q_2^T P_{\mathcal{X}}^V = 0.$$

Note that P_Q^V does not depend on the precise form of Q_2. However, it will often be computed via an orthogonal factorization of X.

Remark 1.6.1 *Let $\mathcal{Y}$ be the orthogonal complement of $\mathcal{X}$. That is*

$$\mathcal{Y} = \left\{\mathbf{y};\ \mathbf{y} = Y\mathbf{b},\ Y^T X = 0,\ \mathbf{b} \in R^{q-p}\right\},$$

where $Y : R^{q-p} \to R^q$ has full rank $q - p$. Then, for $\mathbf{y} \in \mathcal{Y}$, $\mathbf{x} \in \mathcal{X}$

$$\langle V\mathbf{y}, \mathbf{x} \rangle_V = \mathbf{b}^T Y V V^{-1} X \mathbf{z} = 0.$$

Thus the V^{-1} orthogonal subspace is $V\mathcal{Y}$, and $P_{\mathcal{X}}^V V\mathcal{Y} = 0$.

Properties:

1. Let $P_{\mathcal{X}}$ be the orthogonal projection onto X corresponding to $V = I$ then

$$P_{\mathcal{X}}P_{\mathcal{X}}^V = P_{\mathcal{X}}^V,\ (I - P_{\mathcal{X}})\left(I - P_{\mathcal{X}}^V\right) = I - P_{\mathcal{X}}, \tag{1.6.10}$$
$$P_{\mathcal{X}}^V P_{\mathcal{X}} = P_{\mathcal{X}},\ \left(I - P_{\mathcal{X}}^V\right)(I - P_{\mathcal{X}}) = I - P_{\mathcal{X}}^V. \tag{1.6.11}$$

2. The transpose of $P_{\mathcal{X}}^V$ is also a projection matrix. In particular,

$$\left(P_{\mathcal{X}}^V\right)^T = V^{-1}P_{\mathcal{X}}^V V, \tag{1.6.12}$$
$$I - \left(P_{\mathcal{X}}^V\right)^T = V^{-1}\left(I - P_{\mathcal{X}}^V\right)V.$$

3. Let $\mathcal{Y}$ be the orthogonal complement of $\mathcal{X}$. Then

$$\left(I - \left(P_{\mathcal{X}}^V\right)^T\right)\mathcal{Y} = V^{-1}\left(I - P_{\mathcal{X}}^V\right)V\mathcal{Y} = \mathcal{Y}.$$

Also

$$\left(I - \left(P_{\mathcal{X}}^V\right)^T\right)V^{-1}\mathcal{X} = V^{-1}\left(I - P_{\mathcal{X}}^V\right)\mathcal{X} = 0,$$

so that, as $\mathcal{X}^T V^{-1} V\mathcal{Y} = 0$, it follows that $V^{-1}\mathcal{X}$ is the V orthogonal complement of $\mathcal{Y}$ in the V metric, that

$$P_{\mathcal{X}^\perp}^{V^{-1}} = \left(I - \left(P_{\mathcal{X}}^V\right)^T\right)$$

provides the corresponding projection matrix, and that

$$P_{\mathcal{X}^\perp}^{V^{-1}} + \left(P_{\mathcal{X}}^V\right)^T = I.$$

As an application of these operations consider the problem of attaching a limit to (1.6.7) as $W^{-1} \to 0$. We have

$$\begin{aligned}\lim_{W^{-1}\to 0}\left(V + XWX^T\right)^{-1} &= V^{-1} - V^{-1}X\left(X^TV^{-1}X\right)^{-1}X^TV^{-1},\\ &= V^{-1}\left(I - P_{\mathcal{X}}^V\right) = \left(I - \left(P_{\mathcal{X}}^V\right)^T\right)V^{-1},\\ &= P_{\mathcal{X}^\perp}^{V^{-1}}V^{-1}.\end{aligned}$$

This result can be interpreted in terms of orthogonal projectors using the g-pseudo inverse defined by the relations

$$AA^gA = A, \tag{1.6.13}$$
$$A^gAA^g = A^g. \tag{1.6.14}$$

Lemma 1.4

$$V^{-1}\left(I - P_{\mathcal{X}}^{V}\right) = \left\{(I - P_{\mathcal{X}})\, V\, (I - P_{\mathcal{X}})\right\}^{g}.$$

Proof. To verify the defining relations let

$$A = (I - P_{\mathcal{X}})\, V\, (I - P_{\mathcal{X}})\, V^{-1}\left(I - P_{\mathcal{X}}^{V}\right).$$

Then

$$\begin{aligned} A &= (I - P_{\mathcal{X}})\, V\, (I - P_{\mathcal{X}})\left(I - \left(P_{\mathcal{X}}^{V}\right)^{T}\right) V^{-1}, \\ &= (I - P_{\mathcal{X}})\, V\left(I - \left(P_{\mathcal{X}}^{V}\right)^{T}\right) V^{-1}, \\ &= (I - P_{\mathcal{X}})\left(I - P_{\mathcal{X}}^{V}\right), \\ &= (I - P_{\mathcal{X}}). \end{aligned}$$

The first required expression is

$$A\,(I - P_{\mathcal{X}})\, V\, (I - P_{\mathcal{X}}) = (I - P_{\mathcal{X}})\, V\, (I - P_{\mathcal{X}}).$$

This verifies (1.6.13). The second part considers

$$\begin{aligned} V^{-1}\left(I - P_{\mathcal{X}}^{V}\right) A &= V^{-1}\left(I - P_{\mathcal{X}}^{V}\right)(I - P_{\mathcal{X}}), \\ &= V^{-1}\left(I - P_{\mathcal{X}}^{\mathcal{V}}\right) \end{aligned}$$

by (1.6.11). This verifies (1.6.14).

■■

Chapter 2

Least squares computational problems

2.1 Introduction

This chapter is not intended to be an all-embracing account of computational algorithms for the least squares and constrained least squares problems, nor is it intended to be a thorough account of the associated error analysis. Rather the aim is to summarise aspects of these problems that are relevant to numerical considerations stemming from the developments of the previous chapter. Further information on the general problems can be found in the definitive accounts [11] and [46] respectively.

2.2 Perturbation of least squares problems

2.2.1 Case of fixed perturbations

There is no real restriction in assuming for the moment that the design matrix satisfies $\|A\| = \sqrt{n}$. It has been assumed already that the model structure ensures $\frac{1}{n}A^TA \overset{r.e.}{\underset{n\to\infty}{\to}} G$ where the Gram matrix G is bounded, positive definite and that p is fixed, so this further assumption amounts to a rescaling of the design by a quantity which is asymptotically constant.

We consider the generic perturbed least squares problem (1.1.1) with data

$$\mathbf{r} = (A + \tau E)\,\mathbf{x} - (\mathbf{b} + \tau \mathbf{z})$$

where perturbations E, $\mathbf{z}$ are fixed in the sense that they result from a well defined rule for each n. The perturbation E is assumed to be independent

of any observational error . It is assumed that τ is a small parameter which determines the scale of the perturbations. The component-wise scales of E and $\mathbf{z}$ are fixed by requiring

$$\max_{i,j} \|E_{i,j}\| = \eta \leq 1, \ \|\mathbf{z}\|_\infty \leq 1. \tag{2.2.1}$$

It is assumed also that τ is small enough for both A and $A+\tau E$ to have full rank p. The necessary conditions for the perturbed and unperturbed least squares problems give

$$(A+\tau E)^T \widehat{\mathbf{r}} = 0, \ A^T \mathbf{r}^{(n)} = 0$$

where the $\widehat{\ }$ indicates the solution of the perturbed problem, and the superscript (n) the exact solution of the original problem. Subtracting gives

$$(A+\tau E)^T \left(\widehat{\mathbf{r}} - \mathbf{r}^{(n)}\right) + \tau E^T \mathbf{r}^{(n)} = 0,$$

and substituting for the residual vectors gives the basic relation

$$(A+\tau E)^T (A+\tau E) \left(\widehat{\mathbf{x}} - \mathbf{x}^{(n)}\right) = \tau \left\{ (A+\tau E)^T \left(\mathbf{z} - E\mathbf{x}^{(n)}\right) - E^T \mathbf{r}^{(n)} \right\}. \tag{2.2.2}$$

For small enough τ this gives

$$\begin{aligned} \mathbf{x} - \mathbf{x}^{(n)} &= \tau \left\{ \left(A^T A\right)^{-1} \left(A^T \left(\mathbf{z} - E\mathbf{x}^{(n)}\right) - E^T \mathbf{r}^{(n)}\right) \right\} + O\left(\tau^2\right), \\ &= \tau \left\{ \begin{array}{c} \left(\frac{1}{\sqrt{n}} U\right)^{-1} \frac{1}{\sqrt{n}} Q_1^T \left(\mathbf{z} - E\mathbf{x}^{(n)}\right) \\ -\left(\frac{1}{n} A^T A\right)^{-1} \frac{1}{n} E^T \mathbf{r}^{(n)} \end{array} \right\} + O\left(\tau^2\right), \end{aligned} \tag{2.2.3}$$

where A possesses the orthogonal Q times upper triangular U factorization

$$A = Q \left[\begin{array}{c} U \\ 0 \end{array} \right] = \left[\begin{array}{cc} Q_1 & Q_2 \end{array} \right] \left[\begin{array}{c} U \\ 0 \end{array} \right] = Q_1 U,$$

and Q_1 corresponds to the first p columns of Q. There are two ways of looking at this relation. The first considers n fixed and worries about the size of $\mathrm{cond}\,(A) = \frac{\sigma_p}{\sigma_1}$, the ratio of the largest to smallest singular values of A. This describes the problem sensitivity corresponding to $\tau \to 0$ in the order term. It leads to the basic inequality

$$\left\|\widehat{\mathbf{x}} - \mathbf{x}^{(n)}\right\| \leq \tau \left\{ \frac{\mathrm{cond}\,(A)}{\sqrt{n}} \left\|\mathbf{z} - E\mathbf{x}^{(n)}\right\| + \frac{\mathrm{cond}\,(A)^2}{n} \left\|E^T \mathbf{r}^{(n)}\right\| \right\} + O\left(\tau^2\right), \tag{2.2.4}$$

where the assumption that $\|A\| = \sqrt{n} = \sigma_p$ has been used. The original form of this result is due to [40]. It highlights possible dependence on $\mathrm{cond}\,(A)^2$ which is likely if $\frac{1}{n}\left\|E^T\mathbf{r}^{(n)}\right\|$ is not small. The importance of this inequality is that it is a generic result highlighting what is best possible. For this reason computational algorithms in which the error takes this form are said to have *optimal error structure* . Methods based on orthogonal factorization prove to be important in the development of such optimal algorithms. These techniques go back to [48], [35], and [9]. It follows from (2.2.3) that

$$\begin{aligned}\widehat{\mathbf{r}} - \mathbf{r}^{(n)} &= -\tau\left\{(I-P)\,\mathbf{z} + PE\mathbf{x}^{(n)} + A\left(A^TA\right)^{-1}E^T\mathbf{r}^{(n)}\right\} + O\left(\tau^2\right),\\ &= -\tau\left\{(I-P)\,\mathbf{z} + PE\mathbf{x}^{(n)} + Q_1U^{-1}E^T\mathbf{r}^{(n)}\right\} + O\left(\tau^2\right) \qquad (2.2.5)\end{aligned}$$

where P is the orthogonal projection $A\left(A^TA\right)^{-1}A^T$ onto the range of A. Thus the result of the perturbation is a change of $O\left(\mathrm{cond}\,(A)\right)$ on the residual showing that a more satisfactory result is possible if the residual is the required quantity.

However, there is an alternative way of considering this result which is important when n is large and $\boldsymbol{\varepsilon}$ is a random vector. Here the order terms must be interpreted in the sense that $n \to \infty$ and τ is small enough. The Gram matrix G (1.1.7) can be used to write a limiting form of (2.2.3) as $n \to \infty$. Contributions from the quadrature error terms have been ignored (the following Lemma shows they contribute at most $\tau\left(o\left(1\right)\right)$ given regular sampling), and $G^{1/2}$ is written for the large n approximation to $\frac{1}{\sqrt{n}}U$ in (2.2.3).

$$\widehat{\mathbf{x}} - \mathbf{x}^{(n)} = \tau\left\{G^{-1/2}\frac{1}{\sqrt{n}}Q_1^T\left(\mathbf{z} - E\mathbf{x}^{(n)}\right) - G^{-1}\frac{1}{n}E^T\mathbf{r}^{(n)}\right\} + O\left(\tau\left(o\left(1\right)\right), \tau^2\right). \qquad (2.2.6)$$

We have the following bounds for the interesting terms in this equation.

Lemma 2.1

$$\frac{1}{\sqrt{n}}\left\|Q_1^T\left(\mathbf{z} - E\mathbf{x}^{(n)}\right)\right\| \leq \left\|\mathbf{z} - E\mathbf{x}^{(n)}\right\|_\infty,$$

$$\frac{1}{n}\left\|E^T\mathbf{r}^{(n)}\right\| \leq \sqrt{\frac{p}{n}}\eta\left\|\mathbf{r}^{(n)}\right\|_\infty.$$

Proof. The first part applies the standard inequality relating the 2 and ∞ norms. The second part follows in similar fashion from the inequality

$$\|E\| \leq \sqrt{np}\eta, \qquad (2.2.7)$$

where the the right hand side is a simple bound for the Frobenius norm of E. ■

Also it is important when fixing the order of dependence on τ that the n dependence of the $O(\tau)$ terms is appropriately bounded as $n \overset{r.e.}{\to} \infty$. The key is the following result.

Lemma 2.2

$$\frac{1}{n}(A+\tau E)^T(A+\tau E) \underset{n\to\infty}{\overset{r.e.}{\to}} G + O(\tau).$$

It follows that the normal matrix associated with the perturbed least squares problem has a suitably bounded inverse under regular sampling.

Proof.

$$\begin{aligned}\frac{1}{n}(A+\tau E)^T(A+\tau E) = \frac{1}{n}U^T\left\{I+\tau\left\{Q_1^T EU^{-1}+U^{-T}E^TQ_1\right\}\right.\\ \left.+\tau^2U^{-T}E^TEU^{-1}\right\}U.\end{aligned} \tag{2.2.8}$$

To show that the terms multiplying both τ and τ^2 in this expression are $O(1)$, $n \overset{r.e.}{\to} \infty$, requires a bound for $\|EU^{-1}\|$ valid for large n. Note $\|EU^{-1}\| \geq \left\|Q_1^TEU^{-1}\right\|$. The required bound can be constructed as follows:

$$\begin{aligned}\left\|U^{-T}E^TEU^{-1}\right\| &= \sup_{\mathbf{v}}\frac{\mathbf{v}^TU^{-T}E^TEU^{-1}\mathbf{v}}{\mathbf{v}^T\mathbf{v}},\\ &= \sup_{\mathbf{w}}\frac{\mathbf{w}E^TE\mathbf{w}}{\mathbf{w}U^TU\mathbf{w}},\\ &\leq \frac{\|E\|^2}{n\sigma_{min}\left\{\frac{1}{n}A^TA\right\}},\\ &\leq \frac{p\eta^2}{\sigma_{min}\{G\}}+o(1),\ n\overset{r.e.}{\to}\infty,\end{aligned}$$

where the estimate of $\|E\|$ obtained in the previous Lemma has been used. Thus

$$\left\|\frac{1}{n}(A+\tau E)^T(A+\tau E)-G\right\| \leq \tau\left\{3\|G\|^{1/2}\sqrt{p\,\mathrm{cond}(G)}+o(1)\right\},\ n\overset{r.e.}{\to}\infty.$$

The last step uses $\tau^2\|EU^{-1}\|^2 \leq \tau\|EU^{-1}\|$ when $\tau\|EU^{-1}\| \leq 1$. ■

Remark 2.2.1 *This result shows that provided a regular sampling scheme applies then all terms in the basic relation (2.2.6) have the orders claimed as $n \to \infty$. More can be said if the law of large numbers (see Appendix 3.7) can be applied to estimate $E^T\mathbf{r}^{(n)}$. It follows from the necessary conditions that*

$$A^T\mathbf{r}^{(n)} = 0 \Rightarrow Q_2Q_2^T\mathbf{r}^{(n)} = \mathbf{r}^{(n)}.$$

This means that in the case of an exact model the necessary conditions give

$$\mathbf{r}^{(n)} = Q_2Q_2^T\left(A\left(\mathbf{x}^{(n)} - \mathbf{x}^*\right) - \boldsymbol{\varepsilon}\right) = -Q_2Q_2^T\boldsymbol{\varepsilon}.$$

so that

$$\begin{aligned} \frac{1}{n}E^T\mathbf{r}^{(n)} = -\frac{1}{n}E^TQ_2Q_2^T\boldsymbol{\varepsilon} &\overset{a.s.}{\underset{n\to\infty}{\to}} 0, \\ \Rightarrow \frac{1}{n}G^{-1}E^T\mathbf{r}^{(n)} &\overset{a.s.}{\underset{n\to\infty}{\to}} 0. \end{aligned} \tag{2.2.9}$$

by the law of large numbers . This provides a sense in which the term in $G^{-1/2}$ dominates in(2.2.6) for large n.

Remark 2.2.2 *The above discussion has been in terms of prescribed perturbations and exact arithmetic. However, at least as important is the question what information, if any, this discussion can give regarding the behaviour when E, $\mathbf{z}$, τ are determined by the nature of the computational procedure and the characteristics of floating point arithmetic and n is large.*

1. *The scale τ is determined by the requirement that the component-wise scaling conditions (2.2.1) are satisfied. If these are set by reference to worst case error analysis (for example,[46], Theorem, 19.3) , then this suggests $\tau = \gamma_n u$ where u is unit roundoff and $\gamma_n = O(n)$. This is not compatible with the previous asymptotic results.*

2. *The application of the law of large numbers requires that $\boldsymbol{\varepsilon}$ be independent of E, $\mathbf{z}$. This cannot be strictly true here as right hand side values must influence rounding error behaviour to some extent.*

3. *The values of E, $\mathbf{z}$ depend on the detail of the particular algorithm implemented.*

Putting aside the setting of τ for the moment, some progress can be made on other matters. Consider the Golub orthogonal factorization algorithm based on Aitken–Householder transformations [35]. A suitable form of error analysis is given in [51] for the hypothetical case in which the orthogonal matrix Q is estimated by computing the product of the component

Aitken–Householder transformations explicitly. This shows that the potential $\mathrm{cond}\,(A)^2$ *contribution comes from a term*

$$\Delta = U^{-1}\delta Q \frac{1}{\sqrt{n}}\mathbf{r}^{(n)}, \tag{2.2.10}$$

where δQ *is the error in the computed orthogonal transformation. The computation of the factorization matrix* Q *does not involve the problem right hand side so the potential rounding error/stochastic error interactions can only contribute to potential* $\mathrm{cond}\,(A)$ *terms. Now* Δ *can be estimated using the law of large numbers provided the individual elements of* δQ *have an* $O\left(\frac{1}{\sqrt{n}}\right)$ *estimate, a magnitude typical of the elements of an* $n \times n$ *orthogonal matrix.*

A corresponding result to 2.2.10 does not appear available for more usual methods of treating Q*. However, it suggests that the previous analysis could be applied here provided* τ *is small. This requires systematic cancellation not allowed for in the setting of* $\tau = \gamma_n u$ *based on worst case analysis. There is more hope from informal observations which would seem to suggest that the cumulative effects of rounding errors prove relatively small in large computations. This could indicate something like a weak-mixing form of a law of large numbers (weak mixing because there is certainly some local rounding error interaction). Such a law need not depend on the precise statistics of individual rounding errors, and could be compatible with worst case analysis in the sense of allowing certain exceptional cases by analogy with almost sure convergence*

A similar program can be carried out in the generalised least squares case [85]. Some aspects will be considered here first under the assumptions that the principal interest is in the conditioning of the design matrix A, and that there is inherent structural stability in the weighting matrix V so that the rank is fixed by the problem structure, as in the case of least squares subject to equality constraints, and is invariant under allowed perturbations. Let perturbations of the orthogonal factorization of the design matrix be given by

$$A + \tau E = [Q_1 + \tau P_1]\,[U + \tau R]$$

where

$$[Q + \tau P]^T\,[Q + \tau P] = I \Rightarrow P_1^T Q_1 + Q_1^T P_1 = 0$$

to first order in τ. Then the first order term in the perturbation of the minimum variance solution operator T is given by

$$-U^{-1}RT + TQP^T + U^{-1}\left[\begin{array}{cc} 0 & \Delta \end{array}\right] Q^T$$

where Δ is the contribution from the term involving $V_{12}V_{22}^{-1}$. The dependence on $\text{cond}(A)^2$ must occur here also, and this is most easily seen by considering the case $V = I \Rightarrow T = U^{-1}Q_1^T$, $\Delta = 0$. We have

$$\begin{aligned}
-U^{-1}RT + TQP^T &= -U^{-1}Q_1^T ET + U^{-1}Q_1^T P_1 UT + TQP^T, \\
&= -U^{-1}Q_1^T ET - U^{-1}P_1^T Q_1 Q_1^T + U^{-1}P_1^T, \\
&= -U^{-1}Q_1^T ET + U^{-1}P_1^T \left(I - Q_1 Q_1^T\right), \\
&= -U^{-1}Q_1^T ET + U^{-1}U^{-T}\left(E^T - R^T Q_1^T\right)\left(I - Q_1 Q_1^T\right), \\
&= -U^{-1}Q_1^T ET + U^{-1}U^{-T}E^T\left(I - Q_1 Q_1^T\right).
\end{aligned}$$

This agrees with (2.2.3) on noting that $T\mathbf{b} = \mathbf{x}^{(n)}$, $\left(I - Q_1 Q_1^T\right)\mathbf{b} = \mathbf{r}^{(n)}$, and adding in the contribution from the right hand side. The other extreme corresponds to the assumption that V is perturbed, but $\left[\begin{array}{cc} A & \mathbf{b} \end{array}\right]$ remains fixed. In this case it is the perturbation to Δ that matters, and the important term is $\text{cond}\left(Q_2^T V Q_2\right)$. This conclusion is worse than that obtained in the analysis in [85]. The result there gives a dependence on $\text{cond}\left(Q_2^T V Q_2\right)^{\frac{1}{2}}$ by an analysis of an algorithm for solving (1.2.4) which avoids the use of Lagrange multipliers.

Exercise 2.2.1 *Show that the norm assumptions* $\|A\| = \sqrt{n}$, $\|E\|_\infty = 1$ *are compatible with the requirement* $\frac{1}{n}A^T A \to G \in R^p \to R^p$, *bounded.*

2.2.2 Rounding error implications

Perhaps the most interesting result of the fixed perturbation analysis centres on the use of the law of large numbers to show that the role of the term involving the square of the condition number is deemphasised if n is large enough. The other component of this argument is that the condition number, although certainly it can be large in commonly used models, is bounded as $n \to \infty$. When the perturbations are attributed to rounding error , the second part of this argument remains steady enough. However, the first part becomes more conjectural because it is no longer valid to assume that the perturbations are independent of the observational error . The conjecture that this result remains true has to be based on the expectation that the mixing between the two processes is weak enough for a form of the law of large numbers to continue to hold. The argument is that the rounding should mostly be independent of the observational error if there is no significant relative amplification of this relative to the signal (but note that significant is here a "value loaded" descriptor).

Apart from this distinct awkwardness the perturbation development is useful. For example, the rounding error analysis for orthogonal factorization is summarised in [46] where it is presented in just the right form for mimicking the above development. Key inequalities are:

$$\frac{\|\Delta U\|_F}{\|U\|_F} \leq c_n \operatorname{cond}(A) \frac{\|\Delta A\|_F}{\|A\|_F},$$
$$\|\Delta Q\|_F \leq c_n \operatorname{cond}(A) \frac{\|\Delta A\|_F}{\|A\|_F},$$

where $\|.\|_F$ is the Frobenius norm, c_n is a constant, and the scaling of each of the perturbations $\Delta Q,\ \Delta U,\ \Delta A$ has the form eps $f(n,p)$ where eps is the machine precision, and $f(n,p)$ is characteristic of the role of the particular quantity in the algorithm. Each contribution is required to be small enough for the argument leading to these inequalities to hold.

2.3 Main computational algorithms

2.3.1 Cholesky factorization

The classical computational algorithm for solving the linear least squares problem involved forming the normal matrix $M = A^T A$, factorising this into a lower triangular matrix times its transpose

$$M = LL^T, \tag{2.3.1}$$

and then performing the forward and back substitutions

$$L\mathbf{z} = A^T\mathbf{b};\ L^T\mathbf{x} = \mathbf{z}.$$

This approach has an unavoidable worst case conditioning dependence on $\kappa(M) = \kappa(A)^2$. This is potentially much worse than the $\kappa(A)$ sensitivity of the underlying problem in the cases that errors are random or when $\|\mathbf{r}\|$ is small, so this is an algorithm that does not have optimal error structure . The source of this difficulty occurs already in forming M and $A^T\mathbf{b}$. These computations lead to results which can be represented as

$$\text{fl}(M) = (A+\tau E_1)^T (A+\tau E_1);\ \text{fl}\left(A^T\mathbf{b}\right) = (A+\tau E_2)^T \mathbf{b}$$

where the differing subscripts on the perturbation terms indicate the rounding errors arise from independent computations. Here the standard backward analysis for the Cholesky factorisation leads to a computed solution [46] which is the exact solution of

$$\left((A+\tau E_1)^T (A+\tau E_1) + \tau\Delta M\right)\widehat{\mathbf{x}} = (A+\tau E_2)^T \mathbf{b}$$

where $\tau\Delta M$ takes account of the rounding contributions in the solution process. The leading term in the error in this case is

$$\Delta\mathbf{x} = \tau M^{-1}\left(E_2^T\mathbf{b} - \left(E_1^T A + A^T E_1 + \Delta M\right)\widehat{\mathbf{x}}\right)$$

which clearly shows the $\kappa(M)$ dependence for fixed n. As a consequence of this unfavourable condition estimate which is certainly valid when $\|r^{(n)}\|$ is small the Cholesky algorithm has fallen out of favour as a method for "standard" linear least squares problem. However, the method is important a priori in reducing the generalised least squares problem to the form (1.2.4) when the covariance matrix $V \neq I$ is given. It also has potential advantages when the design matrix is large and sparse or possesses other structural properties which need to be exploited to reduce computational cost.

The Cholesky factorisation is a recursive procedure which at each intermediate stage gives the factors of a leading principal sub-matrix. Let M_{i-1} be the $(i-1)\times(i-1)$ principal sub-matrix. Then $L_{i-1} = M_{i-1} = 0$ for $i = 1$, and for $i = 2, 3, \cdots, n$ the factors of M_i are determined by the relation

$$\begin{bmatrix} L_{i-1} & \\ \mathbf{l}_i^T & l_{ii} \end{bmatrix}\begin{bmatrix} L_{i-1}^T & \mathbf{l}_i \\ & l_{ii} \end{bmatrix} = \begin{bmatrix} M_{i-1} & \mathbf{m}_i \\ \mathbf{m}_i^T & m_{ii} \end{bmatrix} = M_i.$$

We have

$$L_{i-1}\mathbf{l}_i = \mathbf{m}_i \tag{2.3.2}$$

which gives $\mathbf{l}_i$ by a forward substitution. Then

$$\mathbf{l}_i^T\mathbf{l}_i + l_{ii}^2 = m_{ii}. \tag{2.3.3}$$

Positive definiteness of M ensures that $m_{ii} > \mathbf{l}_i^T\mathbf{l}_i$ in exact arithmetic, but this property can be destroyed by rounding error in ill conditioned cases. Also, there is an ambiguity in the sign of l_{ii} which typically is removed by taking the positive square root

$$l_{ii} = \sqrt{m_{ii} - \mathbf{l}_i^T\mathbf{l}_i},\ l_{ii} > 0.$$

This factorisation process lends itself to several different organisations. In the above discussion L is built up a row at a time, but it can equally well be found a column at a time. The quantities that are required to be computed to complete stage i of the recursion are $l_{ii}, l_{(i+1)i}, \cdots, l_{ni}$. These are found from the relation

$$\begin{bmatrix} L_{i-1} & 0 \\ \mathbf{l}_i^T & l_{ii} \end{bmatrix}\begin{bmatrix} L_{i-1}^T & \mathbf{l}_i & \mathbf{l}_{i+1}^{(i)} & \cdots & \mathbf{l}_n^{(i)} \\ 0 & l_{ii} & l_{(i+1)i} & \cdots & l_{ni} \end{bmatrix} =$$
$$\begin{bmatrix} M_{i-1} & \mathbf{m}_i & \mathbf{m}_{i+1}^{(i)} & \cdots & \mathbf{m}_n^{(i)} \\ \mathbf{m}_i^T & m_{ii} & m_{i(i+1)} & \cdots & m_{in} \end{bmatrix}$$

The previous step of the recursion gives

$$L_{i-1}\mathbf{l}_k^{(i)} = \mathbf{m}_k^{(i)},\ k = i, i+1, \cdots, n,$$

so that l_{ii} is found as before, while the l_{ki} satisfy the relations

$$\mathbf{l}_i^T\mathbf{l}_k^{(i)} + l_{ii}l_{ki} = m_{ki},\ k = i+1, \cdots, n.$$

This completes the computation of the i'th row of L^T and simultaneously gives

$$\mathbf{l}_k^{(i+1)} = \begin{bmatrix} \mathbf{l}_k^{(i)} \\ l_{ki} \end{bmatrix},\ k = i+1, \cdots, n,$$

ready for the next stage of the computation.

An important variant forms the factorisation in modified form

$$M = LDL^T \tag{2.3.4}$$

where the diagonal elements of L are set equal to 1. In this case the factorisation is unique and can be computed without extracting square roots. It can be written down immediately from the standard form, but it is convenient to develop a modified recursion. To do this let M_k be the submatrix factorised at step k. Then the component terms in the partial factorization in column oriented form satisfy

$$M - \begin{bmatrix} L_k & \\ L_2^{(k)} & I \end{bmatrix} \begin{bmatrix} D_k & \\ & 0 \end{bmatrix} \begin{bmatrix} L_k^T & L_2^{(k)T} \\ & I \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & M^k \end{bmatrix}. \tag{2.3.5}$$

Equating terms gives

$$\begin{aligned} L_kD_kL_k^T &= M_k, \\ L_kD_kL_2^{(k)T} &= M_{(1:k)(k+1:n)}, \\ M^k &= M_{(k+1:n)(k+1:n)} - L_2^{(k)}D_kL_2^{(k)T}, \\ &= M_{(k+1:n)(k+1:n)} - M_{(k+1:n)(1:k)}L_k^{-T}D_k^{-1}D_kD_k^{-1}L_k^{-1}M_{(1:k)(k+1:n)}, \\ &= M_{(k+1:n)(k+1:n)} - M_{(k+1:n)(1:k)}M_k^{-1}M_{(1:k)(k+1:n)}. \end{aligned}$$

This shows that M^k is the Schur complement of M_k in M. It is a standard result that $M^k \succ 0$ if $M \succ 0$. To develop a recursion for the factors set

$$\begin{aligned} D_k &= \begin{bmatrix} D_{k-1} & \\ & d_k \end{bmatrix}, \\ L_k &= \begin{bmatrix} L_{k-1} & \\ \mathbf{l}_1^{(k)T} & 1 \end{bmatrix}, \\ L_2^{(k)} &= \begin{bmatrix} \left(L_2^{(k-1)}\right)_{(2:n-k+1)*} & \mathbf{l}_2^{(k)} \end{bmatrix}. \end{aligned}$$

Here $\mathbf{l}_1^{(k)T} = \mathbf{e}_1^T L_2^{(k-1)}$ is already known in a column oriented method, and it is necessary to compute d_k and $\mathbf{l}_2^{(k)}$ so that (2.3.5) holds. Equating the remaining terms gives

$$\begin{aligned} d_k &= m_{kk} - \mathbf{l}_1^{(k)T} D_{k-1} \mathbf{l}_1^{(k)}, \\ M_{(k+1:n)k} &= \left[\begin{array}{cc} \left(L_2^{(k-1)}\right)_{(2:n-k+1)*} & \mathbf{l}_2^{(k)} \end{array} \right] D_k L_k^T \mathbf{e}_k, \\ &= d_k \mathbf{l}_2^{(k)} + \left(L_2^{(k-1)}\right)_{(2:n-k+1)*} D_{k-1} \mathbf{l}_1^{(k)}. \end{aligned}$$

Also the elements of D can be generated recursively. Let

$$d_k^0 = m_{kk},\ d_k^j = d_k^{j-1} - d_{j-1} \left(L_{k(j-1)}\right)^2,\ j = 1, 2, \cdots, k, \tag{2.3.6}$$

where $d_0 = L_{k0} = 0$. Then

$$d_k^j \geq d_k^k > 0$$

as a consequence of $M_k \succ 0$, and

$$d_k = d_k^k.$$

We have seen that the generalised least squares problem can be solved under weaker conditions than $V \succ 0$. This makes it of interest to determine the Cholesky factors in the semi-definite case. If $\text{rank}\,(M) = r$ then the factorisation will have the form

$$\left[\begin{array}{cc} L_r & \\ L_2^{(r)} & I \end{array} \right] \left[\begin{array}{cc} D_r & \\ & 0 \end{array} \right] \left[\begin{array}{cc} L_r^T & L_2^{(r)T} \\ & I \end{array} \right] = M,$$

so that $M^r = 0$. This corresponds to the conditions

$$d_k^j = 0,\ j = r+1, r+2, \cdots, n.$$

If there is a need to determine information on the rank of M then there is a definite advantage in modifying the algorithm to permit a pivoting strategy in which rows and columns of M^k are exchanged in order to bring the maximum element (necessarily on the diagonal) to the leading position before the next stage of the factorisation. This can be done equivalently by fixing the new pivotal row and column by selecting the maximum element of the set $\left\{d_j^k\right\}_{j=k+1}^n$ to become d_{k+1}in the permuted matrix as these elements are non-increasing as a function of k. This pivoting strategy ensures $d_1 \geq d_2 \geq \cdots \geq d_r > 0$. This pivoting strategy can be thought of as eliminating the most significant elements as part of a greedy strategy to reduce

M^k to zero. Criteria for the success of this factorisation procedure in real arithmetic are discussed in [46]. They boil down to the requirement that the smallest eigenvalue of M_k, when scaled to have diagonal elements unity, is large enough for $k \leq r$. Note that this is a criterion for the success of the factorisation - not a criterion for the success of the rank determination. Further analysis of the triangular factors may be required to obtain this information [43]. What is achieved here when the computations are carried out in real arithmetic are factors for a perturbed and permuted matrix whose rank is known. If this perturbation is small then this result is frequently all that is required.

If the interchange applied to M^k after step k involves exchanging row and column $j > k+1$ of the full matrix then, if P_{kj} is the symmetric elementary permutation matrix interchanging rows $k+1$ and j, $P_{kj} = I - (\mathbf{e}_{k+1} - \mathbf{e}_j)(\mathbf{e}_{k+1} - \mathbf{e}_j)^T$, the exchange is effected by

$$\begin{aligned} P_{kj} \begin{bmatrix} 0 & 0 \\ 0 & M^k \end{bmatrix} P_{kj} &= P_{kj} M P_{kj} - P_{kj} \begin{bmatrix} L_k & \\ L_2^{(k)} & I \end{bmatrix} P_{kj} P_{kj} \begin{bmatrix} D_k & \\ & 0 \end{bmatrix} \\ &\quad P_{kj} P_{kj} \begin{bmatrix} L_k^T & L_2^{(k)T} \\ & I \end{bmatrix} P_{kj}, \\ &= P_{kj} M P_{kj} - \begin{bmatrix} L_k & \\ \widetilde{L}_2^{(k)} & I \end{bmatrix} \begin{bmatrix} D_k & \\ & 0 \end{bmatrix} \begin{bmatrix} L_k^T & \widetilde{L}_2^{(k)T} \\ & I \end{bmatrix}, \end{aligned}$$

where $\widetilde{L}_2^{(k)}$ is obtained from $L_2^{(k)}$ by interchanging the first and $(j-k)+1$ rows and the lower triangular form of the factor is preserved. Thus the rank revealing factorisation in the sense indicated above produces an LDL^T factorisation of PMP^T where the permutation matrix P is given by

$$P = P_{(n-2)j(n-2)} \cdots P_{0j(0)} = \prod_{k=n-2}^{0} P_{kj(k)}$$

where $j(k)$ is the index of the maximum diagonal element of the permuted form of M^k.

Remark 2.3.1 *Positive definiteness requires that $d_k^j > 0$ in (2.3.6). It follows that in the rank revealing factorisation*

$$0 < \frac{d_k^{j-1}}{d_j} - (L_{kj})^2 \Rightarrow |L_{kj}| < 1$$

as $d_k^{j-1} \leq d_j$ as a consequence of the diagonal pivoting . This condition puts a limit on the possible ill-conditioning in L and serves to concentrate this in

D in the case that M is ill-conditioned. A simple case is illustrated in the next example.

Example 2.3.1 *The use of diagonal pivoting helps to concentrate ill-conditioning in M into the diagonal matrix D. Consider the identity*

$$\begin{bmatrix} 1 & \\ \gamma & 1 \end{bmatrix} \begin{bmatrix} 1 & \\ & 1 \end{bmatrix} \begin{bmatrix} 1 & \gamma \\ & 1 \end{bmatrix} = \begin{bmatrix} 1 & \gamma \\ \gamma & 1+\gamma^2 \end{bmatrix}.$$

Here the matrix has been defined by a factorisation which has non-decreasing elements in D, which is well conditioned, but has distinctly ill-conditioned L - $\mathrm{cond}(L) = O(\gamma)$ for large γ. On the other hand the modified Cholesky factorisation with diagonal pivoting gives

$$\begin{bmatrix} 1+\gamma^2 & \gamma \\ \gamma & 1 \end{bmatrix} = \begin{bmatrix} 1 & \\ \frac{\gamma}{1+\gamma^2} & 1 \end{bmatrix} \begin{bmatrix} 1+\gamma^2 & \\ & \frac{1}{1+\gamma^2} \end{bmatrix} \begin{bmatrix} 1 & \frac{\gamma}{1+\gamma^2} \\ & 1 \end{bmatrix}.$$

Example 2.3.2 *It is important that structure be preserved in cases when n is large. This need not be achieved by the rank revealing factorisation. Consider the LDL^T factorisation*

$$M = \begin{bmatrix} 5 & 5 & & & \\ 5 & 9 & 4 & & \\ & 4 & 7 & 3 & \\ & & 3 & 5 & 2 \\ & & & 2 & 3 \end{bmatrix} = LDL^T.$$

where

$$L = \begin{bmatrix} 1 & & & & \\ 1 & 1 & & & \\ & 1 & 1 & & \\ & & 1 & 1 & \\ & & & 1 & 1 \end{bmatrix}, \; D = \begin{bmatrix} 5 & & & & \\ & 4 & & & \\ & & 3 & & \\ & & & 2 & \\ & & & & 1 \end{bmatrix}.$$

Let M be scaled to have diagonal elements 1*. This gives*

$$M_S = \begin{bmatrix} 1 & .745 & & & \\ .745 & 1 & .504 & & \\ & .504 & 1 & .507 & \\ & & .507 & 1 & .516 \\ & & & .516 & 1 \end{bmatrix}$$

The permutation sequence induced by diagonal pivoting is $\{1,3,5,4,2\}$. *The permuted matrix is*

$$PMP^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & .745 \\ 0 & 1 & 0 & .507 & .504 \\ 0 & 0 & 1 & .516 & 0 \\ 0 & .507 & .516 & 1 & 0 \\ .745 & .504 & 0 & 0 & 1 \end{bmatrix}$$

and the factors are

$$\widetilde{L} = \begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & 0 & 1 & & \\ 0 & .507 & .516 & 1 & \\ .745 & .504 & 0 & -.537 & 1 \end{bmatrix}, \ \widetilde{D} = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & .476 & \\ & & & & .053 \end{bmatrix}.$$

Note, in particular, that $\widetilde{L}$ *has more non-zero elements than* L. *This pattern is a property of this family of tri-diagonal* M *and continues as* n *increases with the asymptotic ratio of the number of non-zero elements approaching* 2.

Exercise 2.3.1 *Show that the Schur complement* M^k *in equation (2.3.5) is positive definite if* M *is.*

2.3.2 Orthogonal factorisation

The algorithm that replaced Cholesky factorisation as the preferred method for the linear least squares problem (1.1.1) is based on the reduction of the design matrix A to the product of an orthogonal matrix times a trapezoidal matrix with upper triangular and zero blocks (1.1.9) by means of a sequence of orthogonal transformations. The two families of transformations considered lead to algorithms with optimal error structure. The transformations are:

1. Aitken—Householder transformations based on elementary othogonal matrices
$$H = \left(I - 2\mathbf{w}\mathbf{w}^T\right), \ \mathbf{w}^T\mathbf{w} = 1 \tag{2.3.7}$$
(note the matrix is symmetric); and

2. plane rotation based transformations based on elementary reflectors
$$H = \begin{bmatrix} I & & & \\ & c & & s \\ & & I & \\ & s & & -c \\ & & & & I \end{bmatrix}, \ c^2 + s^2 = 1. \tag{2.3.8}$$

The first class is preferred for the direct reduction of a dense design matrix. The second class is more flexible as it just mixes two rows or columns so a sequence of these transformations is typically required to achieve the same result as a single Aitken—Householder transformation. However, plane rotations possess a flexibility that can prove valuable in sparse problems and in problems in which the design is systematically modified by the sequential addition or deletion of observations and by the sequential addition or deletion of variables. Several examples are given subsequently.

Elementary orthogonal matrices were introduced for systematic matrix reduction in [106]. The appreciation of their use in developing computational algorithms is due to Householder [48], and their popularisation to Golub [35]. The typical operation that defines the algorithmic use of an elementary orthogonal matrix H as a function of its defining vector $\mathbf{w}$ is the transformation of a given vector $\mathbf{v}$ into a multiple of a unit vector:

$$H\mathbf{v} = \left(I - 2\mathbf{w}\mathbf{w}^T\right)\mathbf{v} = \theta \left\|\mathbf{v}\right\| \mathbf{e}_1,\ \theta = \pm 1. \tag{2.3.9}$$

It is immediate that

$$\mathbf{w} = \gamma \left(\mathbf{v} - \theta \left\|\mathbf{v}\right\| \mathbf{e}_1\right)$$

where

$$1 = \mathbf{w}^T\mathbf{w} = 2\gamma^2 \left(\left\|\mathbf{v}\right\|^2 - \theta \left\|\mathbf{v}\right\| v_1\right).$$

Possible cancellation in computing $\mathbf{w}$ is minimized by choosing $\theta = -\operatorname{sgn}(v_1)$). This gives

$$\mathbf{w} = \frac{1}{\sqrt{2\left\|\mathbf{v}\right\|\left(\left\|\mathbf{v}\right\| + \left|v_1\right|\right)}} \left(\mathbf{v} - \theta \left\|\mathbf{v}\right\| \mathbf{e}_1\right). \tag{2.3.10}$$

Note that $H^2 = I$, and $\det(H) = -1$ so that H is a reflector.

To construct the factorisation (1.1.9) assume an intermediate step with the partial factorisation

$$A = Q_{i-1} \begin{bmatrix} U_{i-1} & U_{12}^{i-1} \\ 0 & A_i \end{bmatrix}.$$

Define the elementary reflector H_i by requiring it to map the first column of A_i to a multiple of $\mathbf{e}_1$. Then

$$H_i A_i = \begin{bmatrix} \theta \left\|(A_i)_{*1}\right\| & \mathbf{u}_i^T \\ 0 & A_{i+1} \end{bmatrix},$$

and set

$$Q_i = Q_{i-1} \begin{bmatrix} I & \\ & H_i \end{bmatrix},\ U_i = \begin{bmatrix} U_{i-1} & \left(U_{12}^{i-1}\right)_{*1} \\ & \theta \left\|(A_i)_{*1}\right\| \end{bmatrix},\ U_{12}^i = \begin{bmatrix} \left(U_{12}^{i-1}\right)_{*(2:p-i)} \\ \mathbf{u}_i^T \end{bmatrix}.$$

Then

$$A = Q_i \begin{bmatrix} U_i & U_{12}^i \\ & A_{i+1} \end{bmatrix}.$$

If A has its full rank $p < n$ then the required factorization is produced after p steps. in this case the resulting orthogonal transformation Q can be partitioned as

$$Q = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \tag{2.3.11}$$

where

$$A = Q_1 U,\ Q_2^T A = 0. \tag{2.3.12}$$

If $p = n$ corresponding to a square design then the number of transformations required is reduced to $p - 1$.

Remark 2.3.2 *The analog of the diagonal pivoting strategy in the Cholesky factorisation is a pivoting strategy that chooses the largest column norm of A_i to determine the column to be reduced at the current stage . As orthogonal transformations preserve length, column norms cannot be increased in the passage $i \to i+1$. It follows that column pivoting forces the diagonal elements of U to be non-increasing in magnitude. As these elements must be identical up to sign with the corresponding diagonal elements of L in the diagonal pivoting Cholesky factorisation the $L_{ii}, i = 1, p$ must be non-increasing in magnitude also.*

The fundamental theorem governing the error analysis of orthogonal factorisation based on Aitken–Householder transformations is due to Wilkinson [115], [116], and can be stated as follows.

Theorem 2.1 *Let $\widehat{U}_p$ be the computed upper triangular factor in (1.1.9). Then there exists an orthogonal matrix $\widehat{Q}_p$ such that*

$$A + \tau E = \widehat{Q}_p \begin{bmatrix} \widehat{U}_p \\ 0 \end{bmatrix}$$

where $\|E\|_F \leq np \|A\|_F$ and τ is of the order of the machine precision provided quite weak restrictions on the size of n are satisfied. Here $\widehat{Q}_p$ can be constructed as the product of the Aitken–Householder transformations that would be generated by carrying out each stage exactly using the actually computed numbers.

To solve the least squares problem using the Householder factorisation note that the invariance of the Euclidean norm under orthogonal transformation gives

$$\begin{aligned}\min_{\mathbf{x}} \|A\mathbf{x}-\mathbf{b}\|^2 &= \min_{\mathbf{x}} \left\| \begin{bmatrix} U_p \\ 0 \end{bmatrix} \mathbf{x} - Q_p^T \mathbf{b} \right\|^2, \\ &= \min_{\mathbf{x}} \left\| \begin{bmatrix} U_p\mathbf{x} - (Q_p)_1^T \mathbf{b} \\ (Q_p)_2^T \mathbf{b} \end{bmatrix} \right\|^2.\end{aligned}$$

It follows that the minimiser is

$$\mathbf{x} = U_p^{-1} (Q_p)_1^T \mathbf{b}, \tag{2.3.13}$$

while the norm of the residual satisfies

$$\|\mathbf{r}\| = \left\| (Q_p)_2^T \mathbf{b} \right\|.$$

The error analysis is very satisfactory [46]. This is summarised in the next result which assumes that A has full rank p.

Theorem 2.2 *The computed solution $\widehat{\mathbf{x}}^{(n)}$ is the exact least squares solution of a close by problem*

$$\min_{\mathbf{x}} \|(A+\tau E)\,\mathbf{x} - (\mathbf{b}+\tau\mathbf{z})\|_2^2$$

where the perturbations satisfy $\|E\|_F \leq np\,\|A\|_F$, $\|\mathbf{z}\|_2 \leq np\,\|\mathbf{b}\|$, and τ is of the order of the machine precision provided quite weak restrictions on the size of n are satisfied.

This is a backward error result, but it follows directly from this that the algorithm has optimal error structure .

The solution method based on the Aitken–Householder transformation based orthogonal factorisation has the minor disadvantage that neither Q_1, the orthogonal basis for range (A), nor $\mathbf{r}^{(n)}$, the optimal residual vector, is given explicitly. The hat matrix (1.1.10) provides one example of an important quantity that requires this information. However, both Q_1 and U can be computed explicitly by Gram-Schmidt orthogonalisation. Here the sequence in which the computations are performed is important for the rounding error analysis [9], the preferred form being referred to as the modified Gram-Schmidt algorithm (MGS). Consider

$$\begin{bmatrix} A & \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1, \cdots, \mathbf{a}_p & \mathbf{b} \end{bmatrix}.$$

The first step sets

$$\mathbf{a}_i^2 = \mathbf{a}_i - \frac{\mathbf{a}_1^T\mathbf{a}_i}{\|\mathbf{a}_1\|^2}\mathbf{a}_1,\ i = 2, 3, \cdots p,$$
$$\mathbf{b}^2 = \mathbf{b} - \frac{\mathbf{a}_1^T\mathbf{b}}{\|\mathbf{a}_1\|^2}\mathbf{a}_1,\ \mathbf{q}_1 = \mathbf{a}_1/\|\mathbf{a}_1\|.$$

This can be represented in matrix form as

$$\left[\begin{array}{cc} A & \mathbf{b} \end{array}\right] = \left[\begin{array}{cccc} \mathbf{q}_1 & \mathbf{a}_2^2, \cdots, \mathbf{a}_p^2 & \mathbf{b}^2 \end{array}\right] \left[\begin{array}{ccc} \|\mathbf{a}_1\| & \cdots \frac{\mathbf{a}_1^T\mathbf{a}_i}{\|\mathbf{a}_1\|} \cdots & \frac{\mathbf{a}_1^T\mathbf{b}}{\|\mathbf{a}_1\|} \\ & I & \\ & & 1 \end{array}\right].$$

Subsequent steps orthogonalise $\mathbf{a}_{i+1}^i, \cdots \mathbf{a}_p^i, \mathbf{b}^i$ to $\mathbf{a}_i^i$ for $i = 2, 3, \cdots, p$ in similar fashion. The result is

$$\left[\begin{array}{cc} A & \mathbf{b} \end{array}\right] = \left[\begin{array}{cc} Q_1 & -\mathbf{r}^{(n)} \end{array}\right] \left[\begin{array}{cc} U & \mathbf{z} \\ & 1 \end{array}\right], \tag{2.3.14}$$

where $U_{ii} = \|\mathbf{a}_i^i\|$, $U_{ij} = \frac{\mathbf{a}_i^{iT}\mathbf{a}_j^i}{\|\mathbf{a}_i^i\|}$, $\mathbf{z} = Q_1^T\mathbf{b}$. The solution of the linear least squares problem is now computed as

$$\mathbf{x}^{(n)} = U^{-1}\mathbf{z}.$$

The resulting algorithm has very satisfactory numerical properties. This can be seen most easily from the observation that the MGS procedure is identical with the application of p steps of Aitken–Householder transformations to the modified design matrix in $R^{p+1} \to R^{n+p}$ [10]

$$\left[\begin{array}{cc} \widetilde{A} & \widetilde{\mathbf{b}} \end{array}\right] = \left[\begin{array}{cc} 0 & 0 \\ A & \mathbf{b} \end{array}\right] \in R^{p+1} \to R^{n+p}.$$

By (2.3.10) the elementary orthogonal matrix in the first step is determined by the partitioned vector in R^{n+p}

$$\mathbf{w}_1 = \frac{1}{\sqrt{2}\,\|\mathbf{a}_1\|}\left[\begin{array}{c} -\|\mathbf{a}_1\|\,\mathbf{e}_1 \\ \mathbf{a}_1 \end{array}\right].$$

When applied to the k'th column it gives

$$\left(I - 2\mathbf{w}_1\mathbf{w}_1^T\right)\left[\begin{array}{c} 0 \\ \mathbf{a}_k \end{array}\right] = \left[\begin{array}{c} \frac{\mathbf{a}_1^T\mathbf{a}_k}{\|\mathbf{a}_1\|}\mathbf{e}_1 \\ \mathbf{a}_k - \frac{\mathbf{a}_1^T\mathbf{a}_k}{\|\mathbf{a}_1\|^2}\mathbf{a}_1 \end{array}\right],$$

indicating the manner in which the factorisation is developed. The final result is

$$\left[\begin{array}{cc} \widetilde{A} & \widetilde{\mathbf{b}} \end{array}\right] \rightarrow \left[\begin{array}{cc} U & \mathbf{z} \\ Q_1 & -\mathbf{r}^{(n)} \end{array}\right]. \tag{2.3.15}$$

An interesting feature of the implementation of MGS by means of elementary orthogonal transformations is that while the product of the Aitken–Householder matrices provides a computed result which is close to an orthogonal matrix, there is possible loss of orthogonality in the computed realisation of Q_1.

The result (2.3.15) shows that this implementation of MGS provides a lot of useful additional information. An alternative formulation provides even more. This represents a development of the sweep methods described in [65]. This starts with a different augmented matrix (note that replacing $\mathbf{b}$ by $-\mathbf{b}$ requires replacing $\mathbf{z}$ by $\mathbf{z} = -Q_1^T\mathbf{b}$)

$$\left[\begin{array}{cc} A & -\mathbf{b} \\ I & 0 \end{array}\right] \in R^{p+1} \rightarrow R^{n+p}. \tag{2.3.16}$$

Making use of the MGS equations (2.3.14) gives

$$\begin{aligned} \left[\begin{array}{cc} A & -\mathbf{b} \\ I & 0 \end{array}\right]\left[\begin{array}{cc} U & \mathbf{z} \\ 0 & 1 \end{array}\right]^{-1} &= \left[\begin{array}{cc} A & -\mathbf{b} \\ I & 0 \end{array}\right]\left[\begin{array}{cc} U^{-1} & -U^{-1}\mathbf{z} \\ 0 & 1 \end{array}\right], \\ &= \left[\begin{array}{cc} Q_1 & -\mathbf{b} - Q_1\mathbf{z} \\ U^{-1} & \mathbf{x}^{(n)} \end{array}\right], \\ &= \left[\begin{array}{cc} Q_1 & \mathbf{r}^{(n)} \\ U^{-1} & \mathbf{x}^{(n)} \end{array}\right]. \end{aligned} \tag{2.3.17}$$

Remark 2.3.3 *If $A_{*1} = \mathbf{e}$ then the first MGS orthogonalisation step automatically centres the remaining columns. Thus the MGS algorithm provides an elegant procedure for the inclusion of an intercept term in a statistical model. See subsection 1.1.4.*

The implementation of the algorithm is very simple.Each orthogonalisation step is computed as usual (this involves just the design matrix component of the augmented system) , but it is then applied to the full set of columns in a sweep step. The work is basically the same as in the implementation based on Aitken–Householder transformations, but the information gained is even more useful in form. Also, the algorithm produces the solutions to a sequence of partial regressions as it evolves. Thus it lends itself to variable selection computations . Stability of this form of the MGS algorithm has been tested numerically [78] with the results indicating optimal error structure, but a complete error analysis does not appear to be available.

Exercise 2.3.2 *Show that the sum of squares of residuals is produced explicitly by the sweep form of the MGS algorithm if the row*

$$\begin{bmatrix} -\mathbf{b}^T A & \mathbf{b}^T\mathbf{b} \end{bmatrix}$$

is added to the augmented matrix (2.3.16).

2.3.3 Methods for generalised least squares problems

The generalised least squares problem has a unique solution provided the augmented matrix $\begin{bmatrix} V & A \\ A^T & 0 \end{bmatrix}$ has full rank $n+p$. A sufficient condition that makes good sense either in the case of a known, parsimonious model or in an exploratory context is that the design matrix A has full column rank p and $Q_2^T V Q_2$ has full rank $n-p$ (chapter 1, condition 1.1). The factorization (1.2.5) and solution (1.2.6) provide the basis for an effective computational procedure in which the Cholesky factorization is applied to $Q_2^T V Q_2$. The full rank condition on A is an important aspect of an efficient model, while $Q_2^T V Q_2$ regulates the size of the residual. The catch with the generalised problem centres on $V : R^n \to R^n$. Treating this as an arbitrary positive definite matrix runs into problems with the large n (asymptotic limit) case. These problems are of two kinds and may have no easy answers:

theoretical Variance information is assumed available in the generalised least squares problem formulation but it is not the easiest quantity to come by. A nondiagonal form for V implies coupling between observations in the measurement process. Typically this results in narrow banded block structures in sequential experiments and in special structure reflecting symmetries in other cases. Methods for estimating V are considered subsequently.

practical The potential size of V requires that computational methods that exploit a priori structure such as sparsity are important. The characteristic feature of sparsity most often encountered here is that V depend on at most $O(n)$ distinct elements.

Standard factorizations

If V is given explicitly then its Cholesky factorisation is a typical first step as it has the advantage that it preserves a banded block structure in V in the sense that if V is $2n_p + 1$ banded then the Cholesky factor L is $n_p + 1$ banded with blocks of the same size. This result may not hold if diagonal

pivoting is used in a rank revealing factorization. If V is well conditioned then the least squares problem (1.2.4)can be written

$$\min \mathbf{s}^T\mathbf{s}; \mathbf{s} = L^{-1}A\mathbf{x} - L^{-1}\mathbf{b}.$$

The usual approach based on orthogonal factorization sets

$$L^{-1}A = \widetilde{Q}_1\widetilde{U}$$

so that

$$\mathbf{x}^{(n)} = \widetilde{U}^{-1}\widetilde{Q}_1^T L^{-1}\mathbf{b}.$$

The storage requirement is dominated by that required for L plus that for the design A.

If the problem does have a well determined solution while V is illconditioned or positive semi definite then direct solution of the augmented system

$$\begin{bmatrix} V & -A \\ -A^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} -\mathbf{b} \\ 0 \end{bmatrix}$$

could be appropriate.This matrix is symmetric but indefinite. Solution methods (for example, the Bunch-Kaufman-Parlett algorithm [46]) are available in both dense and sparse cases.

Paige's domain of ideas

The idea here is to provide an algorithm that avoids any dependence on a rank revealing factorization of V, and is numerically stable provided the generalised least squares problem has a well determined solution. The starting point is the constrained form (1.2.4)

$$\min_{\mathbf{x},\mathbf{s}} \mathbf{s}^T\mathbf{s} : \; L\mathbf{s} = A\mathbf{x} - \mathbf{b}.$$

Let $X : R^{n-p} \to R^n$ have full column rank and satisfy

$$X^T A = 0.$$

This implies that $X = Q_2 B$ where $B : R^{n-p} \to R^{n-p}$ has full rank and Q_2 has orthogonal columns. In particular, Q_2 can be constructed via the standard orthogonal factorization

$$A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} U \\ 0 \end{bmatrix}.$$

It follows that $X^T V X$ has full rank $n-p$ under the second order sufficiencey condition (1.1). The main result is the following.

Theorem 2.3 *Equation (1.2.4) can be stated in the equivalent form which does not involve* $\mathbf{x}$

$$\min_{\mathbf{s}} \mathbf{s}^T\mathbf{s} : \ X^TL\mathbf{s} = -X^T\mathbf{b}. \tag{2.3.18}$$

Proof. The necessary conditions for an optimum in (1.2.4) give

$$\begin{bmatrix} \mathbf{s}^T & 0 \end{bmatrix} = \boldsymbol{\mu}^T \begin{bmatrix} L & -A \end{bmatrix}$$

where $\boldsymbol{\mu}$ is the vector of Lagrange multipliers. An immediate consequence is that

$$\boldsymbol{\mu} = Q_2\boldsymbol{\gamma}$$

for some $\boldsymbol{\gamma} \in R^{n-p}$ where Q_2 is the orthogonal complement of the range of A. The values of $\mathbf{s}$ and $\boldsymbol{\gamma}$ can be found from

$$L^TQ_2\boldsymbol{\gamma} = \boldsymbol{s}; \ Q_2^TL\boldsymbol{s} = -Q_2^T\mathbf{b}.$$

It follows from the necessary conditions for (2.3.18) that the optimum is identical to the optimal residual vector in (1.2.4). ■■

The necessary conditions on (2.3.18) give

$$\mathbf{s}^T = \boldsymbol{\lambda}^TX^TL,$$

so the Lagrange multiplier vector $\boldsymbol{\lambda}$ satisfies the nonsingular linear system

$$X^TVX\boldsymbol{\lambda} = -X^T\mathbf{b}.$$

If $X = Q_2$ then the system becomes

$$V_{22}\boldsymbol{\lambda} = -Q_2^T\mathbf{b}$$

. Given $\boldsymbol{\lambda}$ then $\mathbf{s}$ is available and $\mathbf{x}$ can be found by multiplying (1.2.4) by Q_1^T and then solving the resulting upper triangular system. A significant problem is the requirement that V_{22} be formed explicitly as the matrix multiplications required are likely to disturb any sparsity pattern in V. For this reason an alternative approach is sought [85], [86]. Let the orthogonal factorization of $L^TQ_2 : R^{n-p} \rightarrow R^n$ be given by

$$L^TQ_2 = S\begin{bmatrix} R \\ 0 \end{bmatrix}$$

where S is orthogonal and R is upper triangular. Then the constraint equation becomes

$$\begin{bmatrix} R^T & 0 \end{bmatrix} S^T\mathbf{s} = -Q_2^T\mathbf{b}.$$

It follows that the conditions determining the minimum sum of squares are

$$S_1^T\mathbf{s} = -R^{-T}Q_2^T\mathbf{b},$$
$$S_2^T\mathbf{s} = 0.$$

Thus

$$\mathbf{s} = -S_1R^{-T}Q_2^T\mathbf{b},$$

and

$$\mathbf{x}^{(n)} = U^{-1}Q_1^T\left(\mathbf{b} - LS_1R^{-T}Q_2^T\mathbf{b}\right). \tag{2.3.19}$$

This should be compared with (1.2.6). The identity follows on noting that

$$\left(Q_2^TL\right)^+ = S_1R^{-1} = L^TQ_2\left(Q_2^TVQ_2\right)^{-1}$$

where + indicated the generalised inverse. Here it seems that sparsity would be destroyed in forming Q_2^TL, but Paige [86] noted that it does not have to be formed explicitly. He suggests the factorization

$$Q^T\begin{bmatrix}\mathbf{b} & A & L\end{bmatrix}\begin{bmatrix}1 & & \\ & I & \\ & & S\end{bmatrix} = \begin{bmatrix}0 & 0 & L_1 & 0 & \\ \eta & 0 & \mathbf{g}^T & \rho & \\ \mathbf{z} & U^T & L_{21} & \mathbf{m} & L_2\end{bmatrix},$$

where Q, S are orthogonal, $\rho \neq 0$, $\mathbf{z}, \mathbf{m} \in R^p$, $\mathbf{g} \in R^{n-p-1}$, $L_1 : R^{n-p-1} \to R^{n-p-1}$, $L_{21} \in R^{n-p-1} \to R^p$, $L_2, U : R^p \to R^p$, and L_1, L_2 and U^T are lower triangular. Given this we have

$$\begin{aligned}0 &= \begin{bmatrix}\mathbf{b} & A & L\end{bmatrix}\begin{bmatrix}1 \\ -\mathbf{x} \\ \mathbf{s}\end{bmatrix}, \\ &= Q\begin{bmatrix}0 & 0 & L_1 & 0 & \\ \eta & 0 & \mathbf{g}^T & \rho & \\ \mathbf{z} & U^T & L_{21} & \mathbf{m} & L_2\end{bmatrix}\begin{bmatrix}1 & & \\ & I & \\ & & S^T\end{bmatrix}\begin{bmatrix}1 \\ -\mathbf{x} \\ \mathbf{s}\end{bmatrix}, \\ &= \begin{bmatrix}0 & 0 & L_1 & 0 & \\ \eta & 0 & \mathbf{g}^T & \rho & \\ \mathbf{z} & U^T & L_{21} & \mathbf{m} & L_2\end{bmatrix}\begin{bmatrix}1 \\ -\mathbf{x} \\ \mathbf{s}_1 \\ s_2 \\ \mathbf{s}_3\end{bmatrix}.\end{aligned}$$

where $\left\|\begin{bmatrix}\mathbf{s}_1 \\ s_2 \\ \mathbf{s}_3\end{bmatrix}\right\| = \|\mathbf{s}\|$. The key step is to design the factorization and, in particular, the choice of S so that L_1 is forced to be nonsingular. Then the first block row gives

$$L_1\mathbf{s}_1 = 0$$

so that

$$\begin{aligned} \mathbf{s}_1 &= 0, \\ \eta + \rho s_2 &= 0, \\ \mathbf{z} - U^T\mathbf{x} + s_2\mathbf{m} + \mathbf{L}_2\mathbf{s}_3 &= 0. \end{aligned}$$

The minimum value of $\|\mathbf{s}\| = |s_2|$ is achieved when $\mathbf{s}_3 = 0$, and these equations can then be solved for $\mathbf{x}$, ρ. The identifiability conditions A, V_{22} full rank correspond to U, $\begin{bmatrix} L_1 & 0 \\ \mathbf{g}^T & \rho \end{bmatrix}$ nonsingular. Thus $\rho \neq 0$ is required.

To develop the transformation Paige suggests the following sequence of operations:

```
for i=1 to n-1
    for j=1 to min(i,p+1)
        transformation from left. This loop works from right to left.
            mix rows i-j+1,i-j+2 to zero element (i-j+1,p-j+2)
            repeat j
        for j=1 to min(i,p+1)
        transformation from right. This loop works from left to right.
            k=i-min(i,p+1)+j
            mix columns p+k+1,p+k+2 to zero element (k,p+k+2)
            repeat j
        repeat i
```

It is easy to see that these do transform the data array in the desired manner. This is illustrated in the following diagram in which the array is split into data and covariance components and corresponds to $p = 2$, $n = 6$.

$$\begin{array}{cccccccccc} 3 & 2 & 1 & x & 1-3 & & & & & \\ 4 & 3 & 2 & x & x & 2-4 & & & & \\ 5 & 4 & 3 & 1 & x & x & 3-5 & & & \\ x & 5 & 4 & 3 & 2 & x & x & 4-5 & & \\ x & x & 5 & \mathbf{4} & 4 & 3 & x & x & 5 & \\ x & x & x & \mathbf{5} & \mathbf{5} & 5 & 4 & x & x & \end{array}$$

Here the integers in the data array $\begin{bmatrix} \mathbf{b} & A \end{bmatrix}$ indicate the stage at which the corresponding elements are zeroed. In the covariance component L the range in the super diagonal elements indicate the steps in which transformations from the left introduce fill in this position to be removed by the transformations from the right. The subdiagonal elements show the propagation of

fill in the case that L is bidiagonal. Note that it apparently fills the lower triangular matrix showing no apparent sparsity advantage. However, information from the fill only propagates to the right while the column is involved in removing the superdiagonal fill. After this it has no effect on the quantities of interest $\begin{bmatrix} \rho \\ \mathbf{m} \end{bmatrix}$ in determining $\mathbf{x}^{(n)}$. Elements that have no effect in this way are indicated in bold and there is no point in computing them unless additional information on the transformed covariance matrix is required.

Transformations leaving the covariance matrix invariant

The need to develop computational algorithms for the generalised least squares problem raises the question of the possibility of constructing nonsingular transformations of the model equations in order to facilitate their solution

$$\mathbf{r} = A\mathbf{x} - \mathbf{b} \;\rightarrow\; \mathbf{s} = JA\mathbf{x} - J\mathbf{b},$$

where typically JA would be upper triangular, while preserving any useful structure in V. This idea has been exploited in developing the application of orthogonal factorization of the design matrix in the least squares case where $J = Q^T$. In the present context the idea has been explored by, in particular, [41]. It leads naturally to consideration of transformations which leave the covariance matrix invariant corresponding to $Q^TQ = I$ in the least squares case. Because V has the dimension of the data space it must be allowed to be large so the preservation of its structure in cases where it is already in a convenient form is particularly important. The conditions for the solution of the generalised least squares system in the form (1.2.8) which permits certain cases where the covariance matrix V is singular has been summarised in Condition 1.1 as

$$A = Q \begin{bmatrix} U \\ 0 \end{bmatrix},\; Q_2^T V Q_2 \text{ nonsingular.}$$

A convenient starting point that includes the possibility of this generality proves to be equation (1.3.7) defining the Gauss Markov solution operator. The constraints on allowable transformations are:

1. the solution operator must transform by $T \rightarrow TJ^{-1}$ to take account of the transformation $\mathbf{b} \rightarrow J\mathbf{b}$ of the right hand side of the model equations; and

2. the transformed equations must have a symmetric matrix if it is to be interpretable as an augmented matrix .

This suggests the the transformation

$$\begin{aligned}&\begin{bmatrix} T & \Lambda \end{bmatrix} \begin{bmatrix} J^{-1} & \\ & I \end{bmatrix} \begin{bmatrix} J & \\ & I \end{bmatrix} \begin{bmatrix} V & -A \\ -A^T & 0 \end{bmatrix} \begin{bmatrix} J^T & \\ & I \end{bmatrix} \\ &= \begin{bmatrix} 0 & -I \end{bmatrix} \begin{bmatrix} J^T & \\ & I \end{bmatrix} = \begin{bmatrix} 0 & -I \end{bmatrix},\end{aligned}$$

leading to the system

$$\begin{bmatrix} TJ^{-1} & \Lambda \end{bmatrix} \begin{bmatrix} JVJ^T & -JA \\ -A^TJ^T & 0 \end{bmatrix} = \begin{bmatrix} 0 & -I \end{bmatrix}. \tag{2.3.20}$$

Thus the condition for invariance makes the requirement

$$JVJ^T = V \tag{2.3.21}$$

on the covariance matrix [41]. We say that the nonsingular matrix J is V-invariant . This terminology is convenient here, but J^T will be recognised as generating a congruence transformation, and these have been widely studied.

Remark 2.3.4 *The formal objective in the generalised least squares problem is* $\mathbf{r}^TV^{-1}\mathbf{r}$ *while the* V*-invariance condition (2.3.21) is a condition on transformations applied to* V*. This is convenient not only because it is consistent with the extra generality of the Gauss-Markov theory but also because the transformations are required to reduce the linear system if they are to be useful. The argument in [41] goes as follows. Let* J *be nonsingular. Then*

$$\min_{\mathbf{x}} (A\mathbf{x}-\mathbf{b})^T V^{-1} (A\mathbf{x}-\mathbf{b}) = \min_{\mathbf{x}} (JA\mathbf{x}-J\mathbf{b})^T \left(J^{-T}V^{-1}J^{-1}\right) (JA\mathbf{x}-J\mathbf{b})$$

It follows that

$$J^{-T}V^{-1}J^{-1} = V^{-1} \Leftrightarrow JVJ^T = V$$

if V *is nonsingular.*

Remark 2.3.5 *Let* J_1 *and* J_2 *be* V*-invariant. Then* J_1^{-1}*,* J_2^{-1}*,* J_1J_2 *and* J_2J_1 *are* V*-invariant. If* V *is nonsingular then* J_1^T *and* J_2^T *are* V^{-1}*-invariant. If* V *is singular then assume it has nullity* k*. In this case it has the reduced form*

$$V = \begin{bmatrix} 0 & 0 \\ 0 & V_2 \end{bmatrix} \tag{2.3.22}$$

where $V_2 \in R^{n-k} \rightarrow R^{n-k}$*, and* $V_2 \succ 0$*. Here* J *can be* V*-invariant if and only if*

$$J = \begin{bmatrix} J_{11} & \\ J_{21} & J_{22} \end{bmatrix},\ J_{22}V_2J_{22}^T = V_2,$$

and J_{11} *and* J_{22} *are nonsingular.*

Example 2.3.3 *Let $J = I - 2\mathbf{u}\mathbf{v}^T$, then J is an elementary reflector ($J^2 = I,\ \det(J) = -1$) provided $\mathbf{v}^T\mathbf{u} = 1$ and*

$$JVJ^T = V - 2\left(\mathbf{u}\mathbf{v}^T V + V\mathbf{v}\mathbf{u}^T\right) + 4\left(\mathbf{v}^T V\mathbf{v}\right)\mathbf{u}\mathbf{u}^T = V. \tag{2.3.23}$$

Let $\mathbf{v}, \mathbf{v}^T V\mathbf{v} \neq 0$ be given. Then this gives

$$\mathbf{u} = \frac{V\mathbf{v}}{\mathbf{v}^T V\mathbf{v}},\ J = I - 2\frac{V\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T V\mathbf{v}}, \tag{2.3.24}$$

and defines a V-invariant transformation. Also $\mathbf{v}^T\mathbf{u} = 1$ so that J is an elementary reflector. There is a corresponding formula for $\mathbf{v}$ if $\mathbf{u}$ is given.

$$\mathbf{v} = \frac{V^{-1}\mathbf{u}}{\mathbf{u}^T V^{-1}\mathbf{u}}. \tag{2.3.25}$$

If V is singular and $V\mathbf{v} = 0$ then it follows from (2.3.23) that J is V-invariant for arbitrary $\mathbf{u}$. In the special case in which V is given by (2.3.22) then a V-invariant elementary reflector is obtained by setting

$$J = I - 2\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}\begin{bmatrix} \mathbf{v}_1^T & 0 \end{bmatrix},\ \mathbf{u}_1^T\mathbf{v}_1 = 1, \tag{2.3.26}$$

where $\mathbf{v}_1 \neq 0$ is arbitrary.

Example 2.3.4 *Let $D = \begin{bmatrix} d_1 & \\ & d_2 \end{bmatrix}$. Then*

$$J = \begin{bmatrix} \cos\theta & \sqrt{\mu}\sin\theta \\ \frac{\sin\theta}{\sqrt{\mu}} & -\cos\theta \end{bmatrix},\ \mu = \frac{d_1}{d_2}, \tag{2.3.27}$$

satisfies

$$JDJ^T = D.$$

It provides the invariant reflector which corresponds to the plane reflector in linear least squares problems. It can be written in the standard form (2.3.24):

$$J = I - 2\begin{bmatrix} \sin\theta/2 \\ \frac{\cos\theta/2}{\sqrt{\mu}} \end{bmatrix}\begin{bmatrix} \sin\theta/2 & \sqrt{\mu}\cos\theta/2 \end{bmatrix}.$$

Remark 2.3.6 *It is of interest to consider the case $V \succeq 0$ as a limiting case of $V \succ 0$. Let*

$$V_\delta = \begin{bmatrix} \delta W & \\ & V_2 \end{bmatrix}$$

where $V_\delta \in R^k \to R^k$ is a positive diagonal matrix, and δ is a small parameter. Making use of the V_δ-invariant transformation specified in terms of $\mathbf{u}$ *(2.3.25) derived from (2.3.24) gives*

$$\begin{aligned} J_\delta &= I - \frac{2}{\mathbf{u}^T V_\delta^{-1}\mathbf{u}}\mathbf{u}\mathbf{u}^T V_\delta^{-1}, \\ &= I - \frac{2}{\mathbf{u}^T \delta V_\delta^{-1}\mathbf{u}}\mathbf{u}\mathbf{u}^T \delta V_\delta^{-1}, \\ &\to I - \frac{2}{\mathbf{u}_1^T W^{-1}\mathbf{u}_1}\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}\begin{bmatrix} \mathbf{u}_1^T W^{-1} & 0 \end{bmatrix}, \end{aligned}$$

as $\delta \to 0$. Note that the action of this transformation can be considered to be on the scale defined by δ.

The use of J in computational algorithms requires control of the numbers entering and so, by implication, of its norm. The following result is presented in [41] and provides relevant information:

Lemma 2.3 *Let J be an elementary reflector and set $\eta = \|\mathbf{u}\|\,\|\mathbf{v}\| \geq 1$. Then the spectral norm of the V-invariant, elementary reflector is*

$$\|J\| = \eta + \sqrt{\eta^2 - 1}.$$

Thus the norm is determined by η.

Proof. The norm is determined by the largest eigenvalue of

$$J^T J\mathbf{w} = \mu\mathbf{w}.$$

Because J is an elementary reflector this is equivalent to

$$J\mathbf{w} = \mu J^T\mathbf{w}. \tag{2.3.28}$$

If $\mathbf{w}$ is orthogonal to $\mathbf{u}$ and $\mathbf{v}$ then $\mu = 1$. It follows that the eigenvectors associated with the interesting eigenvalues have the form

$$\mathbf{w} = \alpha\mathbf{u} + \beta\mathbf{v}.$$

Equating the coefficients of $\mathbf{u}$ and $\mathbf{v}$ shows that a nontrivial solution of (2.3.28) in the subspace spanned by $\mathbf{u}$, $\mathbf{v}$ is possible only if

$$\begin{vmatrix} -1-\mu & -2\mathbf{v}^T\mathbf{v} \\ 2\mu\mathbf{u}^T\mathbf{u} & 1+\mu \end{vmatrix} = 0.$$

This gives the largest eigenvalue as

$$\begin{aligned}\mu &= 2\eta^2 - 1 + 2\eta\sqrt{\left(\eta^2 - 1\right)}, \\ &= \left(\eta + \sqrt{\left(\eta^2 - 1\right)}\right)^2.\end{aligned}$$

■

Remark 2.3.7 *In the case of (2.3.24) then*

$$\eta = \frac{\|\mathbf{v}\| \, \|V\mathbf{v}\|}{\mathbf{v}^T V \mathbf{v}} \tag{2.3.29}$$

so the relative size of $\mathbf{v}^T V \mathbf{v}$ *is significant. Note that* $\|J\| = 1$ *if and only if* $V = I$*. The corresponding V-invariant transformations are orthogonal.*

To construct the transformed solution operator $TJ^{-1} = \widetilde{T} = \left[\begin{array}{cc} \widetilde{T}_1 & \widetilde{T}_2 \end{array} \right]$ where $\widetilde{T}_1 \in R^p \to R^p$, $\widetilde{T}_2 \in R^{n-p} \to R^p$ consider (2.3.20) in the form

$$\left[\begin{array}{cc} \left[\begin{array}{cc} \widetilde{T}_1 & \widetilde{T}_2 \end{array} \right] & \Lambda \end{array} \right] \left[\begin{array}{c} \left[\begin{array}{ccc} 0 & 0 & 0 \\ 0 & V_{11} & V_{12} \\ 0 & V_{21} & V_{22} \end{array} \right] \quad - \left[\begin{array}{c} U \\ 0 \end{array} \right] \\ - \left[\begin{array}{cc} U^T & 0 \end{array} \right] \qquad\qquad 0 \end{array} \right] = \left[\begin{array}{cc} 0 & -I \end{array} \right],$$

where $U \in R^p \to R^p$ is upper triangular,

$$JA = \left[\begin{array}{c} U \\ 0 \end{array} \right], \; V_2 = \left[\begin{array}{cc} V_{11} & V_{12} \\ V_{21} & V_{22} \end{array} \right], \; V_{11} \in R^{p-k} \to R^{p-k}, \; V_{22} \in R^{n-p} \to R^{n-p}.$$

This gives the sequence of equations

$$\begin{aligned} \widetilde{T}_1 \left[\begin{array}{cc} 0 & 0 \\ 0 & V_{11} \end{array} \right] + \widetilde{T}_2 \left[\begin{array}{cc} 0 & V_{21} \end{array} \right] - \Lambda U^T &= 0, \\ \widetilde{T}_1 \left[\begin{array}{c} 0 \\ V_{12} \end{array} \right] + \widetilde{T}_2 V_{22} &= 0, \\ \widetilde{T}_1 U &= I. \end{aligned}$$

These can be solved to give the quantities of interest

$$\widetilde{T}_1 = U^{-1}, \; \widetilde{T}_2 = -U^{-1} \left[\begin{array}{c} 0 \\ V_{12} \end{array} \right] V_{22}^{-1}, \tag{2.3.30}$$

$$\Lambda = U^{-1} \left\{ \left[\begin{array}{cc} 0 & 0 \\ 0 & V_{11} \end{array} \right] - \left[\begin{array}{c} 0 \\ V_{12} \end{array} \right] V_{22}^{-1} \left[\begin{array}{cc} 0 & V_{21} \end{array} \right] \right\} U^{-T}, \tag{2.3.31}$$

$$\mathbf{x}^{(n)} = U^{-1} \left[\begin{array}{cc} I & - \left[\begin{array}{c} 0 \\ V_{12} \end{array} \right] V_{22}^{-1} \end{array} \right] J\mathbf{b}. \tag{2.3.32}$$

Both the form of the solution given by (2.3.32) and the construction of the factorization when V is non singular follows the same procedure as that in the orthogonal case. At the start of the i'th step assume a partial reduction to upper triangular form

$$J_{i-1}A = \begin{bmatrix} U_{i-1} & U_{12}^i \\ 0 & A_i \end{bmatrix}$$

has been obtained. Let the i'th (pivotal) column be

$$J_{i-1}A\mathbf{e}_i = \begin{bmatrix} \mathbf{U}_1^i \\ \mathbf{a}_i \end{bmatrix}.$$

Then $\mathbf{v}$ is chosen such that the elementary, V-invariant reflector J^i satisfies

$$J^i \begin{bmatrix} \mathbf{U}_1^i \\ \mathbf{a}_i \end{bmatrix} = \left[I - \frac{2}{\mathbf{v}^T V \mathbf{v}} V\mathbf{v}\mathbf{v}^T \right] \begin{bmatrix} \mathbf{U}_1^i \\ \mathbf{a}_i \end{bmatrix} = \begin{bmatrix} \mathbf{U}_1^i \\ \gamma\mathbf{e}_1 \end{bmatrix},$$

giving the updated partial factorization matrix $J_i \leftarrow J^i J_{i-1}$. An appropriate choice of $\mathbf{v}$ is obtained by setting

$$V\mathbf{v} = \begin{bmatrix} 0 \\ \mathbf{a}_i - \gamma\mathbf{e}_1 \end{bmatrix}. \tag{2.3.33}$$

To fix γ note that, as J^i is a reflector,

$$\begin{bmatrix} \mathbf{U}_1^i \\ \mathbf{a}_i \end{bmatrix} = \left[I - \frac{2}{\mathbf{v}^T V \mathbf{v}} V\mathbf{v}\mathbf{v}^T \right] \begin{bmatrix} \mathbf{U}_1^i \\ \gamma\mathbf{e}_1 \end{bmatrix}.$$

Taking the scalar product with $\mathbf{v}$ gives

$$\mathbf{v}^T \begin{bmatrix} \mathbf{U}_1^i \\ \mathbf{a}_i \end{bmatrix} = -\mathbf{v}^T \begin{bmatrix} \mathbf{U}_1^i \\ \gamma\mathbf{e}_1 \end{bmatrix}$$

so that

$$\mathbf{v}^T \begin{bmatrix} 2\mathbf{U}_1^i \\ \mathbf{a}_i + \gamma\mathbf{e}_1 \end{bmatrix} = 0. \tag{2.3.34}$$

This gives a quadratic equation for γ so this stage of the factorization can be completed, at least formally.

Remark 2.3.8 *In this generality there are apparent disadvantages in this approach. Perhaps the most obvious is the need to solve linear equations with matrix V in order to compute $\mathbf{v}$. This appears to engage exactly the problem that it is hoped to avoid. Also, if off-diagonal elements of V are not all 0 then $\mathbf{v}$ is potentially a full vector so there is coupling involving $\mathbf{u}_1^i$ in the calculation of J^i. This appears an unwelcome complication.*

The transformation of the design matrix to upper triangular form is made easier by the structural assumption that V has the reduced form (2.3.22) where $V_2 \in R^{n-k} \to R^{n-k}$, and

$$V_2 = \text{diag}\left\{\nu_{k+1}, \nu_{k+2}, \cdots, \nu_n\right\},\ 0 < \nu_{k+1} \leq \nu_{k+2} \leq \cdots \leq \nu_n. \qquad (2.3.35)$$

If V is a positive semidefinite matrix in general form then it is suggested in [100] that this could be achieved by:

1. First scaling V so that the nonzero diagonal elements of the transformed matrix are unity. Here $V \to S\overline{V}S$ where $S = \text{diag}\left\{V_{11}^{1/2}, \cdots, V_{nn}^{1/2}\right\}$.

2. Then making a diagonal pivoting (rank revealing) Cholesky factorization (2.3.4) applied to $\overline{V}$ to construct a matrix of known factorization and rank which closely approximates $\overline{V}$

$$P\overline{V}P^T = LDL^T.$$

   Note that the ordering achieved by this factorization gives the elements of D in decreasing order of magnitude which is the inverse of that required in (2.3.35).

3. Next transforming the problem to one with covariance matrix D

$$\mathbf{r} \to L^{-1}PS^{-1}\mathbf{r} \Rightarrow A \to L^{-1}PS^{-1}A,\ \mathbf{b} \to L^{-1}PS^{-1}\mathbf{b}.$$

   Here Remark 2.3.1 which indicates that illconditioning in V tends to be concentrated in D rather than in the triangular factor L should be noted.

4. Finally permuting the covariance into increasing order with permutation matrix Q

$$D \to Q^TDQ,\ \mathbf{r} \to Q^T\mathbf{r}.$$

Let $V_2 = \begin{bmatrix} V_{11} & \\ & V_{22} \end{bmatrix}$, $V_{11} = \text{diag}\left\{\nu_{k+1}, \cdots, \nu_{i-1}\right\}$, $V_{22} = \text{diag}\left\{\nu_i, \cdots, \nu_n\right\}$, where the partitioning is chosen to correspond to the i'th factorization step with $i > k$. Then

$$J^i = \begin{bmatrix} I_k & & \\ & I_1 & \\ & & I_2 - \frac{2}{\mathbf{v}^TV_{22}\mathbf{v}}V_{22}\mathbf{v}\mathbf{v}^T \end{bmatrix}$$

is V-invariant where I_k is the $k \times k$ unit matrix and I_1 and I_2 are unit matrices conformable with V_{11} and V_{22} respectively. The calculation of J^i can be specialised in much the same way as in (2.3.33) by setting

$$V_{22}\mathbf{v} = \mathbf{a}_i - \gamma\mathbf{e}_1.$$

In this case (2.3.34) reduces to

$$\left(\mathbf{a}_i - \gamma\mathbf{e}_1\right)^T V_{22}^{-1}\left(\mathbf{a}_i + \gamma\mathbf{e}_1\right) = 0,$$

so that

$$\gamma = \theta\sqrt{\left\{\nu_i \mathbf{a}_i^T V_{22}^{-1}\mathbf{a}_i\right\}}. \tag{2.3.36}$$

The argument used here mirrors the corresponding argument in specifying a Householder transformation. This suggests that to minimize cancellation the choice $\theta = -\operatorname{sgn}\left(\mathbf{a}_i\right)_1$ is appropriate (note (2.3.37) below). An interesting feature is the appearance of the term $\nu_i V_{22}^{-1}$ which means that γ is independent of the scale of V. To evaluate the denominator in the transformation:

$$\begin{aligned}\mathbf{v}^T V_{22}\mathbf{v} &= \left(\mathbf{a}_i - \gamma\mathbf{e}_1\right)^T V_{22}^{-1}\left(\mathbf{a}_i - \gamma\mathbf{e}_1\right), \\ &= \mathbf{a}_i^T V_{22}^{-1}\mathbf{a}_i - 2\gamma\left(\mathbf{a}_i\right)_1/\nu_i + \gamma^2/\nu_i, \\ &= 2\left|\gamma\right|\left(\left|\gamma\right| + \left|\left(\mathbf{a}_i\right)_1\right|\right)/\nu_i. \end{aligned} \tag{2.3.37}$$

This leads to a convenient form

$$J = I - \frac{\left(\mathbf{a} - \gamma\mathbf{e}_1\right)\left(\mathbf{a} - \gamma\mathbf{e}_1\right)^T N^i}{\left|\gamma\right|\left(\left|\gamma\right| + \left|\left(\mathbf{a}_i\right)_1\right|\right)}$$

where

$$N^i = \nu_i V_{22}^{-1} = \left\{1, \frac{\nu_i}{\nu_{i+1}}, \cdots, \frac{\nu_i}{\nu_n}\right\},$$

is the diagonal matrix of scaled weights. These are nonincreasing and bounded by 1.

In this case the stability criterion (2.3.29) is given by

$$\eta = \frac{\left\|V_{22}^{-1}\left(\mathbf{a}_i - \gamma\mathbf{e}_1\right)\right\|\left\|\mathbf{a}_i - \gamma\mathbf{e}_1\right\|}{\left(\mathbf{a}_i - \gamma\mathbf{e}_1\right)^T V_{22}^{-1}\left(\mathbf{a}_i - \gamma\mathbf{e}_1\right)}.$$

The denominator is given by (2.3.37), and the terms in the numerator can be estimated by

$$\begin{aligned}\left\|V_{22}^{-1}\left(\mathbf{a}_i - \gamma\mathbf{e}_1\right)\right\| &\geq \left(\left|\gamma\right| + \left|\left(\mathbf{a}_i\right)_1\right|\right)/\nu_i, \text{ and} \\ \left\|\mathbf{a}_i - \gamma\mathbf{e}_1\right\| &\geq \left\|\mathbf{a}_i\right\|,\end{aligned}$$

where the first inequality is most accurate if $\nu_i \ll \nu_{i+1}$, so that

$$\eta \geq \frac{\|\mathbf{a}_i\|}{2\,|\gamma|}.$$

It follows that the transformation will have large norm corresponding to an illconditioned case if

$$|\gamma| \ll \|\mathbf{a}_i\|\,. \tag{2.3.38}$$

Remark 2.3.9 *Control over conditioning is possible by using column pivoting . Equation(2.3.36) suggests that possible candidate values of* $|\gamma|$ *in (2.3.38) are given by* $\rho_i(A,j) = \sum_{q=i}^{n} N_q^i A_{qj}^2$, $j = i, i+1, \cdots, p$. *Then (2.3.38) suggests that the pivotal column at the current stage should be chosen to maximize* $\rho_i(A,j)/\|\mathbf{a}_j\|^2$. *However, while it is not difficult to economise on the calculation of the* ρ_i *using simple recursions (see the example below) the* $\|\mathbf{a}_j\|^2$ *are not invariant under* V*-invariant transformations if* $V \neq I$ *and so cannot be economised in the same way in general. Thus it is usual to base the selection of the pivotal column on the size of the* $\rho_i(A,j)$ *alone.*

If $k > 0$ so that the form (2.3.22) is assumed for V with V_2 diagonal, then the key mapping

$$J\mathbf{a}_l = \gamma \mathbf{e}_l,\ l \leq k,$$

needs to make use of the second family of V-invariant transformations (2.3.26). There is no loss of generality in choosing $l = 1$ corresponding to the first step of the factorization. Insight is provided by considering it as the limit as $\delta \to 0$ of the case

$$\nu_1 = \nu_2 = \cdots = \nu_k = \delta,\ \nu_{k+1} \gg \delta. \tag{2.3.39}$$

Here

$$\begin{aligned} \lim_{\delta \to 0} \nu_1 V^{-1} &= \begin{bmatrix} I_k & \\ & 0 \end{bmatrix}, \\ \lim_{\delta \to 0} |\gamma| &= \left\|\mathbf{a}_1^1\right\|_2, \end{aligned} \tag{2.3.40}$$

where $\mathbf{a}_1 = \begin{bmatrix} \mathbf{a}_1^1 \\ \mathbf{a}_1^2 \end{bmatrix}$ to match the partitioning of V. The resulting transformation matrix $J = I - 2\mathbf{c}\mathbf{d}^T$ has the generic form (2.3.26) with

$$\begin{aligned} \sqrt{2}\mathbf{c} &= \left(\mathbf{a}_1 + \operatorname{sgn}\left(\left(\mathbf{a}_1^1\right)_1\right)\left\|\mathbf{a}_1^1\right\|_2 \mathbf{e}_1\right) / \left\|\mathbf{a}_1^1\right\|_2, \\ \sqrt{2}\mathbf{d} &= \begin{bmatrix} \mathbf{a}_1^1 + \operatorname{sgn}\left(\left(\mathbf{a}_1^1\right)_1\right)\left\|\mathbf{a}_1^1\right\|_2 \mathbf{e}_1 \\ 0 \end{bmatrix} / \left(\left\|\mathbf{a}_1^1\right\|_2 + \left|\left(\mathbf{a}_1^1\right)_1\right|\right). \end{aligned}$$

This transformation will have large elements if

$$\left\|\mathbf{a}_1^1\right\|_2 \ll \|\mathbf{a}_1\|_2 .$$

This is the limiting case of (2.3.38) which characterizes illconditioning in the other class of transformations.

Example 2.3.5 *The possibility of column pivoting in solution methods for the standard least squares problem has been indicated (Remark 2.3.2). It is unlikely to be a critical factor in the numerical solution of problems with well conditioned design matrices. However, this need no longer be true for the V-invariant solution methods when the structure of V supports different scales. An important example is provided by equality constrained problems. In this case Remark 2.3.9 indicates that the column to be chosen as pivotal column at the i'th step is the one that maximizes $\rho_i(A,j) = \sum_{q=i}^{n+m} N_q^i A_{qj}^2$, $j = i, i+1, \cdots, p$, where m is the number of equality constraints and N^i is the scaled diagonal weighting matrix. To illustrate the requirement consider the design matrix $A_2 \in R^p \to R^n$ given by*

$$\begin{aligned} (A_2)_{*1} &= S\mathbf{e}, \\ (A_2)_{i(2j)} &= S_i \cos(2\pi j (i-1) h), \\ (A_2)_{i(2j+1)} &= S_i \sin(2\pi j (i-1) h), \\ i = 1,2,\cdots,n,\ j &= 1,2,\cdots,k, \end{aligned}$$

where $p = 2k+1$, $S = \operatorname{diag}\left\{1/\sqrt{2}, 1, \cdots, 1, 1/\sqrt{2}\right\}$, $h = 1/(n-1)$. The columns of A_2 are orthogonal and similarly scaled so a least squares problem with A_2 as design is very well conditioned. Let the constraint matrix $A_1 \in R^2 \to R^p$ be given by

$$(A_1)_{1*} = \mathbf{e}^T,\ (A_1)_{2*} = \mathbf{e}^T - p\mathbf{e}_{k+1}^T.$$

Then the constrained least squares problem

$$\min_{\mathbf{s},\mathbf{x}} \mathbf{s}^T\mathbf{s};\ \begin{bmatrix} 0 & \\ & I \end{bmatrix}\mathbf{s} = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}\mathbf{x} - \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix},$$

with

$$\begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} A_1\mathbf{e} \\ A_2\mathbf{e} + \boldsymbol{\varepsilon} \end{bmatrix},\ \varepsilon_i = \sin(2\pi(k+1)(i-1)h),\ i = 1,4,\cdots,n,$$

has the solution $\mathbf{x} = \mathbf{e}$. This follows because $\boldsymbol{\varepsilon}^T A_2 = 0$ so $\mathbf{e}$ solves the unconstrained problem and also satisfies the constraints. However, the leading

2×2 submatrix of A_1 is singular if $k \geq 2$. This causes a breakdown of the V-invariant factorization at the second step as a consequence of the form of N^2 (2.3.40) if column interchanges are not used. In the particular case corresponding to $n = 8$, $p = 5$, the ordering resulting from column pivoting is $\{3, 2, 1, 4, 5\}$. The first interchange is a consequence of the largest coefficient in the second constraint, and it succeeds in forcing the leading 2×2 submatrix to be nonsingular.

Column sums can be computed recursively by noting that

$$\rho_i(A, j) = \left(\rho_{i-1}(A, j) - A^2_{(i-1)j}\right) / N_i^{i-1},\ i = 2, \cdots, p-1,\ j = i, \cdots, p. \tag{2.3.41}$$

However, the advantage of this recurrence is economy, and it needs to be monitored carefully. In the above example $N^i = \nu_i \operatorname{diag}\{\nu_i, \nu_{i+1}, \cdots \nu_{n+m}\}^{-1}$ changes character when i increases from $i = 2$, corresponding to the last zero element in V, to $i = 3$ corresponding to the first nonzero. Here $N^2 = \operatorname{diag}\{1, 0, \cdots, 0\}$, while $N^3 = \operatorname{diag}\{1, 1, \cdots, 1\}$. At this point the perturbation approach (2.3.39) is not satisfactory and the $\rho_i(A, j)$ must be recomputed. However, there is also some potential for cancellation in the recurrence (2.3.41), and this must be watched.

Does the use of the LDL^T factorization of V make sense?

A rank-revealing Cholesky factorization of V has the form

$$V \to L \operatorname{diag}\{D_n, D_{n-1}, \cdots, D_1\} L^T$$

where the diagonal pivoting ensures that

$$D_n \geq D_{n-1} \geq \cdots \geq D_1.$$

This order is the reverse of that required here. As a consequence it must be inverted to be used to construct the factorization of the design matrix based on V-invariant transformations. Conditions for the rank revealing factorization to provide a satisfactory basis for implementing the V-invariant factorization are [46]

$$\begin{aligned}
&\Delta_1 = \operatorname{diag}\{D_1, D_2, \cdots, D_k\} \text{ small},\\
&D_k \ll D_{k+1},\\
&\Delta_2 = \operatorname{diag}\{D_{k+1}, \cdots, D_n\} \text{ not small},\\
&\text{where } k \leq p.
\end{aligned}$$

It would be expected that the values $\{D_1, D_2, \cdots, D_k\}$ could have high relative error as a result of cancellation as they are computed at the final stages of the Cholesky decomposition. Does this matter? The following argument suggests strongly that it does not. First note that the case $D = \text{diag}\{0, \cdots, 0, D_{k+1}, \cdots, D_n\}$ corresponds to the equality constrained problem (1.2.7):

$$\min_{\mathbf{x}} \mathbf{s}^T\mathbf{s};\ \begin{bmatrix} 0 & \\ & \Delta_2^{1/2} \end{bmatrix} \mathbf{s} = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}.$$

This is the limiting problem as $\lambda \to \infty$ associated with the penalised objective

$$\min_{\mathbf{x}} \left\{ \mathbf{r}_2^T \Delta_2^{-1} \mathbf{r}_2 + \lambda \mathbf{r}_1^T \mathbf{r}_1 \right\};\ \mathbf{r} = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} \tag{2.3.42}$$

which has the alternative form

$$\min_{\mathbf{x}} \mathbf{s}^T\mathbf{s};\ \begin{bmatrix} \lambda^{-1/2} I & \\ & \Delta_2^{1/2} \end{bmatrix} \mathbf{s} = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}.$$

From penalty function theory we expect that $\|\mathbf{x}(\lambda) - \widehat{\mathbf{x}}\| = O(1/\lambda),\ \lambda \to \infty$ [29]. To justify this note that the necessary conditions for the penalty problem (2.3.42) are

$$\mathbf{r}_2^T \Delta_2^{-1} A_2 + \lambda \mathbf{r}_1^T A_1 = 0.$$

If we set $\tau = 1/\lambda$ and define

$$\tau \mathbf{u} = A_1\mathbf{x} - \mathbf{b}_1\ (= \mathbf{r}_1)$$

then we can find differential equations defining a trajectory satisfied by $\mathbf{x}(\tau), \mathbf{u}(\tau)$ for small τ by differentiating these relations. This gives

$$\begin{aligned} A_2^T \Delta_2^{-1} A_2 \frac{d\mathbf{x}}{d\tau} + A_1^T \frac{d\mathbf{u}}{d\tau} &= 0, \\ A_1 \frac{d\mathbf{x}}{d\tau} - \tau \frac{d\mathbf{u}}{d\tau} &= \mathbf{u}. \end{aligned}$$

The matrix of this system is nonsingular for τ small enough provided $M = \begin{bmatrix} A_2^T \Delta_2^{-1} A_2 & A_1^T \\ A_1 & 0 \end{bmatrix}$ has full rank. This is a weaker condition than both A_1 and A_2 having full rank. This ensures that the initial value problem for the differential equation system has a well determined solution. Thus it is possible to integrate the system back to $\tau = 0$ starting from the initial condition provided by the solution of the generalised least squares problem for a finite value of τ. The Taylor series expansion of the solution is well

defined at $\tau = 0$. The first two terms give $\mathbf{x}(\tau) = \mathbf{x}(0) + O(\tau)$ corresponding to a small perturbation, and this is just what had to be proved. The size of the derivatives, hidden by the big-O notation, depends on the conditioning of M.

The generalisation of this result to the case where Δ_1 is a matrix of small elements resulting from a rank revealing Cholesky factorization follows directly. It is only necessary to associate with Δ_1 a scaled family of matrices $\tau\widetilde{\Delta}_1$where $\widetilde{\Delta}_1 = \Delta_1 / \|\Delta_1\|$ (in the above argument $\widetilde{\Delta}_1 = I$) and to apply the homotopy argument to the corresponding modified penalty problem. Let $D = \operatorname{diag}\{\Delta_1, \Delta_2\}$. We conclude that the equality constrained problem obtained by setting $\Delta_1 = 0$ has a well defined solution which differs from that based on the LDL^T factorization by $O(\|\Delta_1\|)$.

2.3.4 Updating and downdating methods

Surgery based on orthogonal transformations

This section considers methods for the eficient modification of factorizations developed for the solution of the least squares and generalised least squares problems associated with the design data $[A, \mathbf{b}]$ both to permit the addition of new data (updating) and to enable the deletion of data considered unsatisfactory (downdating) [33]. In the case of the linear least squares problem these operations when applied to the associated Cholesky decomposition of the normal matrix or the orthogonal factorization of the design can be given a unified treatment by considering the relation

$$\begin{bmatrix} L_1 & \\ \mathbf{v}^T & \gamma \end{bmatrix} \begin{bmatrix} L_1^T & \mathbf{v} \\ & \gamma \end{bmatrix} = \begin{bmatrix} L_2 & \mathbf{u} \\ & 1 \end{bmatrix} \begin{bmatrix} L_2^T & \\ \mathbf{u}^T & 1 \end{bmatrix}, \tag{2.3.43}$$

where L_1 and L_2 are lower triangular matrices. This requires that the following equations hold:

$$L_1 L_1^T = L_2 L_2^T + \mathbf{u}\mathbf{u}^T, \tag{2.3.44}$$

$$L_1 \mathbf{v} = \mathbf{u}, \tag{2.3.45}$$

$$\mathbf{v}^T \mathbf{v} + \gamma^2 = 1. \tag{2.3.46}$$

To add a new row to the design let $L_2 L_2^T = A^T A$ then (2.3.44) shows that $\mathbf{u}^T$ corresponds to the new row in the augmented design and that L_1 is the updated Cholesky factor of the normal matrix. The computation of L_1 can be carried out by constructing an orthogonal matrix Q such that

$$Q^T \begin{bmatrix} L_2^T & \\ \mathbf{u}^T & 1 \end{bmatrix} = \begin{bmatrix} L_1^T & \mathbf{v} \\ & \gamma \end{bmatrix}.$$

Here Q can be built up by the sequence of plane rotations $\left(Q = \prod_{j=p}^{1} R_j\right)$ where R_j mixes rows $j, p+1$ of the target matrix and eliminates the $(p+1, j)$ element.

Downdating is a little more complicated. In this case L_1 is the given factor, L_2 is the target, and $\mathbf{u}^T$ is the row of the design to be removed. It follows from (2.3.45) that $\mathbf{v} = L_1^{-1}\mathbf{u}$, and, for consistency, $\gamma^2 = 1 - \mathbf{v}^T\mathbf{v}$ must be positive. This is just the condition that

$$L_2 L_2^T = L_1 L_1^T - \mathbf{u}\mathbf{u}^T = L_1 L_1^T - L_1 \mathbf{v}\mathbf{v}^T L_1^T \succ 0.$$

Note that there is the possibility of cancellation in this step [102]! The requirement this time is the construction of an orthogonal matrix such that

$$Q^T \begin{bmatrix} L_1^T & \mathbf{v} \\ & \gamma \end{bmatrix} = \begin{bmatrix} L_2^T & \\ \mathbf{u}^T & 1 \end{bmatrix}.$$

The computation can be carried out using plane rotations to zero the components of $\mathbf{v}$ in an order that preserves the upper triangular form of L_1^T. In terms of the component plane rotations the transformation has the form $Q = \prod_{j=1}^{p} R_j$ where R_j mixes rows $j, p+1$ in order to introduce a zero in position $(j, p+1)$. For example, in the first step, rows $p, p+1$ are mixed to introduce a zero in position $(p, p+1)$ and fill in position $(p+1, p)$. Next, rows $p-1, p+1$ are mixed to zero the $(p-1, p+1)$ element. This introduces a nonzero in position $(p+1, p-1)$ but preserves the upper triangular form of L_1.

Another application is to certain trust region methods used for nonlinear least squares and maximum likelihood problems (subsection 4.2.3). Here the solution of the system

$$\left[A^T A + \gamma^2 D^T D\right] \mathbf{v} = A^T \mathbf{b}, \tag{2.3.47}$$

where $A : R^p \to R^n$, and $D : R^p \to R^p$ is upper triangular, is required for a sequence of values of γ. For each γ this corresponds to the least squares problem

$$\min_{\mathbf{v}} \mathbf{r}^T\mathbf{r}; \quad \mathbf{r} = \begin{bmatrix} A \\ \gamma D \end{bmatrix} \mathbf{v} - \begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix}. \tag{2.3.48}$$

Let $A = Q \begin{bmatrix} U \\ 0 \end{bmatrix}$, where Q is orthogonal. Then (2.3.47) reduces to

$$\left[U^T U + \gamma^2 D^T D\right] \mathbf{v} = U^T Q_1^T \mathbf{b} = U^T \mathbf{c}_1. \tag{2.3.49}$$

This is requivalent to the least squares problem

$$\min_{\mathbf{v}} \mathbf{s}^T\mathbf{s}; \quad \mathbf{s} = \begin{bmatrix} U \\ \gamma D \end{bmatrix} - \begin{bmatrix} \mathbf{c}_1 \\ 0 \end{bmatrix}. \tag{2.3.50}$$

This has reduced an $(n+p)\times p$ problem to a $2p\times p$ problem once the initial factorization of the design has been carried out. Here orthogonal factorization to solve the reduced least squares problem gives

$$\begin{bmatrix} U \\ \gamma D \end{bmatrix} = Q' \begin{bmatrix} U' \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{c}_1' \\ \mathbf{c}_2' \end{bmatrix} = Q'^T \begin{bmatrix} \mathbf{c}_1 \\ 0 \end{bmatrix}.$$

It is convenient to use elementary reflectors to generate Q', and these can be organised to preserve the upper triangular structure in U while eliminating the elements of D and any associated fill. This is shown in (2.3.51) below.

$$\begin{bmatrix} x & x & x & x \\ & x & x & x \\ & & x & x \\ & & & x \\ 1 & 2 & 3 & 4 \\ & 2 & 3 & 4 \\ & & 3 & 4 \\ & & & 4 \end{bmatrix} \rightarrow \begin{bmatrix} x & x & x & x \\ & x & x & x \\ & & x & x \\ & & & x \\ 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 \\ & & 0 & 0 \\ & & & 0 \end{bmatrix} \tag{2.3.51}$$

Here the numbers indicate which of the first four rows of U is mixed with the current row of D using an elementary reflector to eliminate the indicated element. In each column the eliminations are typically performed in increasing row order. However, usually this would not be essential. In the notation used above the contribution to Q' from eliminations in row k, $k = p+1, \cdots, 2p$ is $\prod_{j=p+1}^{k} R_{jk}$ where R_{jk} mixes rows j and k and eliminates the element in the (k,j) position. For example, in the second step in the above example, the second row is used to zero elements in the $(5,2)$, reflector R_{25}, and $(6,2)$ reflector R_{26}, positions. The exact row order in which the elements in D are eliminated does not affect the result (in exact arithmetic).

Remark 2.3.10 *In the trust region application considered in Chapter 4, subsection 4.2.3, D is diagonal. However, this additional structure does not help because the fill generated in the target row $p+1$ by the first step which eliminates the element $(p+1,1)$ introduces fill in positions $(p+1,2),(p+1,3),\cdots,(p+1,p)$. Similar fill patterns are introduced by subsequent stages and the result is that the same basic elimination pattern has to be followed as in the case D upper triangular.*

Remark 2.3.11 *Each sweeping out of the D matrix costs $O(p^3)$ arithmetic operations. This costs more than the previously discussed examples of variable addition and deletion which cost $O(p^2)$ operations. However, the additional cost is negligable if $n >> p$.*

Surgery based on V-invariant methods

Some generalisation of the above techniques to the application of V-invariant transformations in the solution of generalised least squares problems is possible. Let the D-invariant factorization of the design matrix A be

$$JA = \begin{bmatrix} I \\ 0 \end{bmatrix} U,\ J \begin{bmatrix} D_1 & \\ & \widetilde{D}_1 \end{bmatrix} J^T = \begin{bmatrix} D_1 & \\ & \widetilde{D}_1 \end{bmatrix},$$

where A has full rank p, and $D = \begin{bmatrix} D_1 & \\ & \widetilde{D}_1 \end{bmatrix}$ is diagonal with positive elements. Then, using the properties summarised in Remark 2.3.5,

$$\begin{aligned} A^T D^{-1} A &= A^T J^T J^{-T} D^{-1} J^{-1} J A, \\ &= U^T \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} D_1 & \\ & \widetilde{D}_1 \end{bmatrix}^{-1} \begin{bmatrix} I \\ 0 \end{bmatrix} U, \\ &= U^T D_1^{-1} U. \end{aligned} \tag{2.3.52}$$

If the additional information is

$$\mathbf{r}_2 = A_2\mathbf{x} - \mathbf{b}_2 = \mathbf{y}_2 + \boldsymbol{\varepsilon}_2,\ \mathcal{V}\{\boldsymbol{\varepsilon}_2\} = D_2,$$

where, if necessary, the data $\begin{bmatrix} A_2 & \mathbf{b}_2 \end{bmatrix} : R^{p+1} \to R^q$, $q \geq 1$, has been manipulated to give a diagonal covariance D_2, and $\boldsymbol{\varepsilon}_2$ is assumed to uncorrelated with the original data. In this setting the system corresponding to (2.3.43) is

$$\begin{bmatrix} U_1^T & \\ Z & G \end{bmatrix} \begin{bmatrix} D_1^{-1} & \\ & D_2^{-1} \end{bmatrix} \begin{bmatrix} U_1 & Z^T \\ & G^T \end{bmatrix} = \begin{bmatrix} U_2^T & A_2^T \\ & I \end{bmatrix} \begin{bmatrix} D_1^{-1} & \\ & D_2^{-1} \end{bmatrix} \begin{bmatrix} U_2 & \\ A_2 & I \end{bmatrix},$$

where U_1, U_2 are upper triangular, and G can be taken as lower triangular. The relations which must be satisfied are:

$$U_1^T D_1^{-1} U_1 = U_2^T D_1^{-1} U_2 + A_2^T D_2^{-1} A_2, \tag{2.3.53}$$

$$Z D_1^{-1} U_1 = D_2^{-1} A_2, \tag{2.3.54}$$

$$Z D_1^{-1} Z^T + G D_2^{-1} G^T = D_2^{-1}. \tag{2.3.55}$$

Equation (2.3.53) shows that U_1 is the required factor in the updating step. It can be computed by the D-invariant factorization

$$J_u \begin{bmatrix} U_2 \\ A_2 \end{bmatrix} = \begin{bmatrix} U_1 \\ 0 \end{bmatrix},\ J_u \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix} J_u^T = \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix},$$

where the elementary transformations are organised to preserve the upper triangular structure of U_2. If J_u is made up of products of elementary D-invariant transformations J_k, $J_u = \prod_k J_k$,

$$J_k = I - 2 \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T & \mathbf{v}_2^T \end{bmatrix}, \; J_k \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix} J_k^T = \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix}$$

then $\widetilde{J}_k$,

$$\widetilde{J}_k = I - 2 \begin{bmatrix} \mathbf{u}_1 \\ 0 \\ \mathbf{u}_2 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T & 0 & \mathbf{v}_2^T \end{bmatrix},$$

leaves

$$\widetilde{D} = \begin{bmatrix} D_1 & & \\ & \widetilde{D}_1 & \\ & & D_2 \end{bmatrix}$$

invariant. Let $\widetilde{J}_u = \prod_k \widetilde{J}_k$, then it follows that

$$\widetilde{J}_u \begin{bmatrix} J & \\ & I \end{bmatrix} \begin{bmatrix} A \\ A_2 \end{bmatrix} = \widetilde{J}_u \begin{bmatrix} U_2 \\ 0 \\ A_2 \end{bmatrix} = \begin{bmatrix} U_1 \\ 0 \\ 0 \end{bmatrix}$$

provides the updated factorization.

Downdating requires that $U_1^T D_1^{-1} U_1 - A_2^T D_2^{-1} A_2 \succ 0$. This turns out to be equivalent to the consistency of the defining relations. Equation (2.3.55) requires $D_2^{-1} - Z D_1^{-1} Z^T \succ 0$ in order for G to be defined satisfactorily - it can then be computed by an obvious modification of the Cholesky factorization. This equivalence is the subject of the next two results.

Lemma 2.4 *Let $D_1 : R^p \to R^p$, $D_2 : R^q \to R^q$ be positive diagonal matrices, and $Z : R^p \to R^q$ be given. Then $D_1 - Z^T D_2 Z \succ 0$ if and only if $D_2^{-1} - Z D_1^{-1} Z^T \succ 0$.*

Proof. Let

$$M = \begin{bmatrix} D_1 & Z^T \\ Z & D_2^{-1} \end{bmatrix}.$$

Then M has the block symmetric factorization

$$M = \begin{bmatrix} I & \\ Z D_1^{-1} & I \end{bmatrix} \begin{bmatrix} D_1 & \\ & D_2^{-1} - Z D_1^{-1} Z^T \end{bmatrix} \begin{bmatrix} I & D_1^{-1} Z^T \\ & I \end{bmatrix}.$$

It follows that $M \succ 0$ if and only if $D_2^{-1} - Z D_1^{-1} Z^T \succ 0$. Now let

$$\widetilde{M} = \begin{bmatrix} D_2^{-1} & Z \\ Z^T & D_1 \end{bmatrix}$$

be the symmetric (row and column) permutation of M. Then $\widetilde{M} \succ 0$ if and only if $M \succ 0$. Also, it has the symmetric block decomposition

$$\widetilde{M} = \begin{bmatrix} I & \\ Z^T D_2 & I \end{bmatrix} \begin{bmatrix} D_2^{-1} & \\ & D_1 - Z^T D_2 Z \end{bmatrix} \begin{bmatrix} I & D_2 Z \\ & I \end{bmatrix}.$$

It follows that $\widetilde{M} \succ 0$ if and only if $D_1 - Z^T D_2 Z \succ 0$. ■

Corollary 2.1 *The third relation (2.3.55) has a solution $GD_2^{-1}G^T \succ 0$ if and only if $U_1^T D_1^{-1} U_1 - A_2^T D_2^{-1} A_2 \succ 0$.*

Proof. The defining relations give

$$U_1^T D_1^{-1} U_1 - A_2^T D_2^{-1} A_2 = U_1^T D_1^{-1} \left\{ D_1 - Z^T D_2 Z \right\} D_1^{-1} U_1.$$

Thus $U_1^T D_1^{-1} U_1 - A_2^T D_2^{-1} A_2 \succ 0$ if and only if $D_1 - Z^T D_2 Z \succ 0$. The proposition now follows from the preceding. ■

The downdating step requires the construction of the D-invariant factorization

$$J_d \begin{bmatrix} U_1 & Z^T \\ & G^T \end{bmatrix} = \begin{bmatrix} U_2 & \\ A_2 & I \end{bmatrix}, \; J_d \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix} J_d^T = \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix}.$$

The organisation used in the linear least squares downdating can be followed here, and the D-invariant plane reflectors (2.3.27) possess the flexibility needed to preserve the upper triangular form of U_1.

Remark 2.3.12 *As $\widetilde{D}_1$ does not enter this calculation, the ordering of the elements of D_2 relative to those of $\widetilde{D}_1$ is not a factor in the stability of this reduction. Permutation matrices can be used for local reordering and then the restoring of the original order. Let P be the permutation matrix, and D^* be the reordered diagonal matrix. Then*

$$P^T P D P^T P = P^T D^* P, \; J^* D^* J^{*T} = J^*, \; J^* P A = \begin{bmatrix} U \\ 0 \end{bmatrix}.$$

The reduction sequence becomes:

$$\begin{aligned} \mathbf{r}^T D^{-1} \mathbf{r} &= \mathbf{r}^T \left(P^T P D P^T P \right)^{-1} \mathbf{r}, \\ &= \mathbf{r}^T P^T D^{*-1} P \mathbf{r}, \\ &= \mathbf{r}^T P^T J^{*T} J^{*-T} D^{*-1} J^{*-1} J^* P \mathbf{r}, \\ &= \mathbf{r}^T P^T J^{*T} P \left(P^T D^* P \right)^{-1} P^T J^* P \mathbf{r} \\ &= \mathbf{r}^T J^T D^{-1} J \mathbf{r}, \end{aligned}$$

where

$$J = P^T J^* P.$$

Let $T^* = TJ^{-1}$ *where* T *is the Gauss-Markov operator defined in (1.3.7). Then* T^* *satisfies*

$$\left[\begin{array}{cc} T^* & \Lambda \end{array}\right] \left[\begin{array}{cc} D & -JA \\ -A^T J^T & 0 \end{array}\right] = \left[\begin{array}{cc} 0 & -I \end{array}\right]$$

so that

$$T^* P^T \left[\begin{array}{c} U \\ 0 \end{array}\right] = I,\ T^* D - \Lambda \left[\begin{array}{cc} U^T & 0 \end{array}\right] P = 0.$$

Setting $\left[\begin{array}{cc} \widetilde{T}_1 & \widetilde{T}_2 \end{array}\right] = T^* P^T$ *gives*

$$\begin{aligned} \mathbf{x} = T\mathbf{b} &= T^* J\mathbf{b}, \\ &= \left[\begin{array}{cc} \widetilde{T}_1 & \widetilde{T}_2 \end{array}\right] P\left(P^T J^* P\right)\mathbf{b}, \\ &= \left[\begin{array}{cc} U^{-1} & 0 \end{array}\right] J^* P\mathbf{b}. \end{aligned}$$

This just corresponds to the solution for the permuted diagonal matrix D^**. However, when* A *is structured, the elimination is carried out in several steps to avoid fill. If the scope of the permutations for the different steps overlap then care is required.*

2.3.5 Algorithms for filtering and smoothing

The dynamical system equations provide useful examples to illustrate the developments indicated in the previous subsection. Computational algorithms for the Kalman filter have a long history during which they have gradually evolved to take account of improvements in computational technology. There are two ways of representing the model equations. The first is a rearrangement of (1.4.3), (1.4.4) and can be written

$$\min_{\mathbf{x}} \mathbf{r}^T V_D^{-1} \mathbf{r};\ \mathbf{r} = X\mathbf{x} - \mathbf{y} \tag{2.3.56}$$

where

$$X = \begin{bmatrix} I & & & & & \\ H_1 & & & & & \\ -X_1 & I & & & & \\ & H_2 & & & & \\ & -X_2 & I & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & & -X_{n-1} & I \\ & & & & & H_n \end{bmatrix}, \tag{2.3.57}$$

$$V_D = \text{diag}\left\{S_{1|0}, V_1, R_1, V_2, \cdots, R_{n-1}, V_n\right\}, \tag{2.3.58}$$

$$\mathbf{y}^T = \left\{\mathbf{x}_{1|0}^T, \mathbf{y}_1^T, 0, \mathbf{y}_2^T, \cdots, 0, \mathbf{y}_n^T\right\}. \tag{2.3.59}$$

The solution of this problem gives $\mathbf{x}_{k|n}$, $k = 1, 2, \cdots, n$, so the action of the interpolation smoother is implicit in the solution process and quantities such as the innovations and intermediate covariances must be found by special intervention. Note that X is block bidiagonal and that V_D is block diagonal. Thus this problem has considerable sparsity structure and preservation of this structure has to be an important component of any solution procedure.

The second approach poses the problem in a recursive form corresponding to the steps in the Kalman filter (1.4.5), (1.4.6) as described in Section 1.4. Here the sparsity in (2.3.56) is preserved implicitly. There are two cases to consider. If one starts with the knowledge of $\mathbf{x}_{k|k-1}, S_{k|k-1}$ then the generalised least squares problem

$$\min_{\mathbf{x}_k, \mathbf{x}_{k+1}} \mathbf{r}^T \,\text{diag}\left\{S_{k|k-1}, V_k, R_k\right\}^{-1}\mathbf{r}, \tag{2.3.60}$$

$$\mathbf{r} = \begin{pmatrix} \mathbf{x}_k - \mathbf{x}_{k|k-1} \\ H_k\mathbf{x}_k - \mathbf{y}_k \\ -X_k\mathbf{x}_k + \mathbf{x}_{k+1} \end{pmatrix},$$

has the solution $\mathbf{x}_{k|k}, \mathbf{x}_{k+1|k}$. This form is appropriate for a forward recursion. The alternative form starts with $\mathbf{x}_{k|k}, S_{k|k}$, and considers the generalised least squares problem

$$\min_{\mathbf{x}_k, \mathbf{x}_{k+1}} \mathbf{r}^T \,\text{diag}\{S_{k|k}, R_k, V_{k+1}\}^{-1}\mathbf{r}, \tag{2.3.61}$$

$$\mathbf{r} = \begin{pmatrix} \mathbf{x}_k - \mathbf{x}_{k|k} \\ -X_k\mathbf{x}_k + \mathbf{x}_{k+1} \\ -\mathbf{y}_{k+1} + H_{k+1}\mathbf{x}_{k+1} \end{pmatrix}.$$

The solution gives $\mathbf{x}_{k|k+1}, \mathbf{x}_{k+1|k+1}$. This form is of interest because it updates $\mathbf{x}_{k|k}$ to take account of the new data. Thus it provides a route to determining the interpolation smoother.

Three possibilities are considered here.

1. Reduction of the problem (2.3.56) to the least squares problem

$$\min_{\mathbf{x}} \mathbf{s}^T\mathbf{s};\ \mathbf{s} = L_D^{-1}X\mathbf{x} - L_D^{-1}\mathbf{y},$$

where $L_DL_D^T = V_D$ is the standard square root Cholesky factorization of V_D. If this is solved using orthogonal factorization of the modified design $L_D^{-1}X$ then the information filter of Paige and Saunders results [87]. The requirement to form L_D^{-1} explicitly makes this approach sensitive to ill conditioning of V_D. An important example occurs as a result of the occurence of small elements in V_D corresponding to accurately known components of the state vector . This can occur as a result of the gain of information through the innovation sequence.

2. Application of a V_D-invariant transformation either directly to the design (this reqires that solution of linear equations with matrix V_D is cheap), or following an $L_IDL_I^T$ factorization of V_D. Here L_I is lower triangular with unit diagonal. Because the rank revealing Cholesky can destroy sparsity the algorithm developed here works with the recursive form (2.3.60). The algorithm is due to [99]. This approach has the potential to cope with zero diagonal blocks in V_D corresponding to accurately known quantities.

3. Solution of the recursive system (2.3.61) using a version of the Paige approach. This leads to the square root covariance filter algorithm of Osborne and Prvan [83]. This approach is distinguished in this class of methods by incorporating an effective square root implementation of the interpolation smoother.

The information filter

Let the Cholesky factorization of the covariance matrix V_D be

$$V_D = L_DL_D^T,\ L_D = \mathrm{diag}\left\{L_1, M_1, L_2, M_2, \cdots, L_n, M_n\right\}.$$

Then the transform of the design matrix (2.3.57) can be written

$$L_D^{-1}X = \begin{bmatrix} L_1^{-1} & & & & & & \\ \widetilde{H}_1 & & & & & & \\ -\widetilde{X}_1 & L_2^{-1} & & & & & \\ & \widetilde{H}_2 & & & & & \\ & -\widetilde{X}_2 & L_3^{-1} & & & & \\ & & \cdots & \cdots & \cdots & & \\ & & & & & -\widetilde{X}_{n-1} & L_n^{-1} \\ & & & & & & \widetilde{H}_n \end{bmatrix}, \qquad (2.3.62)$$

where $\widetilde{H}_i = M_i^{-1}H_i$, $\widetilde{X}_i = L_{i+1}^{-1}X_i$. The algorithm now proceeds as an ordinary least squares problem, and the standardization of the error term to an iid vector should be noted. The modified design matrix is block bidiagonal with rectangular blocks

$$L_{i(i-1)} = \begin{bmatrix} -\widetilde{X}_{i-1} \\ 0 \end{bmatrix}, \; L_{ii} = \begin{bmatrix} L_i^{-1} \\ \widetilde{H}_i \end{bmatrix}, \; i = 1, 2, \cdots, n.$$

The result of the orthogonal factorization is a block bidiagonal, upper triangular matrix. For example, the first step takes

$$\begin{bmatrix} L_1^{-1} & 0 \\ \widetilde{H}_1 & 0 \\ -\widetilde{X}_1 & L_2^{-1} \end{bmatrix} \rightarrow \begin{bmatrix} U_1 & U_{12} \\ 0 & 0 \\ 0 & \widetilde{U}_2 \end{bmatrix} \rightarrow P \begin{bmatrix} U_1 & U_{12} \\ 0 & \widetilde{U}_2 \\ 0 & 0 \end{bmatrix}$$

where the tilda indicates that $\widetilde{U}_2$ will enter the factorization at the next stage, and P is a permutation matrix interchanging block rows to build up the upper triangular structure. There are several ways of organising this computation, but one starts by combining L_1^{-1} and $\widetilde{H}_1$ to sweep out the elements in the $\widetilde{H}_1$ position followed by the use of the modified matrix in the L_1^{-1} position to sweep out $-\widetilde{X}_1$. This approach introduces the structured elements in the second block column. The next step takes

$$\begin{bmatrix} \widetilde{U}_2 & 0 \\ \widetilde{H}_2 & 0 \\ -\widetilde{X}_2 & L_3^{-1} \end{bmatrix} \rightarrow \begin{bmatrix} U_2 & U_{23} \\ 0 & 0 \\ 0 & \widetilde{U}_3 \end{bmatrix}.$$

This shows that the algorithm has the recursive pattern characteristic of a filter algorithm. If a back substitution is performed at the k'th stage then the result is

$$\left\{\mathbf{x}_{1|k}^T, \mathbf{x}_{2|k}^T, \cdots, \mathbf{x}_{k|k}^T, \mathbf{x}_{k+1|k}^T\right\}.$$

This shows the dependence of past state value estimates $\mathbf{x}_{i|k}$ on all the current data characteristic of a smoothing algorithm. Because the errors are now iid, it follows that

$$\mathcal{V}\left\{\mathbf{x}_{k+1} - \mathbf{x}_{k+1|k}\right\} = \widetilde{U}_k^{-1}\widetilde{U}_k^{-T} = S_{k+1|k}. \tag{2.3.63}$$

At this point it should be clear that there is recursive procedure in operation as a consequence of the block bidiagonal structure of the design matrix, and that each of the steps indicated above solves the generalised least squares problem (2.3.60). The appearance of the inverse in (2.3.63) provides the origin of the term information filter. This algorithm provided an interesting alternative to the then standard algorithms (for example [7]) both in recurring the inverse and in being one of the first to introduce a technology based on orthogonal factorization in the filtering context. The information filter has the useful attribute that it can run when a diffuse prior corresponding to $S_{1|0} = \infty$ is used to initialise the computation. In this case $L_1^{-1} = 0$.

A V_D-invariant filter

The information filter has the distinguishing characteristic of coping with diffuse initial conditions, but makes the explicit assumption that the component covariances in V_D are all nonsingular. This assumption can be weakened by using V-invariant transformations. The advantage of this approach is that certain cases in which exact information is available on certain state variable combinations ($R_i, S_{i|i-1}$ singular) and/or certain variable combinations are observed without error (V_i singular) can be treated in routine fashion. The complication with the use of the V-invariant methods to factorise (2.3.57) is that reordering may be required both in the rank revealing Cholesky factorization ($PV_DP^T = LDL^T$), and in the reordering of the components of D to meet the requirement that the elements of D are increasing. This reordering has the potential to destroy the block bidiagonal structure. However, we have seen that the block bidiagonal structure has the recursively defined form of the filter given by (2.3.60) associated with it. Each of the recursively defined subproblems can be solved by V-invariant techniques and potential fill is much less serious. Let

$$W_i = \text{diag}\left\{S_{i|i-1}, V_i, R_i\right\} = K_i\overline{W}_iK_i$$

define the scaled covariance and diagonal scaling matrix. Then the rank revealing Cholesky is

$$P_i\overline{W}_iP_i^T = L_iD_iL_i^T,$$

the transformed design becomes

$$\left[\begin{array}{cc} A_i & \mathbf{b}_i \end{array}\right] = Q_i L_i^{-1} P_i K_i^{-1} \left[\begin{array}{c} \left[\begin{array}{cc} I & 0 \\ H_i & 0 \\ -X_i & I \end{array}\right] \left[\begin{array}{c} \mathbf{x}_{i|i-1} \\ \mathbf{y}_i \\ 0 \end{array}\right] \end{array}\right],$$

$\overline{D}_i = Q_i D_i Q_i^T$ is the required diagonal weighting matrix with increasing elements, and Q_i is the corresponding permutation matrix. Let the $\overline{D}_i$-invariant transformation give

$$J_i A_i C_i = \left[\begin{array}{c} U_i \\ 0 \end{array}\right], \; J_i \overline{D}_i J_i^T = \overline{D}_i,$$

where $C_i = \prod_{j=1}^{p-1} P_j$ is the permutation matrix summarising the column interchanges. Then

$$C_i^T \left[\begin{array}{c} \mathbf{x}_{i|i} \\ \mathbf{x}_{i+1|i} \end{array}\right] = U_i^{-1} \left[\begin{array}{cc} I & 0 \end{array}\right] J_i \mathbf{b}.$$

It follows from (1.3.28) that the variance is given by

$$\mathcal{V}\left\{\left[\begin{array}{c} \mathbf{x}_i - \mathbf{x}_{i|i} \\ \mathbf{x}_{i+1} - \mathbf{x}_{i+1|i} \end{array}\right]\right\} = C_i U_i^{-1} \left[\begin{array}{cc} I & 0 \end{array}\right] \overline{D}_i \left[\begin{array}{c} I \\ 0 \end{array}\right] U_i^{-T} C_i^T.$$

A covariance filter

The algorithm developed here is based on the solution of the recursively defined problem (2.3.61). A particular feature is the ready availability of the interpolation gain. Here the design matrix $X : R^{2p} \to R^{2p+m}$ and data vector $\mathbf{y} \in R^{2p+m}$ are given by

$$X = \left[\begin{array}{cc} I & 0 \\ -X_{i-1} & I \\ 0 & H_i \end{array}\right], \; \mathbf{y} = \left[\begin{array}{c} \mathbf{x}_{i-1|i-1} \\ 0 \\ \mathbf{y}_i \end{array}\right].$$

The corresponding generalised least squares problem, making use of the Cholesky factorization of the covariance, is

$$\min \mathbf{s}^T \mathbf{s}; \; \text{diag}\left\{S_{i-1|i-1}^{1/2}, R_{i-1}^{1/2}, V_i^{1/2}\right\} \mathbf{s} = X\mathbf{x} - \mathbf{y}. \tag{2.3.64}$$

Let $Z \in R^m \to R^{2p+m}$,

$$Z = \left[\begin{array}{c} X_{i-1}^T H_i^T \\ H_i^T \\ -I \end{array}\right].$$

Then Z has full column rank m and satisfies $Z^T X = 0$. Thus it can be used in a (nonorthogonal) Paige formulation of the generalised least squares problem as in (2.3.18), and this yields the alternative form

$$\min \mathbf{s}^T\mathbf{s};\ Z^T \operatorname{diag}\left\{S_{i-1|i-1}^{1/2}, R_{i-1}^{1/2}, V_i^{1/2}\right\}\mathbf{s} = -Z^T\mathbf{y} = -\widetilde{\mathbf{y}}_i, \tag{2.3.65}$$

where $\widetilde{\mathbf{y}}_i = H_i X_{i-1}\mathbf{x}_{i-1|i-1} - \mathbf{y}_i$ is the innovation at the current step (1.4.5). Let

$$\operatorname{diag}\left\{S_{i-1|i-1}^{T/2}, R_{i-1}^{T/2}, V_i^{T/2}\right\} Z = Q \begin{bmatrix} U \\ 0 \end{bmatrix}.$$

Then the minimum norm solution (generalised inverse) of (2.3.65) gives

$$\widehat{\mathbf{s}} = -Q_1 U^{-T} Z^T \mathbf{y} = -Q_1 U^{-T}\widetilde{\mathbf{y}}. \tag{2.3.66}$$

Recovery of the state variable $\mathbf{x}$ given $\widehat{\mathbf{s}}$ is simple in this case because the first $2p$ rows of X form a lower triangular matrix with unit diagonal.

The implementation suggested by Osborne and Prvan [83] is based on two stages of orthogonal factorization corresponding to the prediction and update steps in the filter respectively. The prediction transformation Q^p is defined by

$$\begin{bmatrix} S_{i-1|i-1}^{T/2} X_{i-1}^T & S_{i-1|i-1}^{T/2} \\ R_{i-1}^{T/2} & 0 \end{bmatrix} = \begin{bmatrix} Q_1^p & Q_2^p \end{bmatrix} \begin{bmatrix} S_{i|i-1}^{T/2} & K_{i1} \\ 0 & K_{i2} \end{bmatrix}, \tag{2.3.67}$$

while the corresponding update transformation Q^u is given by

$$\begin{bmatrix} S_{i|i-1}^{T/2} & S_{i|i-1}^{T/2} H_i^T \\ 0 & V_i^{T/2} \end{bmatrix} = \begin{bmatrix} Q_1^u & Q_2^u \end{bmatrix} \begin{bmatrix} S_{i|i}^{T/2} & 0 \\ W_i^T & \left[V_i + H_i S_{i|i-1} H_i^T\right]^{T/2} \end{bmatrix}. \tag{2.3.68}$$

Computation of Q^P is a straightforward orthogonal times upper triangular factorization. However, Q^u requires a patterned factorization using elementary reflectors. This is illustrated below for the case $p = 3$, $m = 2$.

$$\begin{bmatrix} x & x & x & x & x \\ . & x & x & x & x \\ . & . & x & x & x \\ . & . & . & x & x \\ . & . & . & . & x \end{bmatrix} \rightarrow \begin{bmatrix} x & x & x & 0_5 & 0_6 \\ . & x & x & 0_3 & 0_4 \\ . & . & x & 0_1 & 0_2 \\ 5 & 3 & 1 & x & x \\ 6 & 4 & 2 & . & x \end{bmatrix}$$

Here the subscripted zeros indicate the stage at which the zero is introduced, while numbers indicate the stage at which fill first occurs. For example, rows

3 and 4 are combined in the first step to introduce a zero in location 0_1 and fill in location 1. The second step combines rows 3 and 5 introducing fill in location 2 and a zero in location 0_2 while preserving the zero in location 0_1.

To verify the identities in (2.3.67) and (2.3.68) we have

$$\begin{bmatrix} X_{i-1}S_{i-1|i-1}^{1/2} & R_{i-1}^{1/2} \\ S_{i-1|i-1}^{1/2} & 0 \end{bmatrix} \begin{bmatrix} S_{i-1|i-1}^{T/2}X_{i-1}^T & S_{i-1|i-1}^{T/2} \\ R_{i-1}^{T/2} & 0 \end{bmatrix} = \begin{bmatrix} X_{i-1}S_{i-1|i-1}X_{i-1}^T + R_{i-1} & X_{i-1}S_{i-1|i-1} \\ S_{i-1|i-1}X_{i-1}^T & S_{i-1|i-1} \end{bmatrix},$$

and

$$\begin{bmatrix} S_{i|i-1}^{1/2} & \\ K_{i1}^T & K_{i2}^T \end{bmatrix} \begin{bmatrix} S_{i|i-1}^{T/2} & K_{i1} \\ 0 & K_{i2} \end{bmatrix} = \begin{bmatrix} S_{i|i-1} & S_{i|i-1}^{1/2}K_{i1} \\ K_{i1}^T S_{i|i-1}^{T/2} & K_{i1}^T K_{i1} + K_{i2}^T K_{i2} \end{bmatrix}$$

for the Q^p transformation. This shows that $S_{i|i-1}$ is updated correctly, and that

$$K_{i1} = S_{i|i-1}^{-1/2} X_{i-1} S_{i-1|i-1} = S_{i|i-1}^{1/2} A_i^T$$

where A_i is the interpolation gain (1.4.12) required for evaluating the smoothed values in (1.4.11). For the Q^u transformation we have

$$\begin{bmatrix} S_{i|i-1}^{1/2} & \\ H_i S_{i|i-1}^{1/2} & V_i^{1/2} \end{bmatrix} \begin{bmatrix} S_{i|i-1}^{T/2} & S_{i|i-1}^{T/2}H_i^T \\ 0 & V_i^{T/2} \end{bmatrix} = \begin{bmatrix} S_{i|i-1} & S_{i|i-1}H_i^T \\ H_i S_{i|i-1} & V_i + H_i S_{i|i-1} H_i^T \end{bmatrix},$$

and

$$\begin{bmatrix} S_{i|i}^{1/2} & W_i \\ & \left[V_i + H_i S_{i|i-1} H_i^T\right]^{1/2} \end{bmatrix} \begin{bmatrix} S_{i|i}^{T/2} & 0 \\ W_i^T & \left[V_i + H_i S_{i|i-1} H_i^T\right]^{T/2} \end{bmatrix} = \begin{bmatrix} S_{i|i} + W_i W_i^T & W_i \left[V_i + H_i S_{i|i-1} H_i^T\right]^{T/2} \\ \left[V_i + H_i S_{i|i-1} H_i^T\right]^{1/2} W_i^T & \left[V_i + H_i S_{i|i-1} H_i^T\right] \end{bmatrix}.$$

Thus

$$W_i = S_{i|i-1} H_i^T \left[V_i + H_i S_{i|i-1} H_i^T\right]^{-T/2},$$

showing that

$$S_{i|i} = S_{i|i-1} - S_{i|i-1} H_i^T \left[V_i + H_i S_{i|i-1} H_i^T\right]^{-1} H_i S_{i|i-1}$$

as required.

The factors Q_1 and U needed to compute $\widehat{\mathbf{s}}$ defined in (2.3.66) can be found by combining (2.3.67) and (2.3.68). The argument can proceed as follows

$$\begin{aligned}
\operatorname{diag}\left\{S_{i-1|i-1}^{T/2}, R_{i-1}^{T/2}, V_i^{T/2}\right\} Z &= \begin{bmatrix} S_{i-1|i-1}^{T/2} X_{i-1} H_i^T \\ R_{i-1}^{T/2} H_i^T \\ -V_i^{T/2} \end{bmatrix}, \\
&= \begin{bmatrix} Q_1^P & Q_2^P & 0 \\ 0 & 0 & -I_m \end{bmatrix} \begin{bmatrix} S_{i|i-1}^{T/2} H_i^T \\ 0 \\ V_i^{T/2} \end{bmatrix}, \\
&= \begin{bmatrix} Q_1^P & Q_2^P & 0 \\ 0 & 0 & -I_m \end{bmatrix} \begin{bmatrix} I_p & 0 & 0 \\ 0 & 0 & I_p \\ 0 & I_m & 0 \end{bmatrix} \begin{bmatrix} S_{i|i-1}^{T/2} H_i^T \\ V_i^{T/2} \\ 0 \end{bmatrix}, \\
&= \begin{bmatrix} Q_1^P & Q_2^P & 0 \\ 0 & 0 & -I_m \end{bmatrix} \begin{bmatrix} I_p & 0 \\ 0 & 0 \\ 0 & I_m \end{bmatrix} Q_2^u \left[V_i + H_i S_{i|i-1} H_i^T\right]^{T/2}, \\
&= \begin{bmatrix} Q_1^P & 0 \\ 0 & -I_m \end{bmatrix} Q_2^u \left[V_i + H_i S_{i|i-1} H_i^T\right]^{T/2}.
\end{aligned}$$

This gives

$$\widehat{\mathbf{s}} = \begin{bmatrix} Q_1^P & 0 \\ 0 & -I_m \end{bmatrix} Q_2^u \left[V_i + H_i S_{i|i-1} H_i^T\right]^{-1/2} \widetilde{\mathbf{y}}_i.$$

The updated state vector can be found from the system

$$\begin{bmatrix} I_p & \\ -X_{i-1} & I_p \end{bmatrix} \begin{bmatrix} \mathbf{x}_{i-1|i} \\ \mathbf{x}_{i|i} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{i-1|i-1} \\ 0 \end{bmatrix} + \begin{bmatrix} S_{i-1|i-1}^{1/2} & 0 \\ 0 & R_{i-1}^{1/2} \end{bmatrix} \widehat{\mathbf{s}}_-,$$

where

$$\widehat{\mathbf{s}}_- = \begin{bmatrix} Q_1^P & 0 \end{bmatrix} Q_2^u \left[V_i + H_i S_{i|i-1} H_i^T\right]^{-1/2} \widetilde{\mathbf{y}}_i.$$

Verification that the update is correct can be carried out directly. For example

$$\begin{aligned}
\mathbf{x}_{i|i} &= X_{i-1}\mathbf{x}_{i-1|i-1} + \begin{bmatrix} X_{i-1} S_{i-1|i-1}^{1/2} & R_{i-1}^{1/2} \end{bmatrix} \widehat{\mathbf{s}}_-, \\
&= \mathbf{x}_{i|i-1} + \begin{bmatrix} S_{i-1|i-1}^{1/2} & 0 \end{bmatrix} Q_2^u \left[V_i + H_i S_{i|i-1} H_i^T\right]^{-1/2} \widetilde{\mathbf{y}}_i, \\
&= \mathbf{x}_{i|i-1} + W_i \left[V_i + H_i S_{i|i-1} H_i^T\right]^{-1/2} \widetilde{\mathbf{y}}_i, \\
&= \mathbf{x}_{i|i-1} + S_{i|i-1} H_i^T \left[V_i + H_i S_{i|i-1} H_i^T\right]^{-1} \widetilde{\mathbf{y}}_i.
\end{aligned}$$

The only condition on the success of this stage of the filter is that $V_i + H_i S_{i|i-1} H_i^T$ has a bounded inverse.

Initialisation of the covariance filter

Values of the initial state $\mathbf{x}_{1|0}$ and state covariance $S_{1|0}$ are required to initialise the covariance filter. The case requiring attention is $\mathbf{x}_{1|0} = 0$, $S_{1|0} = \gamma^2 I$ which is used to approximate the idealisation of a diffuse prior by taking γ^2 large. The generalisation in which the diffuse prior applies only to a subset of the initial state information is also of interest. The limiting process as $\gamma^2 \to \infty$ is known to be well defined in both these cases [4], and the limit is of interest. It turns out that $S_{i|i}$ settles down to bounded values once sufficient observational data has been accumulated. However, the removal of large values in the successive state covariance matrices comes about by a sequence of steps in which cancellation can occur (1.4.10). The concern is that this cancellation could introduce numerical problems for the covariance filter. The Paige and Saunders form of the information filter potentially avoids this as $\gamma \to \infty$ by working with the inverse of $S_{1|0}$ in the case that this is bounded . Subject to this requirement, it provides one approach to initialising the covariance filter.

An alternative is to follow the asymptotic approach of [4] until the leading terms in γ have disappeared. The filter must be followed explicitly for k steps where $p \leq km < p + m$ in order to study the asymptotic behaviour for large γ. The case considered here corresponds to $km = p$. It is convenient to define the following quantities.

$$\begin{aligned}
X_{sl} &= X_{s-1} X_{(s-1)l},\ 1 \leq l < s,\ X_{ll} = I, \\
G_s &= X_{s1}^T H_s^T,\ G_s : R^m \to R^p, \\
G^1 &= G_1 = H_1^T, \\
G^s &= \left[\begin{array}{ccc} G_1 & \cdots & G_s \end{array} \right],\ G^s : R^{sm} \to R^p, \\
\mathbf{y}^1 &= \mathbf{y}_1, \\
\mathbf{y}^{sT} &= \left[\begin{array}{ccc} \mathbf{y}_1^T & \cdots & \mathbf{y}_s^T \end{array} \right],\ \mathbf{y}^s : R \to R^{sm}.
\end{aligned}$$

Let $T_s : R^p \to R^p$ be an orthogonal projection onto the span of G^s, $s = 1, 2, \cdots, k$, $(T_s G^s = G^s, T_0 = 0)$ and define

$$W_s = (I_p - T_{s-1})\, G_s Z_s^{-1}, W_s : R^m \to R^p, \tag{2.3.69}$$

where $Z_s : R^m \to R^m$ is upper triangular and satisfies

$$Z_s^T Z_s = G_s^T\, (I_p - T_{s-1})\, G_s.$$

Conditions under which Z_s is invertible are given subsequently. They are linked closely to the increment in rank(T_s) when $s \to s + 1$. Then

$$W_s^T W_s = Z_s^{-T} G_s^T\, (I_p - T_{s-1})\, G_s Z_s^{-1} = I_m,$$

and

$$W_s W_s^T = W_s W_s^T W_s W_s^T,$$

so that $W_s W_s^T$ is an orthogonal projector. Also

$$W_s W_s^T G_s = (I_p - T_{s-1}) G_s Z_s^{-1} Z_s^{-T} G_s^T (I_p - T_{s-1}) G_s = (I_p - T_{s-1}) G_s,$$

so that

$$\left[T_{s-1} + W_s W_s^T\right] G^s = \left[\; T_{s-1} G^{s-1} \; \middle| \; T_{s-1} G_s + (I_p - T_{s-1}) G_s \; \right].$$

This gives an updating formula for T_s:

$$T_s = T_{s-1} + W_s W_s^T \tag{2.3.70}$$

as the orthogonal projectors on the right hand side map onto orthogonal subspaces.

The quantities defined above permit an orthogonal times upper-triangular factorization of G^s to be defined recursively. Let

$$G^{s-1} = Q_1^{(s-1)} U_{s-1},$$

then

$$G^s = Q_1^s U_s = \begin{bmatrix} Q_1^{(s-1)} & W_s \end{bmatrix} \begin{bmatrix} U_{s-1} & Q^{(s-1)T} G_s \\ & Z_s \end{bmatrix}. \tag{2.3.71}$$

That Q_1^s has orthogonal columns follows because $Q_1^{(s-1)} Q_1^{(s-1)T} = T_{s-1}$ and $T_{s-1} W_s = 0$ by definition. The factorization follows from the definition (2.3.69) of W_s. The interesting calculation is

$$Q_1^{(s-1)} Q_1^{(s-1)T} G_s + W_s Z_s = T_{s-1} G_s + (I_p - T_{s-1}) G_s.$$

In order to simplify the following calculations it is assumed that m divides p ($k = p/m$) where $H_s : R^p \to R^m$, that $\operatorname{rank}(G^s) = sm$ so that each Z_s is invertible, and that each X_s has full rank. The aim is to show that under these conditions it takes exactly k steps for the information from the observation equations to build up to the point that the variances of the state variables settle down as $\gamma \to \infty$.

Lemma 2.5 *The following relations are valid for large γ^2, and $s = 2, \cdots, k$:*

$$S_{s|s-1} = X_{s1} \left\{ \gamma^2 (I_p - T_{s-1}) + E_{s-1} \right\} X_{s1}^T, \tag{2.3.72}$$

$$S_{s|s} = X_{s1} \left\{ \gamma^2 (I_p - T_s) + F_{s-1,} \right\} X_{s1}^T, \tag{2.3.73}$$

$$\mathbf{x}_{s|s} = \mathbf{x}_{s|s}^0 + \frac{1}{\gamma^2} \mathbf{x}_{s|s}^1 + \cdots \tag{2.3.74}$$

where E_s and F_s are defined recursively in (2.3.77) and (2.3.76) respectively. $E_1 = 0$, while E_s, $s < k$, and F_{s-1} are positive definite and $O(1)$ as $\gamma^2 \to \infty$. The leading term in the asymptotic dependence of the state variables is given by

$$\mathbf{x}^0_{s|s} = X_{s1}\boldsymbol{\xi}_s,\ \boldsymbol{\xi}_s = \left[G^{sT}\right]^+ \mathbf{y}^s. \tag{2.3.75}$$

Proof. This proceeds by induction. We assume the form of $S_{s|s-1}$ and evaluate $S_{s|s}$ using (1.4.10). This gives

$$S_{s|s} = S_{s|s-1} - S_{s|s-1}H_s^T \left\{H_s S_{s|s-1} H_s^T + V_s\right\}^{-1} H_s S_{s|s-1},$$

where

$$\begin{aligned}
S_{s|s-1}H_s^T &= X_{s1}\left\{\gamma^2\left(I_p - T_{s-1}\right) + E_{s-1}\right\} X_{s1}^T H_s^T \\
&= X_{s1}\left\{\gamma^2\left(I_p - T_{s-1}\right) + E_{s-1}\right\} G_s, \\
&= X_{s1}\left\{\gamma^2 W_s Z_s + E_{s-1}G_s\right\}, \\
H_s S_{s|s-1} H_s^T + V_s &= G_s^T\left\{\gamma^2 W_s Z_s + E_{s-1}G_s\right\} + V_s, \\
&= \left\{\gamma^2 G_s^T\left(I_p - T_{s-1}\right) W_s Z_s + G_s^T E_{s-1} G_s\right\} + V_s, \\
&= \left\{\gamma^2 Z_s^T Z_s + G_s^T E_{s-1} G_s\right\} + V_s,
\end{aligned}$$

and

$$\begin{aligned}
&S_{s|s-1}H_s^T\left\{H_s S_{s|s-1} H_s^T + V_s\right\}^{-1} H_s S_{s|s-1} \\
&\qquad = \left\{\begin{array}{c} X_{s1}\left\{\gamma^2 W_s Z_s + E_{s-1}G_s\right\} Z_s^{-1} \\ \left\{\gamma^2 I + Z_s^{-T}\left(G_s^T E_{s-1} G_s + V_s\right) Z_s^{-1}\right\}^{-1} \\ Z_s^{-T}\left\{\gamma^2 Z_s^T W_s^T + G_s^T E_{s-1}^T\right\} X_{s1}^T \end{array}\right\}, \\
&\qquad = \frac{1}{\gamma^2} X_{s1} \left\{\begin{array}{c} \left\{\gamma^2 W_s + E_{s-1}G_s Z_s^{-1}\right\} \\ \left\{I - \frac{1}{\gamma^2}L_s + \frac{1}{\gamma^4}L_s^2\right\} \\ \left\{\gamma^2 W_s^T + Z_s^{-T} G_s^T E_{s-1}^T\right\} \end{array}\right\} X_{s1}^T + O\left(\gamma^{-4}\right),
\end{aligned}$$

where

$$L_s = Z_s^{-T}\left(G_s^T E_{s-1} G_s + V_s\right) Z_s^{-1}$$

Thus

$$\begin{aligned}
S_{s|s} &= X_{s1}\left[\gamma^2\left(I_p - T_{s-1} - W_s W_s^T\right) + E_{s-1}\right. \\
&\quad \left. - E_{s-1}G_s Z_s^{-1} W_s^T - W_s Z_s^{-T} G_s E_{s-1} + W_s L_s W_s^T\right] X_{s1}^T + O\left(\frac{1}{\gamma^2}\right), \\
&= X_{s1}\left[\gamma^2\left(I_p - T_s\right) + F_{s-1}\right] X_{s1}^T,
\end{aligned}$$

where

$$F_{s-1} = \left(I - W_s Z_s^{-T} G_s^T\right) E_{s-1} \left(I - G_s Z_s^{-1} W_s^T\right) + W_s Z_s^{-T} V_s Z_s^{-1} W_s^T + O\left(\frac{1}{\gamma^2}\right). \tag{2.3.76}$$

The final step in the argument notes that

$$\begin{aligned} &E_{s-1} - E_{s-1} G_s Z_s^{-1} W_s^T - W_s Z_s^{-T} G_s^T E_{s-1} + W_s Z_s^{-T} G_s^T E_{s-1} G_s Z_s^{-1} W_s^T \\ &= \left(I - W_s Z_s^{-T} G_s^T\right) E_{s-1} \left(I - G_s Z_s^{-1} W_s^T\right). \end{aligned}$$

It now follows from (1.4.9) that $S_{s+1|s}$ has the required form with

$$E_s = F_{s-1} + X_{(s+1)1}^{-1} R_s X_{(s+1)1}^{-T}. \tag{2.3.77}$$

As $k \leq p$, and the boundedness of the inverse matrices is guaranteed by the assumptions made in the preamble of the Lemma, it follows that both E_s and F_s are $O(1) \succ 0$ as $\gamma^2 \to \infty$ for each s, $1 \leq s \leq k-1$. Also, $F_1 \succ 0$ because $V_1 \succ 0$ by (2.3.76), $F_s \succ 0$ implies $E_{s+1} \succ 0$ by (2.3.77), and $E_s \succ 0$ implies $F_s \succ 0$ by (2.3.76).

To develop the recurrence satisfied by the state variables note that the above manipulations give

$$\begin{aligned} S_{s|s-1} H_s^T \left\{H_s S_{s|s-1} H_s^T + V_s\right\}^{-1} = \\ X_{s1} \left\{ \begin{array}{c} \{\gamma^2 W_s + E_s G_s Z_s^{-1}\} \\ \left\{\gamma^2 I + Z_s^{-T} \left(G_s^T E_s G_s + V_s\right) Z_s^{-1}\right\}^{-1} \end{array} \right\} Z_s^{-T}. \end{aligned} \tag{2.3.78}$$

Thus, substituting in (1.4.6),

$$\mathbf{x}_{s|s} = X_{s-1} \mathbf{x}_{s-1|s-1} + X_{s1} \left\{ \begin{array}{c} \{\gamma^2 W_s + E_s G_s Z_s^{-1}\} \\ \left\{\gamma^2 I + Z_s^{-T} \left(G_s^T E_s G_s + V_s\right) Z_s^{-1}\right\}^{-1} \end{array} \right\} Z_s^{-T} \widetilde{\mathbf{y}}_s.$$

Separating out the leading order terms and making use of the representation (2.3.75) of $\mathbf{x}_{s|s}^0$ gives

$$\begin{aligned} \boldsymbol{\xi}_s &= \boldsymbol{\xi}_{s-1} + W_s Z_s^{-T} \left(\mathbf{y}_s - H_s X_{s-1} \mathbf{x}_{s-1|s-1}^0\right), \\ &= \boldsymbol{\xi}_{s-1} + W_s Z_s^{-T} \left(\mathbf{y}_s - G_s^T \boldsymbol{\xi}_{s-1}\right). \end{aligned}$$

Verification of (2.3.75) can be carried out neatly by using induction. Here the generalised inverse of G^{sT} uses the orthogonal factorization (2.3.71) of G^s.

$$\begin{aligned} \left[G^{sT}\right]^+ \mathbf{y}^s &= \left[\begin{array}{cc} Q_1^{s-1} & W_s \end{array} \right] \left[\begin{array}{cc} U_{s-1}^{-T} & \\ -Z_s^{-T} G_s^T Q^{(s-1)} U_{s-1}^{-T} & Z_s^{-T} \end{array} \right] \left[\begin{array}{c} \mathbf{y}^{s-1} \\ \mathbf{y}_s \end{array} \right], \\ &= \left[G^{(s-1)T}\right]^+ \mathbf{y}^{s-1} + W_s Z_s^{-T} \left(\mathbf{y}_s - G_s^T \left[G^{(s-1)T}\right]^+ \mathbf{y}^{s-1}\right), \\ &= \boldsymbol{\xi}_{s-1} + W_s Z_s^{-T} \left(\mathbf{y}_s - G_s^T \boldsymbol{\xi}_{s-1}\right) \\ &= \boldsymbol{\xi}_s. \end{aligned}$$

■ ■

The key result is that the computation settles down after k steps in the sense that the explicit dependence on γ^2 disappears when $\text{rank}\{G^s\} = p \Rightarrow s = k,\ T_s = I$. This means that suitable initial conditions on the filter for $s > k$ are provided at step k by the $\gamma^2 = \infty$ limiting values.

Theorem 2.4 *After exactly k steps*

$$S_{k|k} = X_{k1} F_{k-1}^0 X_{k1}^T = O(1),$$

where $F_{k-1}^0 = \lim_{\gamma^2 \to \infty} F_{k-1}$ is well defined and positive definite.

Proof. This follows from the preceding Lemma on noting that $\text{rank}\, G^k = k \Rightarrow T_k = I$ so that the explicit dependence on γ^2 in (2.3.73) disappears. ■

To compute the dependence of the state variables on all the data for $s < k$ requires a knowledge of the limiting form of the interpolation gain for large γ^2. The derivation makes use of the following result which can be verified by direct calculation.

Lemma 2.6 *Let the projection matrix T_i possess the orthogonal factorization*

$$T_i = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} I_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix},$$

and set

$$Q^T M_s Q = \begin{bmatrix} M_{11}^s & M_{12}^s \\ M_{21}^s & M_{22}^s \end{bmatrix},$$

where M_s is a generic (subscripted) matrix, then $\left(\gamma^2 Q_2 Q_2^T + E_1 + \frac{1}{\gamma^2} E_2 + \cdots\right)$ has the formal inverse $\left(B_1 + \frac{1}{\gamma^2} B_2 + \frac{1}{\gamma^4} B_3 + \cdots\right)$ with leading terms

$$\begin{aligned} B_1 &= Q_1 \left(E_{11}^1\right)^{-1} Q_1^T, \\ B_2 &= Q \begin{bmatrix} \left(E_{11}^1\right)^{-1} \left(E_{12}^1 E_{21}^1 - E_{11}^2\right) \left(E_{11}^1\right)^{-1} & -B_{11}^1 E_{12}^1 \\ -E_{21}^1 B_{11}^1 & I_2 \end{bmatrix} Q^T, \\ Q_2 Q_2^T B_3 &= -E_1 B_2 - E_2 B_1. \end{aligned}$$

To calculate the limiting form of the interpolation gain

$$A_i = S_{i|i} X_i^T S_{i+1|i}^{-1}$$

assume that it has the asymptotic expansion

$$A_i = X_{i1} \left(A_{i1} + \frac{1}{\gamma^2} A_{i2} + \cdots\right) X_{(i+1)1}^{-1}.$$

This expression has to be matched with

$$S_{i|i}X_i^T S_{i+1|i}^{-1} = X_{i1}\left(\gamma^2 Q_2 Q_2^T + F_{i1} + \frac{1}{\gamma^2}F_{i2} + \cdots\right)\left(B_{i1} + \frac{1}{\gamma^2}B_{i2} + \cdots\right)X_{(i+1)1}^{-1},$$

where the B_{ij} are the coefficients in the asymptotic expansion of $S_{i+1|i}^{-1}$ calculated using Lemma 2.6. This gives

$$\begin{aligned} A_{i1} &= (I_p - T_i)\,B_{i2} + F_{i1}B_{i1} = I + \left(F_{i1} - E_{(i+1)1}\right)B_{i1}, \\ A_{i2} &= (I_p - T_i)\,B_{i3} + F_{i1}B_{i2} + F_{i2}B_{i1}, \\ &= \left(F_{i1} - E_{(i+1)1}\right)B_{i2} + \left(F_{i2} - E_{(i+1)2}\right)B_{i1}. \end{aligned}$$

The corresponding recurrence for the variance is

$$S_{i|n} = A_i S_{(i+1)|n}A_i^T + S_{i|i} - A_i S_{(i+1)|i}A_i^T.$$

To show that this gives $O(1)$ terms it is necessary to show that the $O(\gamma^2)$ terms in $S_{i|i} - A_i S_{(i+1)|i}A_i^T$ cancel. This requires

$$(I_p - T_i) - A_{i1}\,(I_p - T_i)\,A_{i1}^T = 0.$$

This follows because

$$(I_p - T_i)\,B_{i2}\,(I_p - T_i) = (I_p - T_i)\,,\ (I_p - T_i)\,B_{i1} = B_{i1}\,(I_p - T_i) = 0.$$

The recurrence for the $O(1)$ term comes down to

$$S_{i|n} = A_i S_{(i+1)|n}A_i^T + X_{i1}C_i X_{i1}^T.$$

where

$$C_i = F_i - A_{i1}E_{i+1}A_{i1}^T - A_{i2}\,(I_p - T_i)\,A_{i1}^T - A_{i1}\,(I_p - T_i)\,A_{i2}^T.$$

Exercise 2.3.3 *If partial initial information is available then* $\mathbf{x}_{1|0} = \mathbf{x}_0$, $S_{1|0} = \gamma^2 T_0 + E_0$ *where* T_0 *is an orthogonal projection and* $E_0 \succeq 0$. *What changes must be made to the initialisation procedure to accommodate this more general situation?*

2.3.6 Methods for structured matrices

2.3.7 Applications of the conjugate gradient algorithm

The conjugate gradient algorithm addresses the problem of solving the system of equations

$$M\mathbf{x} = A^T\mathbf{b} = \mathbf{c},\ M \in R^p \to R^p, \tag{2.3.79}$$

where $M = A^TA$ is positive (semi)definite, by associating it with the optimization problem

$$\min_{\mathbf{x}} F(\mathbf{x});\ F(\mathbf{x}) = -\mathbf{c}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T M\mathbf{x}.$$

Directions which satisfy the conditions

$$\mathbf{d}_i^T M\mathbf{d}_j = \delta_{ij} \tag{2.3.80}$$

are said to be conjugate, and such directions provide a favourable set for descent calculations. Let

$$\mathbf{d}(\boldsymbol{\alpha}) = \sum_{i=1}^{k} \alpha_i \mathbf{d}_i,$$

where the $\mathbf{d}_i$ are conjugate and consider the problem of minimizing $F(\mathbf{x} + \mathbf{d}(\boldsymbol{\alpha}))$. The necessary conditions give

$$\begin{aligned} 0 &= \nabla_{\mathbf{x}} F \left[\begin{array}{cccc} \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_k \end{array} \right], \\ &= \left\{ -\mathbf{c}^T + \left(\mathbf{x} + \sum_{i=1}^{k} \alpha_i \mathbf{d}_i \right)^T M \right\} \left[\begin{array}{cccc} \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_k \end{array} \right], \\ &= \left\{ \nabla_{\mathbf{x}} F(\mathbf{x}) + \boldsymbol{\alpha}^T D_k^T M \right\} D_k, \end{aligned}$$

where

$$D_k = \left[\begin{array}{cccc} \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_k \end{array} \right].$$

Conjugacy (2.3.80) now gives

$$\boldsymbol{\alpha}^T = -\nabla_{\mathbf{x}} F(\mathbf{x})\, D_k.$$

This shows:

1. As $\nabla_{\mathbf{x}} F(\mathbf{x})$ is linear in $\mathbf{x}$, the computation of each α_i depends only on the conjugate direction $\mathbf{d}_i$,

   $$\alpha_i = -\nabla_{\mathbf{x}} F(\mathbf{x})\, \mathbf{d}_i$$

   and

2. The step to the minimum in the subspace is to a point

$$\mathbf{x}_k = \mathbf{x} - D_s D_s^T (M\mathbf{x} - \mathbf{c}) = \left(I - P^k\right)\mathbf{x} + D_k D_k^T \mathbf{c},$$

where $P^s = D_k D_k^T M$ is the projection onto the subspace spanned by the conjugate directions. If $k = p$ then $P^p = I$, and $\mathbf{x}_p = D_p D_p^T \mathbf{c}$ solves (2.3.79). This is readily verified by noting that MD_p is an M^{-1} conjugate set as

$$(MD_p)^T M^{-1} (MD_p) = I.$$

In this basis, as factors of the inverse commute,

$$\begin{aligned}\mathbf{c} &= (MD_p)(MD_p)^T M^{-1}\mathbf{c},\\ &= MD_p D_p^T \mathbf{c} = M\mathbf{x}_p.\end{aligned}$$

It follows that minimization along p mutually conjugate directions provides a basis for a finite algorithm (at least in exact arithmetic), and it is straightforward to generate such sets recursively by familiar orthogonalisation techniques. A suitable form is the following. Fix the initial point $\mathbf{x} = \mathbf{x}_0$, and set $\mathbf{d}_0 = -\mathbf{g}_0 = \mathbf{c} - M\mathbf{x}_0$. Then successive estimates are computed recursively using

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_k \mathbf{d}_k,\ \alpha_k = \frac{-\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{d}_k^T M \mathbf{d}_k},\\ \mathbf{d}_{k+1} &= -\mathbf{g}_{k+1} + \beta_k \mathbf{d}_k,\ \beta_k = \frac{\mathbf{g}_{k+1}^T \mathbf{g}_{k+1}}{\mathbf{g}_k^T \mathbf{g}_k}.\end{aligned}$$

The successive directions $\mathbf{d}_i$ are conjugate

$$\mathbf{d}_{k+1}^T M \mathbf{d}_i = 0,\ i \le k,$$

and the recurrence is verified inductively. The orthogonality properties $\mathbf{g}_{k+1}^T \mathbf{g}_k = 0$, $\mathbf{g}_{k+1}^T \mathbf{d}_k = 0$ are important in this verification.

Although the algorithm is finite in exact arithmetic as at most p descent steps are needed to reach the solution, this property is lost under perturbation in real computation. An important step in improving the algorithm is the use of a pre-conditioner [39]. Let $Z \succ 0$, and let $Z^{-1} M Z^{-1}$ be closer to a multiple of the unit matrix than is M in some sense (for example, better conditioned), then the above sketch for an algorithm can be reworked as an iteration for $\widetilde{\mathbf{x}} = Z\mathbf{x}$ satisfying the system

$$Z^{-1} M Z^{-1} \widetilde{\mathbf{x}} = Z^{-1}\mathbf{c}.$$

The resulting algorithm is most interesting in terms of the original variables. Let $W = Z^2$, $\mathbf{d}_0^W = -\mathbf{g}_0^W$. Then the basic step of the algorithm is

$$\begin{aligned}
\text{Solve } W\mathbf{g}_k^W &= M\mathbf{x}_k - \mathbf{c} \text{ for } \mathbf{g}_k^W, \\
\beta_k &= \frac{\mathbf{g}_k^{WT} W \mathbf{g}_k^W}{\mathbf{g}_{k-1}^{WT} W \mathbf{g}_{k-1}^W}, \\
\mathbf{d}_k^W &= -\mathbf{g}_k^W + \beta_k \mathbf{d}_{k-1}^W, \\
\alpha_k &= \frac{\mathbf{g}_k^{WT} W \mathbf{g}_k^W}{\mathbf{d}_k^{WT} M \mathbf{d}_k^W}, \\
\mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_k \mathbf{d}_k^W.
\end{aligned}$$

The key results associated with this form of the conjugate gradient algorithm are that the vectors $\mathbf{g}_k^W$ are W-conjugate (this generalises the orthogonality results noted above),

$$\mathbf{g}_k^{WT} W \mathbf{g}_j^W = 0, \; j < k,$$

and the vectors $\mathbf{d}_k^W$ are M-conjugate. The power of the method is shown to most effect in the result that the number of steps required in exact arithmetic is equal to the number of distinct eigenvalues of $Z^{-1}MZ^{-1}$. Thus if

$$M = W + N,$$

where N is a low rank correction, then the number of iterations required is $1 + \text{rank}(N)$. Of course it is important that W is readily invertible.

Example 2.3.6 *An example of a problem leading to a matrix which is the sum of an easily inverted matrix and one of low rank is given by the mean model (Section 1.5) which is balanced except for the entries in a few of the cells [36]. In this case, W can be taken as the matrix of the balanced problem, and $\mathbf{g}_k^W$ is known explicitly (1.5.4).*

The preconditioned algorithm permits of an interesting interpretation as a gradient method in a modified metric [50]. This identifies

$$\mathbf{g}^W = 2W^{-1}\{M\mathbf{x} - \mathbf{c}\} \tag{2.3.81}$$

as the gradient of $\|\mathbf{r}\|_2^2$, $\mathbf{r} = M\mathbf{x} - \mathbf{c}$, in the metric defined by W, $\|\mathbf{t}\|_W = \left\{\mathbf{t}^T W \mathbf{t}\right\}^{1/2}$. We have

$$\|\mathbf{r}\|_2^2 = \mathbf{r}^T W W^{-1} \mathbf{r} = \left\langle \mathbf{r}, W^{-1}\mathbf{r} \right\rangle.$$

An infinitesmal displacement $\delta\mathbf{r}$ causes a change

$$\begin{aligned} 2\mathbf{r}^T\delta\mathbf{r} &= 2\delta\mathbf{r}^T WW^{-1}\mathbf{r}, \\ &= 2\left\langle \delta\mathbf{r}, W^{-1}\mathbf{r}\right\rangle, \\ &= \left\langle \delta\mathbf{r}, \mathbf{g}^W\right\rangle. \end{aligned}$$

verifying (2.3.81). Also, in this metric,

$$\beta_k = \frac{\left\|\mathbf{g}_k^W\right\|_W^2}{\left\|\mathbf{g}_{k-1}^W\right\|_W^2},\ \alpha_k = \frac{\left\|\mathbf{g}_k^W\right\|_W^2}{\mathbf{d}_k^{WT} M\mathbf{d}_k^W}.$$

Here the geometric interpretation of α_k is as the step to the minimum of the objective function $F(\mathbf{x})$ in the direction determined by $\mathbf{d}_k^W$.

Chapter 3

The method of maximum likelihood

3.1 Introduction

The basic idea of maximum likelihood is relatively simple and leads to remarkably effective algorithms. It formalises the idea that the event outcomes making up a sequence of experimental observations are, at least in the long run, those more likely to be observed. While this does not mean that the probability of the occurrence of any particular event outcome is necessarily high, it does mean that the occurrence of a significant number of unlikely outcomes is very unlikely. To formalise this idea, let the probability density (probability mass for discrete distributions) associated with the outcome $\mathbf{y}_t \in R^q$ at configuration $\mathbf{t} \in R^l$ be $g\left(\mathbf{y}_t; \boldsymbol{\theta}_t, \mathbf{t}\right)$ where

$$\boldsymbol{\theta}_t = \boldsymbol{\eta}\left(\mathbf{t}, \mathbf{x}\right), \boldsymbol{\theta} \in R^s,\ \mathbf{x} \in R^p,$$

expresses a parametric model of a process in which the current realisation is determined by the values taken by the vector of parameters $\mathbf{x}$. The component values of the parameter vector $\mathbf{x}$ are not observed and it is assumed that these have to be estimated from the event outcomes $\mathbf{y}_t$. If the individual event outcomes are independent then the *likelihood*

$$\mathcal{G}\left(\mathbf{y}; \mathbf{x}, \mathbf{T}\right) = \prod_{\mathbf{t} \in \mathbf{T}} g\left(\mathbf{y}_t; \boldsymbol{\theta}_t, \mathbf{t}\right), \tag{3.1.1}$$

where $\mathbf{T}$ is the set of experimental configurations, is the joint density (probability mass) of the observed events, and the informal principle suggests this density considered as a function of $\mathbf{x}$ should be relatively large in the neighbourhood of the actual outcome. The computational principle is called the

method of maximum likelihood . It is expressed in the form

$$\widehat{\mathbf{x}} = \arg\max_{\mathbf{x}} \mathcal{G}\left(\mathbf{y}; \mathbf{x}, \mathbf{T}\right) \tag{3.1.2}$$

where $\widehat{\mathbf{x}}$ is the required estimate.

Example 3.1.1 *An important family of distributions frequently used in likelihood calculations is the exponential family . Here the density is usually represented in the form*

$$g\left(\mathbf{y}; \begin{bmatrix} \boldsymbol{\theta} \\ \boldsymbol{\phi} \end{bmatrix}\right) = c\left(\mathbf{y}, \boldsymbol{\phi}\right) \exp\left[\left\{\mathbf{y}^T\boldsymbol{\theta} - b\left(\boldsymbol{\theta}\right)\right\}/a\left(\boldsymbol{\phi}\right)\right]. \tag{3.1.3}$$

The corresponding likelihood in the case of independent events is

$$\mathcal{G} = \prod_{\mathbf{t}\in\mathbf{T}} c\left(\mathbf{y}_t, \boldsymbol{\phi}\right) \exp\left[\sum_{\mathbf{t}\in\mathbf{T}} \left\{\mathbf{y}_t^T\boldsymbol{\theta}_t - b\left(\boldsymbol{\theta}_t\right)\right\}/a\left(\boldsymbol{\phi}\right)\right]$$

Familiar cases include:

normal distribution *This is an example of a continuous density:*

$$\begin{aligned} g\left(y; \begin{bmatrix} \mu \\ \sigma \end{bmatrix}\right) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}, \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} e^{\frac{\left(2\mu y - \mu^2\right)}{2\sigma^2}}, \end{aligned} \tag{3.1.4}$$

so that

$$c\left(y, \phi\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}},\ \theta = \mu,\ b\left(\theta\right) = \mu^2/2,\ a\left(\phi\right) = \sigma^2.$$

In the multivariate case set

$$\begin{aligned} g\left(\mathbf{y}, \boldsymbol{\mu}, V\right) &= \frac{1}{\sqrt{\left(2\pi\right)^n \det V}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T V^{-1}(\mathbf{y}-\boldsymbol{\mu})}, \\ &= \frac{1}{\sqrt{\left(2\pi\right)^n \det V}} e^{-\frac{1}{2}\left(\mathbf{y}^T V^{-1}\mathbf{y}\right)} e^{(\boldsymbol{\mu}^T V^{-1}\mathbf{y} - \frac{\boldsymbol{\mu}^T V^{-1}\boldsymbol{\mu}}{2})}. \end{aligned} \tag{3.1.5}$$

Note that the scale has been absorbed into V. Then

$$\begin{aligned} c\left(\mathbf{y}, \phi\right) &= \frac{1}{\sqrt{\left(2\pi\right)^n \det V}} e^{-\frac{1}{2}\left(\mathbf{y}^T V^{-1}\mathbf{y}\right)},\ \boldsymbol{\theta} = V^{-1}\boldsymbol{\mu}, \\ b\left(\boldsymbol{\theta}\right) &= \frac{\boldsymbol{\mu}^T V^{-1}\boldsymbol{\mu}}{2} = \boldsymbol{\theta}^T V \boldsymbol{\theta},\ a\left(\phi\right) = 1. \end{aligned}$$

An alternative form is obtained by setting $\theta_1 = \sigma^{-2}\mu,\ \theta_2 = -\frac{1}{2\sigma^2}$. *Then*

$$g(y;\boldsymbol{\theta}) = e^{\left[\begin{array}{cc} y & y^2 \end{array}\right]\boldsymbol{\theta} - b(\boldsymbol{\theta})},$$

where

$$b(\boldsymbol{\theta}) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log\pi + \frac{1}{2}\log -\theta_2.$$

This form has some advantages. For example, the components of the vector multiplying $\boldsymbol{\theta}$ *contains sufficient statistics for estimating the parameters, and* $b(\boldsymbol{\theta})$ *is the moment generating function for these statistics. Thus:*

$$\begin{aligned}\frac{\partial b}{\partial \theta_1} &= -\frac{\theta_1}{2\theta_2} = \mu,\\ \frac{\partial b}{\partial \theta_2} &= \frac{\theta_1^2}{4\theta_2^2} - \frac{1}{2\theta_2} = \mu^2 + \sigma^2.\end{aligned}$$

The formalism extends to the multivariate case (3.1.5) . Let Set $\boldsymbol{\theta}_1 = V^{-1}\boldsymbol{\mu},\ \boldsymbol{\theta}_2 = -\frac{1}{2}vec\left(V^{-1}\right)$ *where vec is the operation mapping the columns of* V^{-1} *successively into a vector in* R^{n^2} *(the corresponding inverse operation is written invec). Then*

$$g(\mathbf{y},\boldsymbol{\theta}) = e^{\mathbf{y}^T\boldsymbol{\theta}_1 + vec\left(\mathbf{y}\mathbf{y}^T\right)\boldsymbol{\theta}_2 - b(\boldsymbol{\theta})},$$

where

$$b(\boldsymbol{\theta}) = -\frac{1}{4}\boldsymbol{\theta}_1^T invec\left(\boldsymbol{\theta}_2^{-1}\right)\boldsymbol{\theta}_1 + \frac{n}{2}\log\pi + \frac{1}{2}\log\det\left(-invec\left(\boldsymbol{\theta}_2\right)\right).$$

multinomial distribution *This is an example of a discrete distribution (probability mass function):*

$$\begin{aligned} g(\mathbf{n};\pi) &= \frac{n!}{\prod_{j=1}^p n_j!}\prod_{j=1}^p \pi_j^{n_j},\\ &= \frac{n!}{\prod_{j=1}^p n_j!}e^{\sum_{j=1}^p n_j\log\pi_j}, \end{aligned} \tag{3.1.6}$$

where $\sum_{j=1}^p n_j = n$, *and the frequencies* $\pi(j), j = 1, 2, \cdots, p$ *must satisfy the condition*

$$\sum_{j=1}^p \pi_j = 1.$$

Eliminating π_p gives

$$\begin{aligned}\sum_{j=1}^{p} n_j \log \pi_j &= \sum_{j=1}^{p-1} n_j \log \pi_j + \left(n - \sum_{j=1}^{p-1} n_j \right) \log \left(1 - \sum_{j=1}^{p-1} \pi_j \right) \\ &= \sum_{j=1}^{p-1} n_j \log \frac{\pi_j}{1 - \sum_{i=1}^{p-1} \pi_i} + n \log \left(1 - \sum_{j=1}^{p-1} \pi_j \right)\end{aligned}$$

It follows that

$$c(\mathbf{n}) = \frac{n!}{\prod_{j=1}^{p} n_j!},\ a(\phi) = 1, \tag{3.1.7}$$

$$\theta_j = \log \frac{\pi_j}{1 - \sum_{i=1}^{p-1} \pi_i}, \tag{3.1.8}$$

$$\pi_p = \frac{1}{1 + \sum_{i=1}^{p-1} e^{\theta_i}},$$

$$b(\boldsymbol{\theta}) = n \log \left(1 + \sum_{j=1}^{p-1} e^{\theta_j} \right). \tag{3.1.9}$$

Example 3.1.2 *One important model for the event data $\mathbf{y}_t$ is that of a signal observed in the presence of noise . In this case typically there would be a parametric model for the signal given by*

$$\mathcal{E}_{g(\mathbf{y};\boldsymbol{\theta}^*,\mathbf{t})} \{\mathbf{y}\} = \boldsymbol{\mu}(\mathbf{x}^*, \mathbf{t}). \tag{3.1.10}$$

The normal distribution provides one example of a distribution typically used for describing the noise. Here $\boldsymbol{\mu}$ specifies location, while the second parameter σ determines the scale of the noise. In exponential family form for the normal distribution $\frac{db}{d\theta} = \mu$. This result is an instance of the general property of the exponential family that the term $\frac{b(\theta)}{a(\phi)}$, which serves to normalise the density integral, is directly related to a moment generating function . From

$$\int_{R(\mathbf{y})} g\left(\mathbf{y}; \begin{bmatrix} \boldsymbol{\theta} \\ \boldsymbol{\phi} \end{bmatrix} \right) d\mathbf{y} = 1$$

it follows that

$$\int_{R(\mathbf{y})} c(\mathbf{y}, \boldsymbol{\phi})\, e^{\left[\{\mathbf{y}^T \boldsymbol{\theta}\}/a(\boldsymbol{\phi})\right]} d\mathbf{y} = e^{b(\boldsymbol{\theta})/a(\boldsymbol{\phi})}. \tag{3.1.11}$$

Differentiating both sides of (3.1.11) and identifying terms gives

$$\mathcal{E}\{\mathbf{y}\} = \nabla b(\boldsymbol{\theta})^T. \tag{3.1.12}$$

Differentiating (3.1.11) again gives

$$\mathcal{E}\left\{\mathbf{y}\mathbf{y}^T\right\} = \nabla b\left(\boldsymbol{\theta}\right)^T \nabla b\left(\boldsymbol{\theta}\right) + a\left(\boldsymbol{\phi}\right)\nabla^2 b \tag{3.1.13}$$

from which it follows that

$$\mathcal{V}\left\{\mathbf{y}\right\} = a\left(\boldsymbol{\phi}\right)\nabla^2 b\left(\boldsymbol{\theta}\right).$$

If an invertible transformation $\mathbf{h}\left(\boldsymbol{\mu}\right)$ *can be found linking the mean* $\boldsymbol{\mu} = \nabla b^T$ *with a linear model* $\mathbf{h}\left(\boldsymbol{\mu}\right) = A\mathbf{x}$ *then* $\mathbf{h}\left(\boldsymbol{\mu}\right)$ *is called a link function . It is an important component in the analysis of generalised linear models presented in [69]. One reason for this is that special cases can simplify computation. For example, if the distribution is normal, then the choice* $h\left(\mu\right) = \mu$ *leads to a linear least squares problem for the maximum likelihood estimate. More generally, if* $\mathbf{h}\left(\boldsymbol{\mu}\right) = \boldsymbol{\theta}$ *then* $\mathbf{h}$ *is called a canonical link .*

Equation (3.1.2) provides an important example of the application of the general method of maximum likelihood in which unobserved parametric information is estimated by maximizing the joint density (probability mass for discrete distributions) with respect to these parameters. It proves remarkably successful in a wide variety of contexts but needs to be hedged with appropriate reservations. Associated with the likelihood is its logarithm

$$\mathcal{L}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right) = \sum_{\mathbf{t}\in\mathbf{T}} \log g\left(\mathbf{y}_t;\boldsymbol{\theta}_t,\mathbf{t}\right). \tag{3.1.14}$$

Maximizing the log likelihood is equivalent to maximizing the likelihood and this is the strategy usually adopted. Typically the assumptions made are

1. that there exists a true model with parameter vector $\mathbf{x}^*$,
2. the log likelihood is at least two times differentiable with Lipschitz continuous second derivatives as a function of $\mathbf{x}$ in an open region that contains $\mathbf{x}^*$ properly in its interior, and
3. integrals with respect to $\mathbf{y}$, in particular the boundedness of expectations of (products of) the likelihood, log likelihood, derivatives with respect to $\mathbf{x}$ and polynomial terms in the components of $\mathbf{y}$, taken over $S\left(\mathbf{y}\right) = \text{range}\left(\mathbf{y}\right)$, is assumed.

The above examples of distributions satisfy these requirements. Examples which do not include:

1. Uniform distribution on $[0,\theta]$. Here the parameter is θ so convergent estimates must approach the boundary of the allowable range, and

2. Negative exponential distribution

$$g\left(y;\left[\begin{array}{c}\theta\\ \phi\end{array}\right]\right)=\frac{\phi}{2}e^{-\phi|y-\theta|}.$$

Here the density is not smooth when $y=\theta$.

It does not follow that maximum likelihood has no value for distributions of this kind. Just that properties derived under an assumption of smoothness may not fit these cases.

Consequences of the above assumptions include the following results. It is assumed that range $(\mathbf{y})=S(Y)$ is independent of $\mathbf{x}$

Lemma 3.1

$$0=\mathcal{E}\left\{\nabla_{\mathbf{x}}\mathcal{L}\left(\left(\mathbf{y};\mathbf{x},\mathbf{T}\right)\right)\right\}. \tag{3.1.15}$$

Proof. It follows from

$$1=\int_{S(Y)}\mathcal{G}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right)d\mathbf{y},$$

by differentiating under the integral sign, that

$$\begin{aligned}0&=\int_{S(Y)}\nabla_{\mathbf{x}}\mathcal{G}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right)d\mathbf{y},\\&=\int_{S(Y)}\sum_{t\in\mathbf{T}}\left\{\frac{\nabla_{\boldsymbol{\theta}_t}g\left(\mathbf{y}_t;\boldsymbol{\theta}_t,\mathbf{t}\right)}{g\left(\mathbf{y}_t;\boldsymbol{\theta}_t,\mathbf{t}\right)}\nabla_{\mathbf{x}}\boldsymbol{\theta}_t\right\}\mathcal{G}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right)d\mathbf{y},\\&=\mathcal{E}\left\{\nabla_{\mathbf{x}}\mathcal{L}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right)\right\}.\end{aligned}$$

■ ■

This result applies to each term in (3.1.15) separately. Let

$$L_t\left(\mathbf{y}_t;\boldsymbol{\theta}_t,\mathbf{t}\right)=\log g\left(\mathbf{y}_t;\boldsymbol{\theta}_t,\mathbf{t}\right)$$

then

$$0=\mathcal{E}\left\{\nabla_{\mathbf{x}}L_t\left(\mathbf{y}_t;\boldsymbol{\theta}_t,\mathbf{t}\right)\right\}. \tag{3.1.16}$$

Lemma 3.2

$$\mathcal{E}\left\{\nabla_{\mathbf{x}}^2\mathcal{L}\left(\left(\mathbf{y};\mathbf{x},\mathbf{T}\right)\right)\right\}=-\mathcal{E}\left\{\nabla_{\mathbf{x}}\mathcal{L}\left(\left(\mathbf{y};\mathbf{x},\mathbf{T}\right)\right)^T\nabla_{\mathbf{x}}\mathcal{L}\left(\left(\mathbf{y};\mathbf{x},\mathbf{T}\right)\right)\right\} \tag{3.1.17}$$

Proof. This follows by differentiating under the integral sign a second time. This gives

$$\begin{aligned} 0 &= \nabla_{\mathbf{x}} \int_{S(Y)} \nabla_{\mathbf{x}} \mathcal{L}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right)^T \mathcal{G}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right) d\mathbf{y}, \\ &= \int_{S(Y)} \left\{ \nabla_{\mathbf{x}}^2 \mathcal{L}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right) + \nabla_{\mathbf{x}} \mathcal{L}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right)^T \nabla_{\mathbf{x}} \mathcal{L}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right) \right\} \mathcal{G}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right) d\mathbf{y}, \end{aligned}$$

and the result follows directly . ■■

As in the previous lemma the result applies term by term. Here independence is important as the vector result uses

$$\mathcal{E}\left\{ \nabla_{\mathbf{x}} L_i\left(\mathbf{y}_i;\boldsymbol{\theta}_i,\mathbf{t}_i\right)^T \nabla_{\mathbf{x}} L_j\left(\mathbf{y}_j;\boldsymbol{\theta}_j,\mathbf{t}_j\right) \right\} = 0, \quad i \neq j.$$

Example 3.1.3 *These two fundamental identities are readily verified in the case of distributions from the exponential family (3.1.3). Here*

$$L = \log c\left(\mathbf{y},\phi\right) + \left(\mathbf{y}^T\boldsymbol{\theta} - b\left(\boldsymbol{\theta}\right)\right)/a\left(\phi\right).$$

Differenciating and using (3.1.12) gives

$$\nabla_{\boldsymbol{\theta}} L = \left(\mathbf{y}^T - \mathcal{E}\left\{\mathbf{y}\right\}^T\right)/a\left(\phi\right) \Rightarrow \mathcal{E}\left\{\nabla_{\boldsymbol{\theta}} L\right\} = 0$$

which is (3.1.15). Also, it follows that

$$\mathcal{V}\left\{\mathbf{y}\right\} = a^2 \mathcal{E}\left\{\nabla_{\boldsymbol{\theta}} L^T \nabla_{\boldsymbol{\theta}} L\right\}.$$

The second identity (3.1.17) is a consequence of (3.1.13).

$$\nabla_{\boldsymbol{\theta}}^2 L = -\nabla_{\boldsymbol{\theta}}^2 b/a \Rightarrow a^2 \mathcal{E}\left\{\nabla_{\boldsymbol{\theta}}^2 L\right\} = -\mathcal{V}\left\{\mathbf{y}\right\} = -a^2 \mathcal{E}\left\{\nabla_{\boldsymbol{\theta}} L^T \nabla_{\boldsymbol{\theta}} L\right\}.$$

Definition 3.1 *The matrix*

$$\mathcal{I}_n = \frac{1}{n}\mathcal{E}\left\{ \nabla_{\mathbf{x}} \mathcal{L}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right)^T \nabla_{\mathbf{x}} \mathcal{L}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right) \right\} = \mathcal{V}\left\{ \frac{1}{\sqrt{n}} \nabla_{\mathbf{x}} \mathcal{L}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right) \right\}, \tag{3.1.18}$$

where $n = |\mathbf{T}|$, is called the (Fisher) information matrix associated with the parameter vector $\mathbf{x}$.

$\mathcal{I}_n$ is generically positive definite, and the scaling is chosen so the limit as $n \to \infty$ is reasonable under appropriate sampling regimes because then the strong (weak) law of large numbers ensures an almost sure (in probability) limit. The information matrix allows us to get a hold on the variance of unbiassed estimators of $\mathbf{x}$ by providing a minimum variance lower bound . This is the substance of a famous result associated with the names Cramer and Rao , but earlier atributions exist [57].

Theorem 3.1 *Let* $\widehat{\mathbf{x}}\{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n\} = \widehat{\mathbf{x}}(\mathbf{y})$ *be an unbiassed estimator of* $\mathbf{x}^*$ *based on the data* $\{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n\}$ *so that*

$$\mathcal{E}\{\widehat{\mathbf{x}}\} = \int_{S(Y)} \widehat{\mathbf{x}}(\mathbf{y})\, \mathcal{G}(\mathbf{y}; \mathbf{x}^*, \mathbf{T})\, d\mathbf{y} = \mathbf{x}^*.$$

Then

$$\mathcal{V}\left\{\sqrt{n}\,(\widehat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}^*)\right\} \geq \mathcal{I}_n^{-1} \tag{3.1.19}$$

where the inequality (3.1.19) is to be interpreted as indicating that the matrix difference $\mathcal{V}\{\sqrt{n}\,(\widehat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}^*)\} - \mathcal{I}_n^{-1}$ *is positive definite.*

Proof. Differentiating the condition for $\widehat{\mathbf{x}}(\mathbf{y})$ to be an unbiassed estimator gives

$$\int_{S(Y)} \widehat{\mathbf{x}}(\mathbf{y})\, \nabla_{\mathbf{x}^*} \mathcal{G}(\mathbf{y}; \mathbf{x}^*, \mathbf{T})\, d\mathbf{y} = \int_{S(Y)} \widehat{\mathbf{x}}(\mathbf{y})\, \nabla_{\mathbf{x}^*} \mathcal{L}(\mathbf{y}; \mathbf{x}^*, \mathbf{T})\, \mathcal{G}(\mathbf{y}; \mathbf{x}^*, \mathbf{T})\, d\mathbf{y} = I.$$

This just says that, using (3.1.15),

$$\mathcal{C}\left\{\sqrt{n}\,(\widehat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}^*), \frac{1}{\sqrt{n}} \nabla_{\mathbf{x}^*} \mathcal{L}(\mathbf{y}; \mathbf{x}^*, \mathbf{T})\right\} = I.$$

Thus

$$\mathcal{V}\left\{\begin{bmatrix} \sqrt{n}\,(\widehat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}^*) \\ \frac{1}{\sqrt{n}} \nabla_{\mathbf{x}^*} \mathcal{L}(\mathbf{y}; \mathbf{x}^*, \mathbf{T})^T \end{bmatrix}\right\} = \begin{bmatrix} \mathcal{V}\{\sqrt{n}\,(\widehat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}^*)\} & I \\ I & \mathcal{I}_n \end{bmatrix},$$

where the matrix on the right hand side is generically positive definite. It follows that the matrix

$$\begin{bmatrix} \mathcal{V}\{\sqrt{n}\,(\widehat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}^*)\} - \mathcal{I}_n^{-1} & 0 \\ 0 & \mathcal{I}_n \end{bmatrix} = \begin{bmatrix} I & -\mathcal{I}_n^{-1} \\ & I \end{bmatrix} \begin{bmatrix} \mathcal{V}\{\sqrt{n}\,(\widehat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}^*)\} & I \\ I & \mathcal{I}_n \end{bmatrix} \begin{bmatrix} I & \\ -\mathcal{I}_n^{-1} & I \end{bmatrix}$$

is also positive definite. The result follows as its leading principal minors must also be positive definite. ■■

Maximum likelihood estimates are not necessarily unbiassed. However, a consistency result is established in the next section, and this is something like an asymptotic form of the unbiassedness condition. It goes together with a result that shows that asymptotically the estimate satisfies the minimum variance bound. This is generally taken as a strong argument in favour of maximum likelihood estimation.

3.2 Asymptotic properties

3.2.1 Setting the scene

In this section the convergence behaviour of the maximum likelihood estimate $\widehat{\mathbf{x}}_n$ determined by the particular set of observed events corresponding to $|\mathbf{T}| = n$ is studied as n increases without bound. To develop the asymptotic results which provide a basis for discussing such questions as the estimation of the rate of convergence of numerical algorithms it is necessary to provide a mechanism (systematic or random) which allows the specification of a sequence of sets of designed experiments with $n = |\mathbf{T}_n|$, the number of observations per set, tending to ∞. The idea has been introduced in Chapter 1, and a slight generalisation of (1.1.6) to $\mathbf{t} \in R^s$ allows for data collected on spatial grids, for example. It is stressed that the key idea is that designed experiments allow sets of regularly sampled points $\mathbf{t}_i \in \mathbf{T}_n$ to eventually fill out a region $S(\mathbf{T}) \subseteq R^s$ so that

$$\frac{1}{n} \sum_{\mathbf{t} \in \mathbf{T}_n} f(\mathbf{t}) \to \int_{S(\mathbf{T})} f(\mathbf{t}) \rho(\mathbf{t}) \, d\mathbf{t} \tag{3.2.1}$$

for all sufficiently smooth $f(\mathbf{t})$ where $\rho(\mathbf{t})$ is a density function describing the limiting form of the measuring process as the number of observations grows without bound. It was noted in Chapter 1 that $\rho(\mathbf{t}) = 1$ if configuration values are allocated in $S(\mathbf{T})$ either by successive subdivision of a uniform grid or randomly under a uniform distribution, but that the interpretation as a quadrature formula differs in the two cases.

Definition 3.2 *$\widehat{\mathbf{x}}_n$ is a (strong/weak) consistent estimate of $\mathbf{x}^*$ if*

$$\widehat{\mathbf{x}}_n \to \mathbf{x}^*, \; n \to \infty \text{ almost surely (a.s.)/in probability.}$$

The different modes of stochastic convergence prove to be appropriate for different forms (strong/weak) of the law of large numbers . Our considerations will generally have to do with the strong law, and the choice of an appropriate form is discussed in the appendix to this chapter.

The proof of the consistency of estimates given here is by no means weakest possible. For a classical treatment see [110]. However, the method presented here has several advantages:

1. it links the concept of consistency with an algorithmic approach based on Newton's method which will be a paradigm for our computational methods;

2. it makes no assumption about the knowledge of a global maximum of the likelihood function; and

3. it lends itself to discussion of the case when model values can only be estimated approximately.

The connection requires the characterization of $\widehat{\mathbf{x}}_n$ as the solution of a system of nonlinear equations and these are obtained as the necessary conditions for the maximization of (3.1.14). This gives the *estimating equation*

$$\nabla_{\mathbf{x}}\mathcal{L}\left(\mathbf{y};\mathbf{x},\mathbf{T}\right)=0. \tag{3.2.2}$$

It is convenient to work with the log likelihood which leads to equations involving sums of random variables to which the strong law of large numbers can be applied as a consequence of our regularity assumptions. This point is illustrated first by deriving a system of equations satisfied by the true vector of parameters $\mathbf{x}^*$.

Theorem 3.2 *If $\{\mathbf{T}_n\}$ is defined as a sequence of designed experiments for each n as $n\to\infty$ then, under the assumptions governing the regularity of the likelihood, the true vector of parameters $\mathbf{x}^*$ satisfies the system of equations*

$$\int_{S(\mathbf{T})}\mathcal{E}^*\left\{\nabla_{\mathbf{x}}L_t\left(\mathbf{y};\boldsymbol{\theta}_t,\mathbf{t}\right)\right\}\rho\left(\mathbf{t}\right)d\mathbf{t}=0, \tag{3.2.3}$$

where $\mathcal{E}^*$ indicates that the expectation is evaluated using the true density $g\left(\mathbf{y};\boldsymbol{\theta}_t^*,\mathbf{t}\right)$. The corresponding limiting form of the information matrix is

$$\mathcal{I}=-\int_{S(\mathbf{T})}\mathcal{E}^*\left\{\nabla_{\mathbf{x}}^2L_t\left(\mathbf{y};\boldsymbol{\theta}_t,\mathbf{t}\right)\right\}\rho\left(\mathbf{t}\right)d\mathbf{t}. \tag{3.2.4}$$

Proof. Writing out (3.2.2) in terms of its components gives

$$\begin{aligned}\nabla_{\mathbf{x}}\mathcal{L}\left(\mathbf{y};\mathbf{x},\mathbf{T}_n\right)&=\sum_{\mathbf{t}\in\mathbf{T}_n}\nabla_{\mathbf{x}}L_t(\mathbf{y}_t;\boldsymbol{\theta}_t,\mathbf{t}),\\&=\sum_{\mathbf{t}\in\mathbf{T}_n}\left\{\nabla_{\mathbf{x}}L_t(\mathbf{y}_t;\boldsymbol{\theta}_t,\mathbf{t})-\mathcal{E}^*\left\{\nabla_{\mathbf{x}}L_t\left(\mathbf{y}_t;\boldsymbol{\theta}_t,\mathbf{t}\right)\right\}\right\}+\\&\sum_{\mathbf{t}\in\mathbf{T}_n}\mathcal{E}^*\left\{\nabla_{\mathbf{x}}L_t\left(\mathbf{y}_t;\boldsymbol{\theta}_t,\mathbf{t}\right)\right\}.\end{aligned}$$

Applying the strong law of large numbers (appendix to this chapter) to the first term on the right hand side, and the condition (3.2.1) for designed experiments to the second, gives almost surely

$$\frac{1}{n}\nabla_{\mathbf{x}}\mathcal{L}\left(\mathbf{y};\mathbf{x},\mathbf{T}_n\right)\to\int_{S(\mathbf{T})}\mathcal{E}^*\left\{\nabla_{\mathbf{x}}L_t\left(\mathbf{y};\boldsymbol{\theta}_t,\mathbf{t}\right)\right\}\rho\left(\mathbf{t}\right)d\mathbf{t}.$$

It now follows from Lemma 3.1 that (3.2.3) is satisfied when $\mathbf{x} = \mathbf{x}^*$ corresponding to $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^*$. The limiting form of the information matrix follows from (3.1.18) using the condition for designed experiments. ■■
Equation (3.2.3) expresses the condition that $\mathbf{x}^*$ is a stationary point of

$$L^*(\mathbf{x}) = \int_{S(\mathbf{T})} \mathcal{E}^* \left\{ L_t(\mathbf{y}; \boldsymbol{\theta}_t, \mathbf{t}) \right\} \rho(\mathbf{t})\, d\mathbf{t}$$

A sufficient condition for $\mathbf{x}^*$ to be an isolated solution of (3.2.3) is that $L^*(\mathbf{x})$ has an isolated maximum at $\mathbf{x} = \mathbf{x}^*$. This will be the case if $\nabla_{\mathbf{x}}^2 L^*(\mathbf{x}^*)$ is negative definite. From Lemma 3.2 we have

$$\begin{aligned}\nabla_{\mathbf{x}}^2 L^*(\mathbf{x}^*) &= \int_{S(\mathbf{T})} \mathcal{E}^* \left\{ \nabla_{\mathbf{x}}^2 L_t(\mathbf{y}; \boldsymbol{\theta}_t^*, \mathbf{t}) \right\} \rho(\mathbf{t})\, d\mathbf{t} \\ &= -\int_{S(\mathbf{T})} \mathcal{E}^* \left\{ \nabla_{\mathbf{x}} L_t(\mathbf{y}; \boldsymbol{\theta}_t^*, \mathbf{t})^T \nabla_{\mathbf{x}} L_t(\mathbf{y}; \boldsymbol{\theta}_t^*, \mathbf{t}) \right\} \rho(\mathbf{t})\, d\mathbf{t}.\end{aligned}$$

Thus this requirement does not amount to a severe assumption.

3.2.2 Consistency of estimates

The method used to demonstrate consistency makes use of Newton's method to solve the estimating equation (3.2.3). Let

$$\mathcal{J}_n(\mathbf{x}) = \frac{1}{n} \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{y}; \mathbf{x}, \mathbf{T}_n).$$

Then the Newton iteration is defined by

$$\begin{aligned}&\mathbf{h} = -\mathcal{J}_n(\mathbf{x})^{-1} \frac{1}{n} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{y}; \mathbf{x}, \mathbf{T}_n)^T, \\ &\mathbf{x} \leftarrow \mathbf{x} + \mathbf{h}.\end{aligned}$$

Associated with this iteration is the famous Kantorovich theorem [66] , p.277, which not only provides conditions under which the iteration converges but goes further even permitting the existence of a solution to be deduced from calculations made at the initial point of the iteration $\mathbf{x}_0$.

Theorem 3.3 (Kantorovich) *If the following conditions are satisfied in a ball $S_\varrho = \{\mathbf{x}; \|\mathbf{x} - \mathbf{x}_0\| < \varrho\}$:*

(i) $\|\mathcal{J}_n(\mathbf{u}) - \mathcal{J}_n(\mathbf{v})\| \leq K_1 \|\mathbf{u} - \mathbf{v}\|, \ \forall \mathbf{u}, \mathbf{v} \in S_\varrho,$

(ii) $\left\|\mathcal{J}_n(\mathbf{x}_0)^{-1}\right\| = K_2,$

(iii) $\left\| \mathcal{J}_n(\mathbf{x}_0)^{-1} \frac{1}{n} \nabla_x \mathcal{L}(\mathbf{y}; \mathbf{x}_0, \mathbf{T}_n)^T \right\| = K_3$, *and*

(iv) $\xi = K_1 K_2 K_3 < \frac{1}{2}$,

then Newton's method converges to a point $\widehat{\mathbf{x}} \in S_\varrho$ *satisfying (3.2.3), and* $\widehat{\mathbf{x}}$ *is the only root of (3.2.3) in* S_ϱ.

Remark 3.2.1 *A lower bounds on* ϱ *can be obtained by summing estimates of successive Newton corrections. It turns out that*

$$\frac{1}{\xi}\left(1 - \sqrt{1 - 2\xi}\right) K_3 < \varrho \tag{3.2.5}$$

is a consequence of the conditions imposed in the theorem. Also the step to the solution $\widehat{\mathbf{x}}$ *is bounded by*

$$\|\widehat{\mathbf{x}} - \mathbf{x}_0\|_2 < 2K_3. \tag{3.2.6}$$

Theorem 3.4 (Consistency of likelihood) *Let the estimation problem associated with a sequence of designed experiments have a well determined solution in the sense that* $\nabla^2_{\mathbf{x}} L^*(\mathbf{x}^*)$ *is bounded, negative definite. Then for each sequence of experiments* $\{\mathbf{T}_n\}$ *there exists an* n_0 *such that the Newton iteration started at* $\mathbf{x}^*$ *converges to* $\widehat{\mathbf{x}}_n$ *maximizing (possibly just locally)* $\mathcal{L}(\mathbf{y}; \mathbf{x}, \mathbf{T}_n)$ *for almost all* $n > n_0$, *and*

$$\widehat{\mathbf{x}}_n \underset{n\to\infty}{\overset{a.s.}{\to}} \mathbf{x}^*. \tag{3.2.7}$$

Proof. This involves verifying the conditions of the Kantorovich theorem at $\mathbf{x}^*$. The product $K_1 K_2$ is like $\operatorname{cond} \mathcal{J}_n(\mathbf{x}^*)$ and so is scale independent. The trick of adding and subtracting expectations can be used to show that, by the strong law of large numbers,

$$\mathcal{J}_n(\mathbf{x}^*) \overset{\text{a.s.}}{\to} \int_{S(\mathbf{T})} \mathcal{E}\left\{\nabla^2_{\mathbf{x}} L_t(\mathbf{y}; \mathbf{x}^*, \mathbf{T})\right\} \rho(\mathbf{t})\, d\mathbf{t} = -\mathcal{I},\ n \to \infty. \tag{3.2.8}$$

Thus the assumption that $K_1 K_2$ is bounded is equivalent to the assumption that the estimation problem has a well determined solution. It remains to show that

$$K_3 = \left\| \mathcal{J}_n(\mathbf{x}^*)^{-1} \frac{1}{n} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{y}; \mathbf{x}^*, \mathbf{T}_n)^T \right\|$$

gets small almost surely as $n \to \infty$. K_3 is just the magnitude of the Newton correction, and the interesting part is

$$\left\| \frac{1}{n} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{y}; \mathbf{x}^*, \mathbf{T}_n) \right\| \le \left\| \frac{1}{n} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{y}; \mathbf{x}^*, \mathbf{T}_n) - \frac{1}{n} \sum_{\mathbf{t} \in \mathbf{T}_n} \mathcal{E}^*\left\{\nabla_{\mathbf{x}} L_t(\mathbf{y}_t; \boldsymbol{\theta}^*_t, \mathbf{t})\right\} \right\| +$$
$$\left\| \begin{array}{l} \frac{1}{n} \sum_{\mathbf{t} \in \mathbf{T}_n} \mathcal{E}^*\left\{\nabla_{\mathbf{x}} L_t(\mathbf{y}_t; \boldsymbol{\theta}^*_t, \mathbf{t})\right\} - \\ \int_{S(\mathbf{T})} \mathcal{E}^*\left\{\nabla_{\mathbf{x}} L_t(\mathbf{y}; \boldsymbol{\theta}^*_t, \mathbf{t})\right\} \rho(\mathbf{t})\, d\mathbf{t}. \end{array} \right\|,$$

where Lemma 3.1 has been used. Here the first term on the right hand side get small almost surely as $n \to \infty$, while the second gets small as a consequence of the designed experiment assumption. It follows that $\xi \overset{a.s.}{\underset{n\to\infty}{\to}} 0$. Thus $\widehat{\mathbf{x}}_n$ is well defined and lies in S_{ϱ_n} for almost all n large enough. As $\|\widehat{\mathbf{x}}^* - \widehat{\mathbf{x}}_n\|$ shrinks to zero with K_3 by (3.2.6), (3.2.7) is an immediate consequence. ■

The property of consistency permits the limiting form of the distribution of the maximum likelihood estimates to be deduced. Expanding the estimating equation (3.2.2) about the true parameter value $\mathbf{x}^*$ gives

$$0 = \nabla_{\mathbf{x}}\mathcal{L}\left(\mathbf{y};\mathbf{x}^*,\mathbf{T}_n\right)+\left(\widehat{\mathbf{x}}-\mathbf{x}^*\right)^T\nabla^2_{\mathbf{x}}\mathcal{L}\left(\mathbf{y};\mathbf{x}^*,\mathbf{T}_n\right)+\frac{1}{2}\nabla^3_{\mathbf{x}}\mathcal{L}\left(\mathbf{y};\widetilde{\mathbf{x}},\mathbf{T}_n\right)\left(\widehat{\mathbf{x}}-\mathbf{x}^*,\widehat{\mathbf{x}}-\mathbf{x}^*\right),$$

where $\widetilde{\mathbf{x}}$ is a vector of mean values and hence a consistent estimator of $\mathbf{x}^*$. Solving for $\widehat{\mathbf{x}}-\mathbf{x}^*$ gives

$$\widehat{\mathbf{x}}-\mathbf{x}^* = \left[\mathcal{J}_n\left(\mathbf{x}^*\right)+\frac{1}{2n}\nabla^3_{\mathbf{x}}\mathcal{L}\left(\mathbf{y};\widetilde{\mathbf{x}},\mathbf{T}_n\right)\left(\widehat{\mathbf{x}}-\mathbf{x}^*,\cdot\right)\right]^{-1}\frac{1}{n}\nabla_{\mathbf{x}}\mathcal{L}\left(\mathbf{y};\mathbf{x}^*,\mathbf{T}_n\right)^T$$

Under reasonable conditions the mean value term is small compared with the second derivative term in the square brackets. One such condition is the uniform bound assumption [68] .

$$\left\|\frac{1}{n}\nabla^3_{\mathbf{x}}\mathcal{L}\left(\mathbf{y};\widetilde{\mathbf{x}},\mathbf{T}_n\right)\left(\widehat{\mathbf{x}}-\mathbf{x}^*,\cdot\right)\right\| \le \left\{\frac{1}{n}\sum_{\mathbf{t}\in\mathbf{T}}B\left(\mathbf{y}_t,\mathbf{t}\right)\right\}\left\|\widehat{\mathbf{x}}-\mathbf{x}^*\right\|,$$

where the $B\left(\mathbf{y}_t,\mathbf{t}\right)$ are positive random variables with bounded variance. In this case the sum in brackets tends to a limit in the appropriate stochastic convergence mode while the norm term is small as a consequence of consistency. The limiting distribution of $\frac{1}{\sqrt{n}}\nabla_{\mathbf{x}}\mathcal{L}\left(\mathbf{y};\mathbf{x}^*,\mathbf{T}_n\right)^T$ follows from the central limit theorem and is given by

$$\frac{1}{\sqrt{n}}\nabla_{\mathbf{x}}\mathcal{L}\left(\mathbf{y};\mathbf{x}^*,\mathbf{T}_n\right)^T \backsim N\left(0,\mathcal{I}\right),\ n\to\infty,$$

while $\mathcal{J}_n \to \mathcal{I}$ given by (3.2.4). Thus the asymptotic distribution of $\sqrt{n}\left(\widehat{\mathbf{x}}-\mathbf{x}^*\right)$ is ([98], p.209)

$$\sqrt{n}\left(\widehat{\mathbf{x}}-\mathbf{x}^*\right) \backsim N\left(0,\mathcal{I}^{-1}\right). \tag{3.2.9}$$

This can be formulated in terms of the variance to obtain a convergence rate result:

$$\mathcal{V}\left\{\widehat{\mathbf{x}}-\mathbf{x}^*\right\} = O\left(\frac{1}{n}\right). \tag{3.2.10}$$

Exercise 3.2.1 *Show that the nonlinear least squares estimator*

$$\widehat{\mathbf{x}}_n = \arg\min_{\mathbf{x}} \sum_{i=1}^{n} \left(y_i - \mu\left(\mathbf{x}, t_i\right)\right)^2,$$

where $y_i = \mu\left(\mathbf{x}^*, t_i\right) + \varepsilon_i,\ i = 1, 2, \cdots, n$, *and the* ε_i *are independent and have bounded variance, is consistent under the conditions for a designed experiment. This result is due to [52]. It provides an example of a consistent estimator which is not necessarily a maximum likelihood estimator . It is, however, a quasi-likelihood estimator (Remark 3.3.3)*

3.2.3 Working with the wrong likelihood

This subsection examines some consequences of the situation where the true density is $g\left(\mathbf{y}; \boldsymbol{\theta}, \mathbf{t}\right)$ but the likelihood calculations are performed using

$$\mathcal{L}_f\left(\mathbf{y}; \mathbf{x}, \mathbf{T}\right) = \sum_{\mathbf{t} \in \mathbf{T}} \log\left(f\left(\mathbf{y}; \boldsymbol{\theta}_t, \mathbf{t}\right)\right), \tag{3.2.11}$$

where $f\left(\mathbf{y}; \boldsymbol{\theta}_t, \mathbf{t}\right)$ is also a density and $f \neq g$. A first step is to find the system that would be satified by a limiting parameter vector as $n \to \infty$.

Lemma 3.3 *Assume that the problem of maximizing* $\mathcal{L}_f\left(\mathbf{y}; \mathbf{x}, \mathbf{T}_n\right)$ *has a solution* $\mathbf{x}_f^n$ *for all* n *large enough, and that* $\mathbf{x}_f^n \to \mathbf{x}_f$, $n \to \infty$. *Then* $\mathbf{x}_f$ *satisfies the system*

$$\int_{S(\mathbf{T})} \mathcal{E}^* \left\{ \frac{\nabla_{\mathbf{x}} f\left(\mathbf{y}; \boldsymbol{\theta}_t, \mathbf{t}\right)}{f\left(\mathbf{y}; \boldsymbol{\theta}_t, \mathbf{t}\right)} \right\} \rho\left(\mathbf{t}\right) d\mathbf{t} = 0. \tag{3.2.12}$$

Proof. The necessary conditions for a maximum of (3.2.11) give

$$0 = \frac{1}{n} \sum_{\mathbf{t} \in \mathbf{T}_n} \left\{ \frac{\nabla_{\mathbf{x}} f\left(\mathbf{y}_t; \boldsymbol{\theta}_t, \mathbf{t}\right)}{f\left(\mathbf{y}_t; \boldsymbol{\theta}_t, \mathbf{t}\right)} - \mathcal{E}^* \left\{ \frac{\nabla_{\mathbf{x}} f\left(\mathbf{y}; \boldsymbol{\theta}_t, \mathbf{t}\right)}{f\left(\mathbf{y}; \boldsymbol{\theta}_t, \mathbf{t}\right)} \right\} \right\} + \frac{1}{n} \sum_{\mathbf{t} \in \mathbf{T}_n} \mathcal{E}^* \left\{ \frac{\nabla_{\mathbf{x}} f\left(\mathbf{y}; \boldsymbol{\theta}_t, \mathbf{t}\right)}{f\left(\mathbf{y}; \boldsymbol{\theta}_t, \mathbf{t}\right)} \right\} \tag{3.2.13}$$

where it should be noted that the expectation is calculated using the true density. The desired result now follows by letting $n \to \infty$, and using the condition (3.2.1) for designed experiments. ■■

It will be an isolated maximizer if the Hessian of the limiting likelihood is positive definite. To consider this point further let

$$\mathcal{J}_f^n = \frac{1}{n} \nabla_{\mathbf{x}}^2 \mathcal{L}_f\left(\mathbf{y}; \mathbf{x}, \mathbf{T}_n\right).$$

Then the law of large numbers gives

$$\mathcal{J}_f^n - \mathcal{E}^* \left\{ \frac{1}{n} \nabla_{\mathbf{x}}^2 \mathcal{L}_f \left(\mathbf{y}; \mathbf{x}, \mathbf{T}_n \right) \right\} \to 0, \ n \to \infty.$$

Now

$$\mathcal{E}^* \left\{ \mathcal{J}_f^n \right\} = \mathcal{E}_f \left\{ \mathcal{J}_f^n \right\} + \frac{1}{n} \sum_{\mathbf{t} \in \mathbf{T}_n} \int_{\text{range}(\mathbf{y})} (g - f) \nabla_{\mathbf{x}} \left\{ \frac{\nabla_{\mathbf{x}} f^T}{f} \right\} d\mathbf{y}. \quad (3.2.14)$$

The first term on the right hand side is generically positive definite. To see this note that it follows from Lemma 3.2 that

$$\begin{aligned} \mathcal{E}_f \left\{ \mathcal{J}_f^n \right\} &= -\frac{1}{n} \mathcal{E}_f \left\{ \nabla_{\mathbf{x}} \mathcal{L}_f^T \nabla_{\mathbf{x}} \mathcal{L}_f \right\}, \quad (3.2.15) \\ &\to \int_{S(\mathbf{T})} \mathcal{E}_f \left\{ \nabla_{\mathbf{x}} L_f^T \nabla_{\mathbf{x}} L_f \right\} \rho(\mathbf{t}) \, d\mathbf{t}, \, n \to \infty, \end{aligned}$$

where $L_f = \log\left(f\left(\mathbf{y}; \boldsymbol{\theta}_t, \mathbf{t}\right)\right)$. Thus positive definiteness depends on the relative size of the second term. This can be evaluated using the following Lemma.

Lemma 3.4 *Let A be positive definite. Then $A \pm B$, where B is symmetric, is positive definite if $\varpi\left(A^{-1}B\right) < 1$ where ϖ denotes the spectral radius (magnitude of the eigenvalue of largest magnitude) of the indicated matrix.*

Proof. The condition to be satisfied is

$$\mathbf{v}^T (A \pm B) \mathbf{v} > 0, \ \forall \mathbf{v} \neq 0.$$

This is equivalent to

$$\mathbf{v}^T A^{1/2} A^{T/2} \mathbf{v} \pm \mathbf{v}^T A^{1/2} A^{-1/2} B A^{-T/2} A^{T/2} \mathbf{v} > 0, \ \forall \mathbf{v} \neq 0,$$

or to

$$\mathbf{u}^T (I \pm C) \mathbf{u} > 0, \ C = A^{-1/2} B A^{-T/2}, \ \forall \mathbf{u} \neq 0.$$

As C is symmetric, $C = T \Lambda T^T$ where Λ is the diagonal matrix of the eigenvalues, and T is orthogonal. Thus the condition on C reduces to the requirement

$$\mathbf{u}^T T (I \pm \Lambda) T^T \mathbf{u} > 0, \ \forall \mathbf{u} \neq 0,$$

and this is satisfied if and only if all elements of Λ are less than 1 in magnitude $\Rightarrow \varpi(C) < 1$. The desired result now follows as

$$A^{-1} B = A^{-T/2} C A^{T/2}$$

is similar to C. ■■

Theorem 3.5 *Let $\mathbf{x}_f$ be a solution of (3.2.12). Then $\mathbf{x}_f$ is an isolated solution corresponding to a maximum of the limiting problem provided*

$$\varpi \left\{ \begin{array}{c} \left(\int_{S(\mathbf{T})} \mathcal{E}_f \left\{ \nabla_{\mathbf{x}} L_f^T \nabla_{\mathbf{x}} L_f \right\} \rho(\mathbf{t})\, d\mathbf{t} \right)^{-1} \\ \left(\int_{S(\mathbf{T})} \left[\int_{\text{range}(\mathbf{y})} (g-f)\, \nabla_{\mathbf{x}}^2 L_f d\mathbf{y} \right] \rho(\mathbf{t})\, d\mathbf{t} \right) \end{array} \right\} < 1. \tag{3.2.16}$$

Proof. This follows by an application of Lemma 3.4 to (3.2.15) and (3.2.14). ■ ■

Remark 3.2.2 *This result can be used to show that the Newton iteration to maximize $\mathcal{L}_f(\mathbf{y}; \mathbf{x}, \mathbf{T}_n)$ converges to $\mathbf{x}_f^n$ from all starting points close enough to $\mathbf{x}_f$ provided n is large enough, and that $\lim_{n\to\infty} \mathbf{x}_f^n = \mathbf{x}_f$ provided (3.2.16) holds. This property is called consistency with the assumed probability model in [77].*

Conditions under which $\mathbf{x}_f = \mathbf{x}^*$ include:

1. The case $f = g$ is a consequence of Lemma 3.1 which gives

$$\begin{aligned} 0 &= \int_{S(\mathbf{T})} \mathcal{E}^* \left\{ \nabla_{\mathbf{x}} L(\mathbf{y}; \boldsymbol{\theta}^*, \mathbf{t}) \right\} \rho(\mathbf{t})\, d\mathbf{t} \\ &= \int_{S(\mathbf{T})} \mathcal{E}^* \left\{ \nabla_{\mathbf{x}} L_f(\mathbf{y}; \boldsymbol{\theta}(\mathbf{x}_f), \mathbf{t}) \right\} \rho(\mathbf{t})\, d\mathbf{t}. \end{aligned}$$

2. When a signal in noise model is assumed, so that the dependence of $\nabla_{\mathbf{x}} L_f(\mathbf{y}; \boldsymbol{\theta}(\mathbf{x}_f), \mathbf{t})$ on the signal is of the form $\mathbf{y}_t - \boldsymbol{\mu}(\mathbf{x}, \mathbf{t})$, and (3.1.10) holds. If f is normal then a nonlinear least squares problem results.

3.3 Quasi-likelihood formulations

The method of maximum likelihood makes the strong assumption that the exact distribution $g(\mathbf{y}; \boldsymbol{\theta}, \mathbf{t})$ is known. Thus it is of interest to ask if it is possible to preserve some at least of the good properties when this information is not available. If the new objective function is to be chosen to force consistency then analogy suggests an objective function having the generic form

$$\mathcal{K}(\mathbf{y}; \mathbf{x}, \mathbf{T}_n) = \sum_{\mathbf{t} \in \mathbf{T}_n} K_t(\mathbf{y}_t; \boldsymbol{\theta}_t, \mathbf{t}),$$

while meeting the requirements suggested by the form of Theorem 3.2 gives

$$\int_{S(\mathbf{T})} \mathcal{E}^* \left\{ \nabla_{\mathbf{x}} K_t \left(\mathbf{y}_t ; \boldsymbol{\theta}_t^*, \mathbf{t} \right) \right\} \rho \left(\mathbf{t} \right) d\mathbf{t} = \mathbf{0},$$
$$\int_{S(\mathbf{T})} \mathcal{E}^* \left\{ \nabla_{\mathbf{x}}^2 K_t \left(\mathbf{y}_t ; \boldsymbol{\theta}_t^*, \mathbf{t} \right) \right\} \rho \left(\mathbf{t} \right) d\mathbf{t} \text{ bounded, negative definite.}$$

The general theory of quasi-likelihood starts with the estimating equation

$$\nabla_{\mathbf{x}} \mathcal{K}_T \left(\mathbf{y} ; \mathbf{x}, \mathbf{T}_n \right) = 0, \tag{3.3.1}$$

where the $\nabla_{\mathbf{x}} K_t, \mathbf{t} \in \mathbf{T}_n$, satisfy the *differential* condition

$$\mathcal{E}^* \left\{ \nabla_{\mathbf{x}} K_t \left(\mathbf{y}_t ; \boldsymbol{\theta}_t^*, \mathbf{t} \right) \right\} = 0. \tag{3.3.2}$$

In particular, it is not assumed that an explicit form for K is known, only that its gradient is available . A general theory of optimal estimating equations has been developed under these constraints [34]. A less ambitious program characterised initial work on quasi-likelihood [112]. The idea is to work from the general form of likelihoods based on the exponential family (3.1.3), and to abstract a generic structure to be satisfied by $\nabla_{\mathbf{x}} K_t \left(\mathbf{y}_t ; \boldsymbol{\theta}_t, \mathbf{t} \right)$ from this. If the exponential family likelihood terms are given by

$$L_t = \left\{ \mathbf{y}_t^T \boldsymbol{\theta}_t - b \left(\theta_t \right) \right\} / a \left(\phi \right) + \log \left(c \left(\mathbf{y}_t, \phi \right) \right).$$

Then

$$\nabla_{\boldsymbol{\theta}} L_t = \left\{ \mathbf{y}_t^T - \nabla_{\boldsymbol{\theta}} b \left(\boldsymbol{\theta}_t \right) \right\} / a \left(\phi \right)$$

which has the form of a signal in noise model (3.1.10) with $\boldsymbol{\mu} = \nabla_{\boldsymbol{\theta}} b$ (cf example 3.1.2). The form of estimating equation adopted in [112] is based on

$$\nabla_{\boldsymbol{\mu}} K_t \left(\mathbf{y}_t ; \boldsymbol{\mu}_t \left(\mathbf{x} \right), \mathbf{t} \right)^T = V \left(\boldsymbol{\mu}_t \right)^{-1} \left(\mathbf{y}_t - \boldsymbol{\mu}_t \left(\mathbf{x} \right) \right), \tag{3.3.3}$$

where

$$\boldsymbol{\mu}_t \left(\mathbf{x}^* \right) = \mathcal{E}^* \left\{ \mathbf{y}_t \right\} \tag{3.3.4}$$

(so the model is exact) and $V \left(\boldsymbol{\mu}_t \right)$ is assumed to be a function of the mean only to avoid problems with nuisance parameters.

Remark 3.3.1 *If*

$$V \left(\boldsymbol{\mu} \right) = \mathcal{V} \left\{ \mathbf{y} - \boldsymbol{\mu} \right\} \tag{3.3.5}$$

then

$$\mathcal{E} \left\{ \nabla_{\boldsymbol{\mu}} K \right\} = 0,$$
$$\mathcal{E} \left\{ \nabla_{\boldsymbol{\mu}} K^T \nabla_{\boldsymbol{\mu}} K \right\} = -\mathcal{E} \left\{ \nabla_{\boldsymbol{\mu}}^2 K \right\} = V \left(\boldsymbol{\mu} \right)^{-1}.$$

These results correspond to those of Lemmas 3.1 and 3.2 which hold in the case of the likelihood. They guarantee optimality in the sense of [45] for estimating equations having the general form (3.3.1) with $\nabla_{\boldsymbol{\mu}} K$ given by (3.3.3). However, it is not necessary that (3.3.5) hold. If $V(\boldsymbol{\mu}) \neq \mathcal{V}\{\mathbf{y}\} = V^*$ then

$$\mathcal{E}\left\{\nabla_{\boldsymbol{\mu}} K^T \nabla_{\boldsymbol{\mu}} K\right\} = V^{-1} V^* V^{-1}.$$

Thus the analogue of the method of maximum likelihood seeks an estimate $\widehat{\mathbf{x}}_n$ by solving

$$\sum_{\mathbf{t}} \left(\nabla_{\mathbf{x}} \boldsymbol{\mu}_t\right)^T V\left(\boldsymbol{\mu}_t\right)^{-1} \left(\mathbf{y}_t - \boldsymbol{\mu}_t\left(\mathbf{x}\right)\right) = 0. \tag{3.3.6}$$

The consistency of the sequence of estimates $\{\widehat{\mathbf{x}}_n\}$ is now considered. Note first that as a consequence of $\mathcal{E}\{\mathbf{y}_t\} = \boldsymbol{\mu}_t(\mathbf{x}^*)$, K_t trivially satisfies (3.3.2), and that, from the form of (3.3.3), it follows that

$$\frac{1}{n}\nabla_{\mathbf{x}}^2 \mathcal{K}_T\left(\mathbf{y}; \mathbf{x}^*, \mathbf{T}_n\right) = -\frac{1}{n}\sum_{\mathbf{t}} \left\{ \begin{array}{c} \left(\nabla_{\mathbf{x}} \boldsymbol{\mu}_t\right) V_t^{-1} \nabla_{\mathbf{x}} \boldsymbol{\mu}_t \\ + \left(\left(\nabla_{\mathbf{x}} \boldsymbol{\mu}_t\right)^T \nabla_{\mathbf{x}} V_t^{-1} + \nabla_{\mathbf{x}}^2 \boldsymbol{\mu}_t^T V_t^{-1}\right)\left(\mathbf{y}_t^T - \boldsymbol{\mu}_t\right) \end{array} \right\} \tag{3.3.7}$$

$$\rightarrow -\int_{\mathcal{S}(\mathbf{T})} \left(\nabla_{\mathbf{x}} \boldsymbol{\mu}_t\right)^T V_t^{-1} \nabla_{\mathbf{x}} \boldsymbol{\mu}_t \rho\left(\mathbf{t}\right) d\mathbf{t}. \tag{3.3.8}$$

Let this limit be written $-\mathcal{I}_K$. The requirement that $\mathcal{I}_K$ be positive definite does not impose a severe restriction. Also, given consistency, this requirement serves to guarantee that $\mathbf{x}^*$ is an isolated limit point. The proof for quasi-likelihood estimates can follow the approach based on the Kantorovich Theorem (Theorem 3.3) used in the likelihood case. Here the argument that produces the limiting equation gives:

$$\frac{1}{n}\nabla_x \mathcal{K}\left(\mathbf{y}; \mathbf{x}, \mathbf{T}_n\right) \rightarrow \int_{S(\mathbf{T})} \nabla_x \boldsymbol{\mu}_t V\left(\boldsymbol{\mu}_t\right)^{-1} \left(\boldsymbol{\mu}_t(\mathbf{x}^*) - \boldsymbol{\mu}_t(\mathbf{x})\right) \rho(\mathbf{t}) d\mathbf{t}. \tag{3.3.9}$$

Here (3.3.4) is used to evaluate $\mathcal{E}^*\{\nabla_x \mathcal{K}\}$. Note that the limiting equation has the solution $\mathbf{x}^*$ essentially independent of the choice of V.

Theorem 3.6 (Consistency of quasi-likelihood) *Let the quasi-likelihood estimation problem associated with a regular sampling regime have a well determined solution in the sense that $\mathcal{I}_K = \int_{\mathcal{S}(\mathbf{T})} \left(\nabla_{\mathbf{x}} \boldsymbol{\mu}_t\right)^T V_t^{-1} \nabla_{\mathbf{x}} \boldsymbol{\mu}_t \rho\left(\mathbf{t}\right) d\mathbf{t}$. is bounded, positive definite when $\mathbf{x} = \mathbf{x}^*$. Then for each sequence of designed experiments $\{\mathbf{T}_n\}$ there exists an n_0 such that, for almost all $n > n_0$, the Newton iteration applied to (3.3.1) and started at $\mathbf{x}^*$ converges to a solution $\widehat{\mathbf{x}}_n$ and $\widehat{\mathbf{x}}_n \overset{a.s.}{\rightarrow} \mathbf{x}^*$ as $n \rightarrow \infty$.*

The direct analogue of the argument used to show asymptotic normality of the distribution of $\sqrt{n}\left(\widehat{\mathbf{x}}_n - \mathbf{x}^*\right)$ where $\widehat{\mathbf{x}}_n$ is the estimate computed by maximizing the likelihood can be carried through in this case as well by expanding the estimating equation about $\mathbf{x}^*$ by Taylor's Theorem. The result is

$$\sqrt{n}\left(\widehat{\mathbf{x}}_n - \mathbf{x}^*\right) \backsim N\left(0, \mathcal{I}_K^{-1}\left[\int_{\mathcal{S}(\mathbf{T})} \left(\nabla_{\mathbf{x}}\boldsymbol{\mu}_t\right)^T \mathcal{V}\left\{K_t\right\} \nabla_{\mathbf{x}}\boldsymbol{\mu}_t \rho\left(\mathbf{t}\right) d\mathbf{t}\right] \mathcal{I}_K^{-1}\right)$$

where

$$\mathcal{V}\left\{K_t\right\} = V_t^{-1}\mathcal{V}\left\{\mathbf{y}_t\right\} V_t^{-1}.$$

Remark 3.3.2 *Starting from a derivative based definition has its disadvantages. Integration to find the form of the objective function requires knowledge of the "constants of integration" - for example, the term* $\log\left(c\left(\mathbf{y}_t, \phi\right)\right)$ *in the exponential family case considered above. This requires additional information. Such information is likely needed for estimation of scale, the typical role of the auxilliary parameter* ϕ *in this example. In [69] it is suggested that*

$$\log\left(c\left(\mathbf{y}_t, \phi\right)\right) = -\frac{1}{2} h_1\left(\sigma^2\right) - h_2\left(\mathbf{y}\right),$$

and that, under suitable restrictions on the size of σ^2 *and higher order cumulants, the estimate* $h_1\left(\sigma^2\right) = \sigma^2$ *is approximately valid. An alternative approach is to estimate the covariance matrix separately using the results of the quasi-likelihood computation. One possibility [119] is the sample variance*

$$\mathcal{V}\left\{\widehat{\mathbf{x}}_n\right\} \leftarrow V_1^{-1} V_2 V_1^{-1}$$

where

$$V_1 = \frac{1}{n}\sum_{\mathbf{t}} \left(\nabla_{\mathbf{x}}\widehat{\boldsymbol{\mu}}_t\right)^T \widehat{V}_t^{-1} \nabla_{\mathbf{x}}\widehat{\boldsymbol{\mu}}_t,$$

$$V_2 = \frac{1}{n}\sum_{\mathbf{t}} \left(\nabla_{\mathbf{x}}\widehat{\boldsymbol{\mu}}_t\right)^T \widehat{V}_t^{-1} \left(\mathbf{y}_t - \widehat{\boldsymbol{\mu}}_t\right)\left(\mathbf{y}_t - \widehat{\boldsymbol{\mu}}_t\right)^T \widehat{V}_t^{-1} \nabla_{\mathbf{x}}\widehat{\boldsymbol{\mu}}_t,$$

and hats indicate evaluation at the quasi-likelihood estimate. This estimate is consistent even when $V_t \neq \mathcal{V}\left\{\mathbf{y}_t\right\}$*. Its robustness is discussed in [114].*

Remark 3.3.3 *The best known example of a quasi-likelihood is given by the method of nonlinear least squares. Here*

$$\nabla_{\boldsymbol{\mu}} K_t\left(\mathbf{y}_t; \boldsymbol{\mu}_t\left(\mathbf{x}\right), \mathbf{t}\right)^T = \mathbf{y}_t - \boldsymbol{\mu}_t\left(\mathbf{x}\right)$$

corresponding to $V(\boldsymbol{\mu}) = I$. This method provides a consistent estimator under quite general conditions (adequate smoothness, density with bounded second moment, and a sampling regime consistent with (3.2.1)) [52]. It is not required that $V(\boldsymbol{\mu})$ be a consistent estimator of the true covariance matrix, but the cost could well be a serious loss of efficiency. If the error distribution is not normal then nonlinear least squares provides an example of the use of the wrong density in a likelihood calculation.

3.4 Equality constrained likelihood

The consequences for the estimation problem resulting from the imposition of a (fixed) finite number m of equality constraints on the the likelihood function are investigated in this section. One possible approach is to use the constraints to eliminate variables, and this is attractive when the elimination is straightforward. Here it is assumed that the elimination is not straightforward and that the constraints must be handled explicitly by the use of Lagrange multiplier techniques. The development follows a similar sequence of steps to that in the analysis of the unconstrained likelihood. That is, first a limiting form of the necessary conditions characterizing the maximum of the constrained likelihood as $n = |\mathbf{T}_n| \to \infty$ is found. Then the Kantorovich Theorem can be used to demonstrate consistency of the estimates, and the corresponding asymptotic distributions are derived.

The constrained form of the likelihood problem is written

$$\max_{\mathbf{x},\mathbf{c}(\mathbf{x})=0} \mathcal{L}(\mathbf{y};\mathbf{x},\mathbf{T}_n), \tag{3.4.1}$$

where $\mathbf{c}(\mathbf{x}) \in R^p \to R^m$, $m < p$, is the constraint vector. It is assumed to be smooth enough (at least twice continuously differentiable) in a set $B_\tau = \{\|\mathbf{x} - \mathbf{x}^*\| < \tau\}$ containing the true parameter vector $\mathbf{x}^*$, and the constraints are assumed to be independent in the sense that $\operatorname{rank}(\nabla_{\mathbf{x}}\mathbf{c}(\mathbf{x})) = m$, $\mathbf{x} \in B_\tau$. The necessary condition for a maximum of (3.4.1) are

$$\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{y};\mathbf{x},\mathbf{T}_n) = \boldsymbol{\zeta}^T \nabla_{\mathbf{x}}\mathbf{c}(\mathbf{x}), \tag{3.4.2}$$

$$\mathbf{c}(\mathbf{x}) = 0, \tag{3.4.3}$$

where $\boldsymbol{\zeta}$ is the vector of Lagrange multipliers. To use the law of large numbers, equation (3.4.2) is written as

$$\frac{1}{n}\left\{\nabla_{\mathbf{x}}\mathcal{L} - \mathcal{E}^*\left\{\nabla_{\mathbf{x}}\mathcal{L}\right\}\right\} + \frac{1}{n}\mathcal{E}^*\left\{\nabla_{\mathbf{x}}\mathcal{L}\right\} = \frac{\boldsymbol{\zeta}^T}{n}\nabla_{\mathbf{x}}\mathbf{c}.$$

This permits the limiting form of (3.4.2), (3.4.3) to be written

$$\int_{S(\mathbf{T})} \mathcal{E}^* \left\{ \nabla_{\mathbf{x}} L_t \left(\mathbf{y}; \boldsymbol{\theta}_t, \mathbf{t} \right) \right\} \rho \left(\mathbf{t} \right) d\mathbf{t} = \widetilde{\boldsymbol{\zeta}}^T \nabla_{\mathbf{x}} \mathbf{c} \left(\mathbf{x} \right), \qquad (3.4.4)$$

$$\mathbf{c} \left(\mathbf{x} \right) = 0,$$

where $\widetilde{\boldsymbol{\zeta}}$ is the limit of the scaled sequence of multiplier estimates $\left\{ \widetilde{\boldsymbol{\zeta}}_n = \widehat{\boldsymbol{\zeta}}_n / n \right\}$.

Remark 3.4.1 *It is tempting to conclude that this system has the solution*

$$\mathbf{x} = \mathbf{x}^*, \ \widetilde{\boldsymbol{\zeta}} = 0, \qquad (3.4.5)$$

as a consequence of Lemma 3.1 and the linear independence of the constraint gradients. However, the reversal of the order of integration and differentiation in Lemma 3.1 requires that the optimum be properly in the interior of the allowable set of independent variables, a property at variance with the equality constraint conditions. To see how this works out consider the variation

$$\mathcal{G} \left(\mathbf{y}, \mathbf{x} + \delta\mathbf{x}, T_n \right) - \mathcal{G} \left(\mathbf{y}, \mathbf{x}, T_n \right) = \nabla_{\mathbf{x}} \mathcal{L} \mathcal{G} \delta\mathbf{x} + O \left(\| \delta\mathbf{x} \|^2 \right),$$

where the constraint $\mathbf{c} = 0$ *forces the condition*

$$\nabla_{\mathbf{x}} \mathbf{c} \delta\mathbf{x} = 0.$$

Let

$$\nabla_{\mathbf{x}} \mathbf{c}^T = \left[\begin{array}{cc} Q_1 & Q_2 \end{array} \right] \left[\begin{array}{c} U \\ 0 \end{array} \right].$$

Then

$$\delta\mathbf{x} = Q_2 \mathbf{h}, \ \| \delta\mathbf{x} \| = \| \mathbf{h} \|.$$

Using this information to differentiate under the integral sign now gives the modified condition:

$$\mathcal{E} \left\{ \nabla_{\mathbf{x}} \mathcal{L} \left(\mathbf{y}, \mathbf{x}, T_n \right) \right\} Q_2 = 0. \qquad (3.4.6)$$

This is directly compatible with (3.4.2) in the sense of leading directly to the same multiplier condition as (3.4.6) implies that for some $\boldsymbol{\eta}$

$$\mathcal{E} \left\{ \nabla_{\mathbf{x}} \mathcal{L} \left(\mathbf{y}, \mathbf{x}, T_n \right) \right\} = \boldsymbol{\eta}^T Q_1^T.$$

Note Q_2 *is a constant matrix if the constraint vector is linear. If the constraint vector is nonlinear then* Q_2 *can be replaced by a projection onto the null space of* $\nabla_{\mathbf{x}} \mathbf{c}^T$*, and in this form it can have a smooth (differentiable) representation. The corresponding form for the conclusion of Lemma 3.2 is*

$$Q_2^T \mathcal{E} \left\{ \nabla_{\mathbf{x}}^2 \mathcal{L} \right\} Q_2 = -Q_2^T \mathcal{E} \left\{ \nabla_{\mathbf{x}} \mathcal{L}^T \nabla_{\mathbf{x}} \mathcal{L} \right\} Q_2 - \dot{Q}_2^{\ T} \mathcal{E} \left\{ \nabla_{\mathbf{x}} \mathcal{L} \right\}^T, \qquad (3.4.7)$$

where the extra term on the right hand side is obtained by differentiating the transformation.

To see that the limiting Lagrange multiplier equations (3.4.4) are satisfied by the true solution $\mathbf{x}^*$ note that the independence of the constraints implies that a subset of m of the variables can be expressed in terms of the remainder using the implicit function theorem . After possible reordering the result is the system of equations

$$\mathbf{c}\left(\mathbf{u}\left(\mathbf{v}\right),\mathbf{v}\right)=0.$$

To construct a representation of the constraint gradient null space differentiate this equation. The result is

$$\nabla_{\mathbf{u}}\mathbf{c}\frac{\partial\mathbf{u}}{\partial\mathbf{v}}+\nabla_{\mathbf{v}}\mathbf{c}=0,$$

showing that

$$\left[\begin{array}{cc}\nabla_{\mathbf{u}}\mathbf{c} & \nabla_{\mathbf{v}}\mathbf{c}\end{array}\right]\left[\begin{array}{c}\frac{\partial\mathbf{u}}{\partial\mathbf{v}}\\ I\end{array}\right]=0. \tag{3.4.8}$$

The log likelihood expressed in terms of the independent variables $\mathbf{v}$ now has an unconstrained maximum. The limiting form (3.2.3) of the corresponding necessary conditions in this case is

$$\begin{aligned}\int_{S(\mathbf{T})}\mathcal{E}^*\left\{\nabla_{\mathbf{u}}L_t\frac{\partial\mathbf{u}}{\partial\mathbf{v}}+\nabla_{\mathbf{v}}L_t\right\}\rho dt&=\int_{S(\mathbf{T})}\mathcal{E}^*\left\{\nabla_{(\mathbf{u},\mathbf{v})}L_t\right\}\rho dt\left[\begin{array}{c}\frac{\partial\mathbf{u}}{\partial\mathbf{v}}\\ I\end{array}\right],\\&=0.\end{aligned}$$

This is equivalent to (3.4.6.

Equations (3.4.2) and (3.4.3) have an isolated solution provided the Jacobian of the system, scaled by dividing by n, is nonsingular for all n large enough. Let the Lagrangian function corresponding to this system be

$$\mathbf{L}_n\left(\mathbf{x},\boldsymbol{\zeta}\right)=\frac{1}{n}\mathcal{L}\left(\mathbf{y};\mathbf{x},\mathbf{T}_n\right)-\boldsymbol{\zeta}^T\mathbf{c}\left(\mathbf{x}\right).$$

If the Jacobian is written as $\mathcal{A}_n\left(\mathbf{x},\boldsymbol{\zeta}\right)$ (the terminology anticipates the standard notation which refers to the Jacobian as the augmented matrix in this case) the condition for an isolated solution for n large enough becomes

$$\begin{aligned}\mathcal{A}_n\left(\mathbf{x}^*,\boldsymbol{\zeta}\right)&=\left[\begin{array}{cc}\nabla_{\mathbf{x}}^2\mathbf{L}_n\left(\mathbf{x}^*,\boldsymbol{\zeta}\right) & -\nabla_{\mathbf{x}}\mathbf{c}\left(\mathbf{x}^*\right)^T\\ -\nabla_{\mathbf{x}}\mathbf{c}\left(\mathbf{x}^*\right) & \end{array}\right],\\ \operatorname{rank}\left(\mathcal{A}_n\left(\mathbf{x}^*,\boldsymbol{\zeta}\right)\right)&=m+p.\end{aligned}$$

This requirement is close to the one met before in Condition 1.1. Setting

$$\begin{aligned}Q\left[\begin{array}{c}U\\0\end{array}\right]&=\nabla_{\mathbf{x}}\mathbf{c}\left(\mathbf{x}^*\right)^T,\\ V&=\frac{1}{n}Q^T\nabla_{\mathbf{x}}^2\mathbf{L}_nQ,\end{aligned}$$

gives the condition in the equivalent form

$$U \text{ nonsingular, } V_{22} = Q_2^T \nabla_{\mathbf{x}}^2 \mathbf{L}_n Q_2 \text{ nonsingular.}$$

The first condition guarantees that the constraints are linearly independent in the case that they are linear, and it provides the obvious extension of independence if they are nonlinear by ensuring that they are (locally) not contradictory. The second condition is satisfied if V_{22} is positive definite. In this form positive definiteness of the restriction of $\frac{1}{n}\nabla_{\mathbf{x}}^2 \mathbf{L}_n$ to the tangent space of the constraints corresponds to the second order sufficiency condition [73].

Theorem 3.7 *Let the estimation problem associated with a regular sampling regime have a well determined solution in the sense that the augmented matrix* $\mathcal{A}_n\left(\mathbf{x}^*, \widetilde{\boldsymbol{\zeta}}\right)$ *has its full rank. Then for each sequence of experiments* $\{\mathbf{T}_n\}$ *there exists an* n_0 *such that the Newton iteration started at* $\left(\mathbf{x}^*, \widetilde{\boldsymbol{\zeta}}\right)$ *converges to* $\left(\widehat{\mathbf{x}}_n, \widetilde{\boldsymbol{\zeta}}_n\right)$ *satisfying the necessary conditions (3.4.2), (3.4.3) for almost all* $n > n_0$*, and*

$$\left(\widehat{\mathbf{x}}_n \overset{a.s.}{\to} \mathbf{x}^*, \widetilde{\boldsymbol{\zeta}}_n \overset{a.s.}{\to} \widetilde{\boldsymbol{\zeta}}\right), \; n \to \infty. \tag{3.4.9}$$

Proof. To use the Newton's method based approach to the question of consistency note that the predicted correction given the initial guess $\mathbf{x}_0 = \mathbf{x}^*$, $\widetilde{\boldsymbol{\zeta}}_0 = \widetilde{\boldsymbol{\zeta}}$ satisfies the system of linear equations

$$\mathcal{A}_n\left(\mathbf{x}^*, \widetilde{\boldsymbol{\zeta}}\right) \begin{bmatrix} \mathbf{h} \\ \Delta\widetilde{\boldsymbol{\zeta}} \end{bmatrix} = \begin{bmatrix} -\frac{1}{n}\nabla_{\mathbf{x}}\mathcal{L}\left(\mathbf{y}; \mathbf{x}^*, \mathbf{T}_n\right)^T + \nabla_{\mathbf{x}}\mathbf{c}^T\widetilde{\boldsymbol{\zeta}} \\ 0 \end{bmatrix}, \tag{3.4.10}$$

where $\mathbf{h} = \mathbf{x} - \mathbf{x}^*$. If we assume that the conditions for the nonsingularity of the augmented matrix are satisfied and the problem data are smooth enough then the Kantorovich conditions ensure that the Newton iteration proceeds satisfactorily provided the right hand side of (3.4.10) is small almost surely for all n large enough. This follows by the same law of large numbers argument as before. ■ ■

To derive the asymptotic distributions let $\left(\widehat{\mathbf{x}}, \widetilde{\boldsymbol{\zeta}}_n\right)$ satisfy the estimating equations (3.4.2), (3.4.3). If these are expanded by Taylor series about $\left(\mathbf{x}^*, \widetilde{\boldsymbol{\zeta}}\right)$ then we obtain

$$\begin{aligned} 0 &= \nabla_{\mathbf{x}}\mathbf{L}_n\left(\mathbf{x}^*, \widetilde{\boldsymbol{\zeta}}\right) + \left(\widehat{\mathbf{x}} - \mathbf{x}^*\right)^T \nabla_{\mathbf{x}}^2 \mathbf{L}_n\left(\mathbf{x}^*, \widetilde{\boldsymbol{\zeta}}\right) + \frac{1}{2}\nabla_{\mathbf{x}}^3 \mathbf{L}_n\left(\overline{\mathbf{x}}, \widetilde{\boldsymbol{\zeta}}\right)\left(\widehat{\mathbf{x}} - \mathbf{x}^*, \widehat{\mathbf{x}} - \mathbf{x}^*\right) \\ &\quad - \Delta\boldsymbol{\zeta}_n^T\left\{\nabla_{\mathbf{x}}\mathbf{c}\left(\mathbf{x}^*\right) + \frac{1}{2}\nabla_{\mathbf{x}}^2\mathbf{c}\left(\overline{\mathbf{x}}\right)\left(\cdot, \widehat{\mathbf{x}} - \mathbf{x}^*\right)\right\}, \\ 0 &= \nabla_{\mathbf{x}}\mathbf{c}\left(\mathbf{x}^*\right)\left(\widehat{\mathbf{x}} - \mathbf{x}^*\right) + \frac{1}{2}\nabla_{\mathbf{x}}^2\mathbf{c}\left(\overline{\mathbf{x}}\right)\left(\widehat{\mathbf{x}} - \mathbf{x}^*, \widehat{\mathbf{x}} - \mathbf{x}^*\right), \end{aligned}$$

where $\overline{\mathbf{x}}$ indicates that a mean value is appropriate. Ignoring the almost surely small terms

$$\nabla_{\mathbf{x}}^3 \mathbf{L}_n \left(\overline{\mathbf{x}}, \widetilde{\boldsymbol{\zeta}} \right) \left(\widehat{\mathbf{x}} - \mathbf{x}^*, \widehat{\mathbf{x}} - \mathbf{x}^* \right), \Delta\boldsymbol{\zeta}_n^T \nabla_{\mathbf{x}}^2 \mathbf{c} \left(\overline{\mathbf{x}} \right) \left(\cdot, \widehat{\mathbf{x}} - \mathbf{x}^* \right), \nabla_{\mathbf{x}}^2 \mathbf{c} \left(\overline{\mathbf{x}} \right) \left(\widehat{\mathbf{x}} - \mathbf{x}^*, \widehat{\mathbf{x}} - \mathbf{x}^* \right)$$

gives a system of equations basically similar to (3.4.10) . Making use of the orthogonal factorization of $\nabla_{\mathbf{x}} \mathbf{c} \left(\mathbf{x}^* \right)^T$ gives for $\widehat{\mathbf{x}} - \mathbf{x}^*$ the equations

$$\begin{aligned} Q_1^T \left(\widehat{\mathbf{x}} - \mathbf{x}^* \right) &= 0, \\ Q_2^T \nabla_{\mathbf{x}}^2 \mathbf{L}_n Q_2 Q_2^T \left(\widehat{\mathbf{x}} - \mathbf{x}^* \right) &= -Q_2^T \nabla_{\mathbf{x}} \mathbf{L}_n^T, \\ &= -\frac{1}{n} Q_2^T \nabla_{\mathbf{x}} \mathcal{L}_n^T, \end{aligned}$$

and these have the solution

$$\begin{aligned} \widehat{\mathbf{x}} - \mathbf{x}^* &= -\frac{1}{n} Q_2 \left(Q_2^T \nabla_{\mathbf{x}}^2 \mathbf{L}_n Q_2 \right)^{-1} Q_2^T \nabla_{\mathbf{x}} \mathcal{L} \left(\mathbf{y}; \mathbf{x}^*, \mathbf{T}_n \right)^T, \\ &= -\frac{1}{n} \nabla_{\mathbf{x}}^2 \mathbf{L}_n^{-1} P_n \nabla_{\mathbf{x}} \mathcal{L} \left(\mathbf{y}; \mathbf{x}^*, \mathbf{T}_n \right)^T, \end{aligned} \tag{3.4.11}$$

where P_n is the oblique projector

$$P_n = \nabla_{\mathbf{x}}^2 \mathbf{L}_n Q_2 \left(Q_2^T \nabla_{\mathbf{x}}^2 \mathbf{L}_n Q_2 \right)^{-1} Q_2^T. \tag{3.4.12}$$

Because $\nabla_{\mathbf{x}}^2 \mathbf{L}_n$ converges almost surely as $n \to \infty$ as a consequence of consistency so does $P_n \to P$. The equation for the multiplier vector is

$$\begin{aligned} -U\Delta\widetilde{\boldsymbol{\zeta}}_n + Q_1^T \nabla_{\mathbf{x}}^2 \mathbf{L}_n Q_2 Q_2^T \left(\widehat{\mathbf{x}} - \mathbf{x}^* \right) &= -Q_1^T \nabla_{\mathbf{x}} \mathbf{L}_n^T + U\widetilde{\boldsymbol{\zeta}}, \\ &= -\frac{1}{n} Q_1^T \nabla_{\mathbf{x}} \mathcal{L}_n^T. \end{aligned}$$

The solution to this equation is

$$\begin{aligned} \widetilde{\boldsymbol{\zeta}}_n - \widetilde{\boldsymbol{\zeta}} &= U^{-1} Q_1^T \nabla_{\mathbf{x}}^2 \mathbf{L}_n Q_2 Q_2^T \left(\widehat{\mathbf{x}} - \mathbf{x}^* \right) + \frac{1}{n} U^{-1} Q_1^T \nabla_{\mathbf{x}} \mathcal{L}_n^T, \\ &= \frac{1}{n} U^{-1} Q_1^T \left(I - P_n \right) \nabla_{\mathbf{x}} \mathcal{L} \left(\mathbf{y}; \mathbf{x}^*, \mathbf{T}_n \right)^T. \end{aligned} \tag{3.4.13}$$

Theorem 3.8 *The asymptotic distributions are as follows:*

1. $\sqrt{n} \left(\widehat{\mathbf{x}} - \mathbf{x}^* \right) \sim N \left(0, \ V_x \right)$ *where*

$$V_x = Q_2 V_L^{-1} Q_2^T \mathcal{I} Q_2 V_L^{-1} Q_2^T$$

and V_L is the almost sure limit of $Q_2^T \nabla_{\mathbf{x}}^2 \mathbf{L}_n Q_2$.

2. $\sqrt{n}\left(\widetilde{\boldsymbol{\zeta}}_n - \widetilde{\boldsymbol{\zeta}}\right) \sim N\left(0, V_\zeta\right)$ *where*

$$V_\zeta = U^{-1} Q_1^T \left(I - P\right) \mathcal{I} \left(I - P\right)^T Q_1 U^{-T}.$$

Proof. That the limiting distributions have mean 0 follows from the central limit theorem applied to $\frac{1}{\sqrt{n}} \nabla_{\mathbf{x}} \mathcal{L}\left(\mathbf{y}; \mathbf{x}^*, \mathbf{T}_n\right)$ via an application of Slutsky's theorem [98]. To compute the variance estimates, the same approach allows the substitution of almost sure limits with small error. ■■

3.5 Separable regressions

Estimation problems where the basic data has the particular form

$$\mathbf{y} = \Phi\left(\boldsymbol{\beta}\right) \boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{3.5.1}$$

are said to be separable with the α_i, $i = 1, 2, \cdots, p$, the linear parameters , and the components of $\boldsymbol{\beta} \in R^m$ the nonlinear parameters . They have received a fair deal of attention, and a number of special methods which first eliminate the linear parameters and then solve a reduced optimization problem to recover an estimate of $\boldsymbol{\beta}$ have been proposed (an early reference is [37]and a recent survey is [38]). One possible approach notes that (3.5.1) is the general solution of a linear ordinary differential equation of order p with fundamental solutions given by the $\phi_i\left(t, \boldsymbol{\beta}\right)$, $i = 1, 2, \cdots, p$. This is discussed in Chapter 5. Two approaches are considered here. Both use orthogonal projection to eliminate the linear parameters from the objective function. They differ in the mode of constructing this projection.

Typical sets of model functions include:

1. Sets of exponentials (here $m = p$)

$$\phi_j\left(t, \boldsymbol{\beta}\right) = e^{-\beta_j t},\ j = 1, 2, \cdots, p.$$

2. Rational functions

$$\phi_j\left(t, \boldsymbol{\beta}\right) = \frac{t^{j-1}}{\sum_{i=1}^m \beta_i t^{i-1}},\ j = 1, 2, \cdots, p.$$

   Here the scale of the β_j remains to be fixed as multiplying numerator and denominator of the rational function by an arbitrary constant does not change $y\left(t\right)$.

To illustrate the separation of the calculation of the linear and nonlinear parameters assume that the observation errors are independent, $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \cdots, n$, and let $Q(\boldsymbol{\beta})$ be an orthogonal matrix such that

$$\left[\begin{array}{cc} Q_1(\boldsymbol{\beta}) & Q_2(\boldsymbol{\beta}) \end{array} \right]^T \Phi(\boldsymbol{\beta}) = \left[\begin{array}{c} U \\ 0 \end{array} \right], \tag{3.5.2}$$

where $\Phi(\boldsymbol{\beta})_{ji} = \phi_i(t_j, \boldsymbol{\beta})$ is assumed to have its full column rank p. Then maximum likelihood estimation, ignoring constant terms including the variance contribution, involves minimizing the sum of squares of residuals

$$\begin{aligned} \mathbf{r}^T\mathbf{r} &= \mathbf{r}^T Q Q^T \mathbf{r}, \\ &= \left\| U\boldsymbol{\alpha} - Q_1^T\mathbf{b} \right\|^2 + \left\| Q_2^T\mathbf{b} \right\|^2, \end{aligned}$$

where $\mathbf{b}$ with components $b_i = y_i + \varepsilon_i$, $i = 1, 2, \cdots, n$ is the signal observed in the presence of noise. A point to notice here is that the noise in the observations is additive, but this property is not preserved in the sum of squares term $\left\| Q_2^T\mathbf{b} \right\|^2$ as the matrix multiplication couples the nonlinear parameters and the noise. Note also that $\boldsymbol{\alpha}$ appears only in the first term on the right hand side. This term can be reduced to zero for all $\boldsymbol{\beta}$ such that $\Phi(\boldsymbol{\beta})$ has full column rank by setting

$$\boldsymbol{\alpha}(\boldsymbol{\beta}) = U^{-1} Q_1^T \mathbf{b}. \tag{3.5.3}$$

Thus

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathbf{r}^T\mathbf{r} = \min_{\boldsymbol{\beta}} \left\| Q_2^T\mathbf{b} \right\|^2 = \min_{\boldsymbol{\beta}} \|P\mathbf{b}\|^2 \tag{3.5.4}$$

where

$$P = Q_2 Q_2^T = I - \Phi\left(\Phi^T\Phi\right)^{-1}\Phi^T \tag{3.5.5}$$

projects onto the orthogonal complement of the range of Φ. Thus the problem has been reduced to a nonlinear least squares problem with a smaller number of parameters. The objective function in the reduced problem is referred to as the *variable projection functional*. There is identity between the minimizers of the full log likelihood and the solution computed from the variable projection formulation (3.5.4). Hence the solution of the reduced problem is consistent.

The key step in eliminating the linear parameters in equation (3.5.4) is the construction of an orthogonal basis for the null space of Φ^T. This provides an explicit form for the orthogonal projector P, and the problem reduces to minimizing $\|P\mathbf{b}\|^2$. Any other method of constructing P can be used, and this provides a way to use additional information on the parametrization.

The reduction is essentially that used in (3.5.4). Let P be the orthogonal projection onto the null space. Then

$$\begin{aligned}\mathbf{r}^T\mathbf{r} &= \mathbf{r}^T(I-P)\mathbf{r} + \mathbf{r}^T P\mathbf{r},\\ &= \mathbf{r}^T\Phi\left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{r} + \|P\mathbf{b}\|^2. \end{aligned} \tag{3.5.6}$$

The projection onto the range of Φ gives a nonnegative term which is minimized when $\boldsymbol{\alpha}(\boldsymbol{\beta})$ satisfies the normal equations so the problem reduces to determining the nonlinear parameters by minimizing $\|P\mathbf{b}\|^2$. A key property which permits the direct elimination of the linear parameters in each of the two examples quoted above is an explicit difference equation linking $p+1$ consecutive values of each of the model functions computed on an equispaced grid with spacing increment τ.

Example 3.5.1 *For the exponential fitting problem this difference equation has the form*

$$\sum_{j=1}^{p+1}\gamma_j e^{-\beta_j(t_k+(j-1)\tau)} = 0,\ k = 1, 2, \cdots, n-p.$$

Here this introduces an alternative parametrization $\boldsymbol{\gamma}$ replacing the original nonlinear parameters , and the β_j, $j = 1, 2, \cdots, p$ are recovered from the roots λ_k, $k = 1, 2, \cdots, p$, of the polynomial equation

$$\sum_{j=1}^{p+1}\gamma_j\lambda^{j-1} = 0$$

by using the relation

$$\lambda_k = e^{-\beta_k\tau},\ k = 1, 2, \cdots, p.$$

The γ_i are the elementary symmetric functions of the λ_k up to an arbitrary scalar multiplier.

Example 3.5.2 *In the rational fitting example the difference equation is given by*

$$\Delta^p\sum_{i=1}^{p}\left\{\left(\sum_{j=1}^{m}\beta_j t_k^{j-1}\right)\alpha_i\phi_i(t_k,\boldsymbol{\beta})\right\} = 0,\ k = 1, 2, \cdots, n-p.$$

where Δ is the forward difference operator defined by

$$\Delta\phi(t) = \phi(t+\tau) - \phi(t).$$

In contrast to the previous example, the nonlinear parameters $\boldsymbol{\beta}$ appear explicitly in the difference equation.

Assume now that a difference equation with coefficients F_k exists satisfying the conditions

$$\sum_{k=1}^{p+1} F_k\left(t_i, \boldsymbol{\gamma}\left(\boldsymbol{\beta}\right)\right) \phi_j\left(t_i + \left(k-1\right)\tau, \boldsymbol{\beta}\right) = 0,\ i = 1, 2, \cdots, n-p,\ j = 1, 2, \cdots, p,$$

and having the property that it is linear in the parametrization $\boldsymbol{\gamma}$ as in the above examples. Application to the data gives

$$\sum_{i=k}^{p+1} F_k\left(t_i, \boldsymbol{\gamma}\left(\boldsymbol{\beta}\right)\right) b_{i+k-1} = \sum_{k=1}^{p+1} F_k\left(t_i, \boldsymbol{\gamma}\left(\boldsymbol{\beta}\right)\right) \varepsilon_{i+k-1},\ i = 1, 2, \cdots, n-p.$$

In matrix form this is

$$F\left(\boldsymbol{\gamma}\right)\mathbf{b} = F\left(\boldsymbol{\gamma}\right)\boldsymbol{\varepsilon},\ F \in R^n \to R^{n-p},$$

where F is a $p+1$ banded, rectangular matrix which is linear and homogeneous in $\boldsymbol{\gamma}$. The property that F_k is exactly $p+1$ banded implies that $\text{rank}(F_k) = n-p$. Now F can be used to compute the projection matrix P. This permits the nonlinear parameters to be computed by minimizing the objective:

$$\Gamma\left(\boldsymbol{\gamma}\right) = \mathbf{b}^T F^T \left(FF^T\right)^{-1} F\mathbf{b} = \|P\mathbf{b}\|^2. \tag{3.5.7}$$

Note that the objective function Γ is independent of the scale of the parametrization so that it is necessary to adjoin a scaling constraint

$$\Psi\left(\boldsymbol{\gamma}\right) = 1 \tag{3.5.8}$$

to completely specify the optimization problem. Because $F^T\left(FF^T\right)^{-1}F$ is a projection matrix of rank $n-p$ it follows that

$$\mathcal{E}\left\{\Gamma\left(\boldsymbol{\gamma}^*\right)\right\} = \left(n-p\right)\sigma^2.$$

The term $F\left(\boldsymbol{\gamma}\right)\mathbf{b}$, being linear in the components of $\boldsymbol{\gamma}$, can be transformed to show the dependence on the parametrization explicitly. Let

$$F\left(\boldsymbol{\gamma}\right)\mathbf{b} = B\boldsymbol{\gamma}, \tag{3.5.9}$$

where $B : R^{p+1} \to R^{n-p}$ Then

$$B = \nabla_\gamma\left(F\left(\boldsymbol{\gamma}\right)\mathbf{b}\right).$$

In the exponential fitting case:

$$B_{ij} = b_{i+j-1},\ i = 1, 2, \cdots, n-p,\ j = 1, 2, \cdots, p+1. \tag{3.5.10}$$

In the rational fitting example

$$B_{ik} = \sum_{j=1}^{p+1} \Delta_j^p t_{i+j-1}^{k-1} b_{i+j-1},\ i = 1, 2, \cdots, n-p,\ k = 1, 2, \cdots, m. \tag{3.5.11}$$

To compute the gradient of the objective we have

$$\frac{\partial \Gamma}{\partial \gamma_i} = 2\mathbf{e}_i^T B^T \left(FF^T\right)^{-1} B\boldsymbol{\gamma} - \boldsymbol{\gamma}^T B^T \left(FF^T\right)^{-1} \frac{\partial \left(FF^T\right)}{\partial \gamma_i} \left(FF^T\right)^{-1} B\boldsymbol{\gamma}.$$

Setting

$$F = \sum_{i=1}^{m} \gamma_i C_i,\ \mathbf{s} = \left(FF^T\right)^{-1} B\boldsymbol{\gamma},$$

then

$$\begin{aligned} \frac{\partial \Gamma}{\partial \gamma_i} &= 2\mathbf{e}_i^T B^T \mathbf{s} - \mathbf{s}^T \left\{ C_i \left(\sum_{i=1}^{m} C_j \gamma_j \right)^T + \left(\sum_{i=1}^{m} C_j \gamma_j \right) C_i^T \right\} \mathbf{s}, \\ &= 2\left(\mathbf{e}_i^T B^T \mathbf{s} - \mathbf{w}_i^T \boldsymbol{\gamma}\right), \end{aligned}$$

where $\left(\mathbf{w}_i\right)_j = \mathbf{s}^T C_i^T C_j \mathbf{s} = \left(\mathbf{w}_j\right)_i$. It follows that the necessary conditions are

$$\left(B^T \left(FF^T\right)^{-1} B - W \right) \boldsymbol{\gamma} = \lambda \nabla \Psi^T, \tag{3.5.12}$$

where W is the symmetric matrix with $W_{i*} = \mathbf{w}_i^T$ and λ is the Lagrange multiplier associated with the scaling constraint. This looks like an eigenvalue problem, but it is nonlinear in $\boldsymbol{\gamma}$ through the term $\left(FF^T\right)^{-1}$. Because Γ is homogeneous of degree 0 as a function of $\boldsymbol{\gamma}$ it follows that

$$\nabla_\gamma \Gamma \boldsymbol{\gamma} = 0.$$

It follows that $\boldsymbol{\gamma}$ satisfying the necessary conditions is associated with the multiplier $\lambda = 0$ provided

$$\nabla_\gamma \Psi \boldsymbol{\gamma} \neq 0,\ \boldsymbol{\gamma} \neq 0,$$

for example, if $\Psi = \|\boldsymbol{\gamma}\|^2$.

There is a connection between the the above development and the classical method of Prony [91] . This seeks to minimize the sum of squares of the transformed residuals.

$$\min_{\gamma} \|F(\boldsymbol{\gamma})\,\mathbf{b}\|^2. \tag{3.5.13}$$

It turns out that this procedure is not even consistent [54]. The problem comes about because the random component in the sum of squares is $F(\boldsymbol{\gamma})\boldsymbol{\varepsilon}$ and there is nontrivial correlation between these elements.

$$\begin{aligned}\mathcal{V}\{F(\boldsymbol{\gamma})\} &= \mathcal{E}\left\{F(\boldsymbol{\gamma})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T F(\boldsymbol{\gamma})^T\right\}, \\ &= \sigma^2 F(\boldsymbol{\gamma}) F(\boldsymbol{\gamma})^T.\end{aligned}$$

In contrast, the correct maximum likelihood formulation is

$$\min_{\boldsymbol{\gamma}} \mathbf{b}^T F(\boldsymbol{\gamma})^T \mathcal{V}\{F(\boldsymbol{\gamma})\}^{-1} F(\boldsymbol{\gamma})\mathbf{b}. \tag{3.5.14}$$

3.6 Analysis of variance

In the analysis of a linear model with normal errors $\backsim N(0,\sigma^2 I)$ the variance σ^2 appears as a factor multiplying the normal equations and so does not interfere with the calculation of the linear parameters. If unknown it can be estimated separately once the linear parameters are known. The problem is more complicated when the covariance matrix depends on a second set of parameters $\boldsymbol{\beta} \in R^m$. It is convenient to write this dependence in distributional form:

$$\mathbf{r}_n \backsim N(A_n\boldsymbol{\alpha}, V_n(\boldsymbol{\beta})). \tag{3.6.1}$$

Structure of $V_n(\boldsymbol{\beta})$ becomes significant in problem analysis in as much as it reflects aspects of matters such as the set up of experiments or equipment performance. For example, the discussion of consistency has involved the assumption of data collected from observations on a sequence of independent events. Independence brings with it the implications that $V_n(\boldsymbol{\beta})$ is block diagonal with n indexing the number of events, and that block sizes are determined by the individual experiments and are bounded a priori. Thus $V_n(\boldsymbol{\beta})$ has a particular form of sparsity, and this has been used explicitly in the analysis based on the law of large numbers. In the application of the analysis of variance to experimental design , where quantifying variation in plot yields is of particular interest, a common assumption [47] is

$$V(\boldsymbol{\beta}) = \beta_1 I + \sum_{j=2}^{m} \beta_j V_j V_j^T.$$

The log likelihood based on (3.6.1) is, up to constant terms, given by

$$-2\mathcal{L} = \mathbf{r}^T V(\boldsymbol{\beta})^{-1}\mathbf{r} + \log\det(V(\boldsymbol{\beta})). \tag{3.6.2}$$

The necessary conditions are

$$\mathbf{r}^T V(\boldsymbol{\beta})^{-1} A = 0, \tag{3.6.3}$$

$$-\mathbf{r}^T V(\boldsymbol{\beta})^{-1} \frac{\partial V}{\partial \beta_i} V(\boldsymbol{\beta})^{-1} \mathbf{r} + \operatorname{trace} V(\boldsymbol{\beta})^{-1} \frac{\partial V}{\partial \beta_i} = 0,\ i = 1, 2, \cdots, m. \tag{3.6.4}$$

The first set of conditions is just the normal equations for $\boldsymbol{\alpha}$ for each fixed $\boldsymbol{\beta}$. This gives the linear parameters $\boldsymbol{\alpha}(\boldsymbol{\beta})$ as functions of the nonlinear parameters . This reduces the problem to one in $\boldsymbol{\beta}$ alone but at a minimum cost of repeatedly solving a generalised least squares problem. Strict separation between the linear and nonlinear parameters can be achieved by a transformation which permits us to work in the null space of A. However, in contrast to (3.5.3), the critical part of the transformation is independent of the parametrization. Let Q be orthogonal,

$$A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} U \\ 0 \end{bmatrix},\ T = \begin{bmatrix} Q_2^T \\ A^T V^{-1} \end{bmatrix}.$$

Then T is nonsingular provided A has full column rank p. If

$$\boldsymbol{\zeta} = \begin{bmatrix} \boldsymbol{\zeta}_1 \\ \boldsymbol{\zeta}_2 \end{bmatrix} = T\mathbf{b},$$

then

$$\mathcal{E}\{\boldsymbol{\zeta}\} = \begin{bmatrix} 0 \\ A^T V^{-1} A\boldsymbol{\alpha} \end{bmatrix},\ \mathcal{V}\{\boldsymbol{\zeta}\} = \begin{bmatrix} Q_2^T V Q_2 & 0 \\ 0 & A^T V^{-1} A \end{bmatrix}.$$

Thus $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2$ are uncorrelated. This implies independence for normally distributed random variables. The log likelihood is the sum of the corresponding log marginal likelihoods which are simply written down given the means and variances. The log likelihood is, up to constant terms,

$$\begin{aligned} -2\mathcal{L} = {} & \mathbf{b}^T Q_2 \left(Q_2^T V Q_2\right)^{-1} Q_2^T \mathbf{b} + \log\det\left(Q_2^T V Q_2\right) \\ & + \mathbf{r}^T V^{-1} A \left(A^T V^{-1} A\right)^{-1} A^T V^{-1} \mathbf{r} + \log\det\left(A^T V^{-1} A\right). \end{aligned} \tag{3.6.5}$$

As the part of T which affects the linear parameters depends on $\boldsymbol{\beta}$ it is worth noting that (3.6.5) gives values that are pointwise identical to (3.6.2) - again up to constant terms. This uses that $P_Q^V + P_A^V = I$ (Appendix 1 of Chapter 1), that the densities of $\mathbf{b}$ and $\boldsymbol{\zeta}$ have densities related by

$$f_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) = f_{\mathbf{b}}\left(T^{-1}\boldsymbol{\zeta}\right) \det(T)^{-1},$$

and makes use of the following Lemma.

Lemma 3.5

$$\det\left(Q_2^T V Q_2\right) = \det\left(V\right)\det\left(A^T V^{-1} A\right)\det\left(A^T A\right)^{-1}.$$

Proof. First note that

$$\begin{aligned}
\det\left(A^T V^{-1} A\right) &= \det\left(A^T Q Q^T V^{-1} Q Q^T A\right), \\
&= \det\left(\left[\begin{array}{cc} U^T & 0 \end{array}\right] Q^T V^{-1} Q \left[\begin{array}{c} U \\ 0 \end{array}\right]\right), \\
&= \det\left(U^T U\right)\det\left(\left(Q^T V^{-1} Q\right)_{11}\right), \\
&= \det\left(U^T U\right)\det\left(\left(Q^T V Q\right)_{11}^{-1}\right),
\end{aligned}$$

and that

$$\det\left(Q_2^T V Q_2\right) = \det\left(\left(Q^T V Q\right)_{22}\right).$$

Jacobi's Theorem on complementary minors [49] gives

$$\begin{aligned}
\det\left(\left(Q^T V Q\right)_{11}^{-1}\right) &= \det\left(\left(Q^T V Q\right)_{22}\right)\det\left(Q^T V Q\right)^{-1}, \\
&= \det\left(\left(Q^T V Q\right)_{22}\right)\det\left(V\right)^{-1}.
\end{aligned}$$

The desired result now follows. ■■

Identity between the likelihoods means that the linear parameters are determined by (3.6.3) and this is readily verified by differentiating (3.6.5) with respect to $\boldsymbol{\alpha}$. Separability has been achieved because given (3.6.3) then the term involving $\boldsymbol{\alpha}$ in (3.6.5) drops out. Differentiating with respect to $\boldsymbol{\beta}$ gives

$$\begin{aligned}
0 = &-\mathbf{b}^T Q_2 \left(Q_2^T V Q_2\right)^{-1} Q_2^T \frac{\partial V}{\partial \beta_i} Q_2 \left(Q_2^T V Q_2\right)^{-1} Q_2^T \mathbf{b} \\
&+ \text{trace}\left(\left(Q_2^T V Q_2\right)^{-1} Q_2^T \frac{\partial V}{\partial \beta_i} Q_2\right) \\
&- \text{trace}\left(\left(A^T V^{-1} A\right)^{-1} A^T \frac{\partial V}{\partial \beta_i} A\right),\ i = 1, 2, \cdots, m.
\end{aligned} \tag{3.6.6}$$

It is tempting to drop the third term in this equation and just work with the terms corresponding to the marginal log likelihood for $\boldsymbol{\zeta}_1$. The marginal log likelihood up to constant terms is given by

$$\mathcal{L}_M(\boldsymbol{\beta}) = \mathcal{L}\left(\boldsymbol{\alpha}(\boldsymbol{\beta}), \boldsymbol{\beta}\right) + \frac{1}{2}\log\det\left(A^T V^{-1} A\right), \tag{3.6.7}$$

and the corresponding necessary conditions are

$$\begin{aligned} 0 = &-\mathbf{b}^T Q_2 \left(Q_2^T V Q_2\right)^{-1} Q_2^T \frac{\partial V}{\partial \beta_i} Q_2 \left(Q_2^T V Q_2\right)^{-1} Q_2^T \mathbf{b} \\ &+ \operatorname{trace}\left(\left(Q_2^T V Q_2\right)^{-1} Q_2^T \frac{\partial V}{\partial \beta_i} Q_2\right),\ i = 1, 2, \cdots, m. \end{aligned} \tag{3.6.8}$$

This approach has been adopted for general problems in experimental design , for example, in [88] where the acronym REML is introduced.

Example 3.6.1 *Let $V = \sigma^2 I$ then*

$$\begin{aligned} \mathbf{b}^T Q_2 \left(Q_2^T V Q_2\right)^{-1} Q_2^T \frac{\partial V}{\partial \sigma^2} Q_2 \left(Q_2^T V Q_2\right)^{-1} Q_2^T \mathbf{b} &= \frac{1}{\sigma^4} \boldsymbol{\varepsilon}^T Q_2 Q_2^T \varepsilon, \\ \operatorname{trace}\left(\left(Q_2^T V Q_2\right)^{-1} Q_2^T \frac{\partial V}{\partial \sigma^2} Q_2\right) &= \frac{n-p}{\sigma^2}. \end{aligned}$$

Thus the REML estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n-p} \mathbf{b}^T Q_2 Q_2^T \mathbf{b}.$$

This contrasts with the full maximum likelihood estimator in being unbiassed:

$$\begin{aligned} \mathcal{E}\left\{\boldsymbol{\varepsilon}^T Q_2 Q_2^T \boldsymbol{\varepsilon}\right\} &= \operatorname{trace}\left(Q_2^T \mathcal{E}\left\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right\} Q_2\right), \\ &= (n-p)\,\sigma^2. \end{aligned}$$

Remark 3.6.1 *For the REML estimate for $\boldsymbol{\beta}^*$ to be consistent when this is true of the maximum likelihood estimate requires that the term*

$$\operatorname{trace}\left(\left(A^T V^{-1} A\right)^{-1} A^T \frac{\partial V}{\partial \beta_i} A\right)$$

be negligible compared to the other terms. Typically this is an $O(p)$ term if $\frac{1}{n}\left(A^T V^{-1} A\right)$ and $\frac{1}{n}\left(A^T \frac{\partial V}{\partial \beta_i} A\right)$ are bounded for large n, conditions which fit well with our standard assumptions. The REML terms can be written as matrix traces each involving $n-p$ terms. These will dominate if the trace arguments involve $O(1)$ terms as was the case in the above example.

To calculate the expected Hessian of the marginal log likelihood $\mathcal{L}_M$ at $\boldsymbol{\beta}^*$ note that $\mathcal{E}\left\{Q_2^T \mathbf{b}\mathbf{b}^T Q_2\right\} = Q_2^T V\left(\boldsymbol{\beta}^*\right) Q_2$ and set

$$\mathbf{z} = \mathcal{E}\left\{\mathbf{b}^T Q_2 \left(Q_2^T V Q_2\right)^{-1} Q_2^T \frac{\partial V}{\partial \beta_i} Q_2 \left(Q_2^T V Q_2\right)^{-1} Q_2^T \mathbf{b}\right\}.$$

Then

$$\begin{aligned}\mathbf{z} &= \text{trace}\left(\left(Q_2^TVQ_2\right)^{-1}Q_2^T\frac{\partial V}{\partial\beta_i}Q_2\left(Q_2^TVQ_2\right)^{-1}\mathcal{E}\left\{Q_2^T\mathbf{b}\mathbf{b}^TQ_2\right\}\right)\\ &= \text{trace}\left(\left(Q_2^TVQ_2\right)^{-1}Q_2^T\frac{\partial V}{\partial\beta_i}Q_2\right).\end{aligned}$$

This shows that the REML necessary conditions involve equating statistics to their expectations.This provides a sense in which the resulting estimating equation can be said to be unbiased. The calculation of the expected Hessian proceeds as follows:

$$\begin{aligned}-2\mathcal{E}\left\{\frac{\partial^2\mathcal{L}_M}{\partial\beta_i\partial\beta_j}\right\} &= \text{trace}\left(\left(Q_2^TVQ_2\right)^{-1}Q_2^T\frac{\partial V}{\partial\beta_i}Q_2\left(Q_2^TVQ_2\right)^{-1}Q_2^T\frac{\partial V}{\partial\beta_j}Q_2\right),\\ &= \text{trace}\left(V^{-1}P_Q^V\frac{\partial V}{\partial\beta_i}V^{-1}P_Q^V\frac{\partial V}{\partial\beta_j}\right),\end{aligned}\tag{3.6.9}$$

where $P_Q^V = VQ_2\left(Q_2^TVQ_2\right)^{-1}Q_2^T$ is defined in the Appendix to Chapter 1.

3.7 Appendix: A form of the law of large numbers

The classic Khintchtine form of the law of large numbers shows that $\frac{1}{n}\sum_{i=1}^n\varepsilon_i \overset{\text{a.s.}}{\to} 0,\ n\to\infty$ where the ϵ_i are mean zero, iid random variables. An extension of this result [52] shows that $\frac{1}{n}\sum_{i=1}^n X_i\varepsilon_i \overset{\text{a.s.}}{\to} 0,\ n\to\infty$ provided the X_i are sufficiently slowly varying. This proves to be in the right form for our considerations in the case of (nonlinear) least squares problems associated with normal likelihoods. It is of interest for introducing a regularity condition which is a form of the designed experiment condition.

Theorem 3.9 *Let $\{f_i(\boldsymbol{\gamma}\}$ be a sequence of continuous functions defined for $\boldsymbol{\gamma}\in\Gamma$ having the property that*

$$\frac{1}{n}\sum_{i=1}^n f_i(\boldsymbol{\gamma}_1)f_i(\boldsymbol{\gamma}_2)\tag{3.7.1}$$

converges uniformly for $\boldsymbol{\gamma}_1,\ \boldsymbol{\gamma}_2\in\Gamma,\ n\to\infty$. Then

$$\frac{1}{n}\sum_{i=1}^n f_i(\boldsymbol{\gamma})\varepsilon_i \overset{a.s.}{\to} 0,\ n\to\infty\tag{3.7.2}$$

for $\boldsymbol{\gamma}\in\Gamma$, where the ε_i are mean 0, iid random variables.

In our application, typically $f_i(\boldsymbol{\gamma}) = f(t_i, \boldsymbol{\gamma}),\ t_i \in \mathbf{T_n}$. In this case (3.7.1) corresponds to a case of the designed experiment condition (3.2.1).

The noise term need not enter linearly in $\nabla_{\mathbf{x}} \mathcal{L}\left((\mathbf{y}; \mathbf{x}, \mathbf{T})\right)$ in more general likelihoods, and in this case a sufficient condition for the validity of the Kolmogorov form [98] can be applied. Let the ε_i be iid random variables, the $X_i = X(t_i, \boldsymbol{\gamma}, \varepsilon_i)$ be continuous functions of their arguments, and let $\mathcal{V}\{X_i\},\ i = 1, 2, \cdots, n$ be uniformly bounded. Then

$$\frac{1}{n} \sum_{i=1}^{n} \{X_i - \mathcal{E}\{X_i\}\} \stackrel{\text{a.s.}}{\to} 0. \tag{3.7.3}$$

Chapter 4

Computational methods for maximum likelihood

4.1 Introduction

This chapter considers in some detail implementation and properties of the Fisher scoring algorithm for maximizing a likelihood. Scoring is widely used in statistical calculations, and there is a general belief in its effectiveness. This context draws attention to the importance of a data analytic setting in analyzing the algorithm, and this is the one that is adopted here. In contrast, not only the orthodox computational literature (for example [24]), but also texts such as [103], which set out computational developments for statisticians, restrict performance considerations to discussion of the Gauss-Newton algorithm in the special case of small residuals. It is stressed here that this analysis can be broadened dramatically by considering the properties of the large sample asymptotics in the correct setting.

The scoring algorithm is no more than a particular variant of Newton's method applied to solve the maximization problem. Recall that the basic step of the Newton algorithm has the form

$$\begin{aligned} &\mathbf{h} = -\mathcal{J}_n(\mathbf{x})^{-1} \frac{1}{n} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{y}; \mathbf{x}, \mathbf{T}_n)^T, \\ &\mathbf{x} \leftarrow \mathbf{x} + \mathbf{h}. \end{aligned} \tag{4.1.1}$$

The modified algorithm has two variants:

1. The Hessian of the likelihood $\mathcal{J}_n(\mathbf{x})$ is replaced by its expectation

$-\mathcal{J}_n(\mathbf{x}^*)$	$\overset{a.s.}{\underset{n\to\infty}{\to}}$	$\mathcal{I}(\mathbf{x}^*)$	
		$\approx$	$\|\mathbf{x}^* - \mathbf{x}\|$ small
$\mathcal{I}_n(\mathbf{x})$	$\overset{r.e.}{\underset{n\to\infty}{\to}}$	$\mathcal{I}(\mathbf{x})$	

Table 4.1: Scoring diagram

$\mathcal{E}\{\mathcal{J}_n(\mathbf{x})\} = -\mathcal{I}_n(\mathbf{x})$ to give the iteration

$$\begin{aligned} \mathbf{h}_I &= \mathcal{I}_n(\mathbf{x})^{-1} \frac{1}{n} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{y}; \mathbf{x}, \mathbf{T}_n)^T, \\ \mathbf{x} &\leftarrow \mathbf{x} + \mathbf{h}. \end{aligned} \tag{4.1.2}$$

This form is the one corresponding to the standard form of the Fisher scoring algorithm. The particular case of nonlinear least squares corresponds to the use of the normal likelihood and gives the Gauss-Newton method. An immediate consequence of (4.1.2) is that

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{y}; \mathbf{x}, \mathbf{T}_n)\, \mathbf{h}_I > 0,\ \mathbf{h}_I \neq 0,$$

so that $\mathbf{h}_I$ is a direction of ascent for maximizing the likelihood.

2. The standard form is fine provided the evaluation of the expectation is straight forward. Otherwise a possible replacement is provided by the sample information

$$\mathcal{S}_n(\mathbf{x}) = \frac{1}{n} \sum_{\mathbf{t} \in \mathbf{T}_n} \nabla_{\mathbf{x}} L_t(\mathbf{y}_t; \boldsymbol{\theta}_t, \mathbf{t})^T \nabla_{\mathbf{x}} L_t(\mathbf{y}_t; \boldsymbol{\theta}_t, \mathbf{t}). \tag{4.1.3}$$

The resulting iteration is

$$\begin{aligned} \mathbf{h}_S &= \mathcal{S}_n(\mathbf{x})^{-1} \frac{1}{n} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{y}; \mathbf{x}, \mathbf{T}_n)^T, \\ \mathbf{x} &\leftarrow \mathbf{x} + \mathbf{h}. \end{aligned} \tag{4.1.4}$$

Remark 4.1.1 *The steps from Newton's method to the scoring algorithm are summarised in Table 4.1. Note the important roles played by the almost sure convergence of the Hessian to a matrix that does not depend on second derivatives of the log likelihood, and by the convergence associated with a regular sequence of experiments.*

Remark 4.1.2 *In the case of a normal likelihood, independent observations, and a signal in noise model (here $\mu_i(\mathbf{x})$ is written for $\mu(t_i, \mathbf{x})$) then the dependence on the standard deviation σ^2 can be suppressed and the log likelihood written:*

$$\nabla_x \mathcal{L} = \sum_i \left(y_i - \mu_i\left(\mathbf{x}\right)\right) \nabla_x \mu_i\left(\mathbf{x}\right)$$
$$\nabla_x^2 \mathcal{L} = \sum_i \left\{-\nabla_x \mu_i^T \nabla_x \mu_i + \left(y_i - \mu_i\left(\mathbf{x}\right)\right) \nabla_x^2 \mu_i\right\}.,$$

Here

$$\mathcal{E}\left\{\nabla_x^2 \mathcal{L}\right\} = \sum_i -\nabla_x \mu_i^T \nabla_x \mu_i.$$

and does not involve random components.

The first result is that the three iterations are asymptotically equivalent under regular sampling regimes. This requires showing that the three different Hessian estimates are asymptotically equivalent in the sense of almost sure convergence for large enough n for then the values of $\mathbf{h}$ generated by the different methods will agree to first order almost surely.

Lemma 4.1 *Assume a regular sampling regime. Then, in the sense of almost sure convergence,*

$$\lim_{n\to\infty} \mathcal{I}_n\left(\mathbf{x}^*\right) = \lim_{n\to\infty} \mathcal{S}_n\left(\mathbf{x}^*\right) = -\lim_{n\to\infty} \mathcal{J}_n\left(\mathbf{x}^*\right) = \mathcal{I}.$$

Proof. The result

$$\lim_{n\to\infty} \mathcal{J}_n\left(\mathbf{x}^*\right) = -\mathcal{I}.$$

is equation (3.2.8). We have

$$\begin{aligned}
\mathcal{I}_n\left(\mathbf{x}^*\right) + \mathcal{J}_n\left(\mathbf{x}^*\right) &= \frac{1}{n} \sum_{\mathbf{t}\in\mathbf{T}_n} \left\{ \begin{array}{c} \nabla_{\mathbf{x}}^2 L_t\left(\mathbf{y}_t; \boldsymbol{\theta}_t^*, \mathbf{t}\right) \\ +\mathcal{E}\left\{\nabla_{\mathbf{x}} L_t\left(\mathbf{y}_t; \boldsymbol{\theta}_t^*, \mathbf{t}\right)^T \nabla_{\mathbf{x}} L_t\left(\mathbf{y}_t; \boldsymbol{\theta}_t^*, \mathbf{t}\right)\right\} \end{array} \right\}, \\
&= \frac{1}{n} \sum_{\mathbf{t}\in\mathbf{T}_n} \left\{\nabla_{\mathbf{x}}^2 L_t\left(\mathbf{y}_t; \boldsymbol{\theta}_t^*, \mathbf{t}\right) - \mathcal{E}\left\{\nabla_{\mathbf{x}}^2 L_t\left(\mathbf{y}_t; \boldsymbol{\theta}_t^*, \mathbf{t}\right)\right\}\right\}, \\
&\to 0,\ n \to \infty,
\end{aligned}$$

by Lemma 3.2, and Jennrich's form of the strong law of large numbers. In similar fashion

$$\begin{aligned}
\mathcal{S}_n\left(\mathbf{x}^*\right) - \mathcal{I}_n\left(\mathbf{x}^*\right) &= \frac{1}{n} \sum_{\mathbf{t}\in\mathbf{T}_n} \left\{ \begin{array}{c} \nabla_{\mathbf{x}} L_t\left(\mathbf{y}_t; \boldsymbol{\theta}_t^*, \mathbf{t}\right)^T \nabla_{\mathbf{x}} L_t\left(\mathbf{y}_t; \boldsymbol{\theta}_t^*, \mathbf{t}\right) \\ -\mathcal{E}\left\{\nabla_{\mathbf{x}} L_t\left(\mathbf{y}_t; \boldsymbol{\theta}_t^*, \mathbf{t}\right)^T \nabla_{\mathbf{x}} L_t\left(\mathbf{y}_t; \boldsymbol{\theta}_t^*, \mathbf{t}\right)\right\} \end{array} \right\}, \\
&\to 0,\ n \to \infty,
\end{aligned}$$

by the strong law of large numbers. ■

There is a similar development for the application of Newton's method to find a zero of the quasilikelihood estimating equation (3.3.6). The basic step is

$$\begin{aligned}\mathbf{h} &= -\left\{\frac{1}{n}\nabla_{\mathbf{x}}^2\mathcal{K}_T\left(\mathbf{y};\mathbf{x},\mathbf{T}_n\right)\right\}^{-1}\left\{\frac{1}{n}\nabla_{\mathbf{x}}\mathcal{K}_T\left(\mathbf{y};\mathbf{x},\mathbf{T}_n\right)^T\right\}, \\ \mathbf{x} &\leftarrow \mathbf{x}+\mathbf{h},\end{aligned} \tag{4.1.5}$$

where

$$\nabla_{\mathbf{x}}\mathcal{K}_T\left(\mathbf{y};\mathbf{x},\mathbf{T}_n\right)^T = \sum_{\mathbf{t}}\left(\nabla_{\mathbf{x}}\boldsymbol{\mu}_t\right)^T V\left(\boldsymbol{\mu}_t\right)^{-1}\left(\mathbf{y}_t-\boldsymbol{\mu}_t\left(\mathbf{x}\right)\right)$$

The scoring algorithm replaces the Hessian term $-\frac{1}{n}\nabla_{\mathbf{x}}^2\mathcal{K}_T$ by its expectation $\mathcal{I}_K^n$ (compare (3.3.8))

$$\mathcal{I}_K^n = \frac{1}{n}\sum_{\mathbf{t}}\left(\nabla_{\mathbf{x}}\boldsymbol{\mu}_t\right)^T V\left(\boldsymbol{\mu}_t\right)^{-1}\nabla_{\mathbf{x}}\boldsymbol{\mu}_t, \tag{4.1.6}$$

to give the basic iteration step

$$\begin{aligned}\mathbf{h}_Q &= \left\{\mathcal{I}_K^n\right\}^{-1}\frac{1}{n}\nabla_{\mathbf{x}}\mathcal{K}_T\left(\mathbf{y};\mathbf{x},\mathbf{T}_n\right)^T, \\ \mathbf{x} &\leftarrow \mathbf{x}+\mathbf{h}.\end{aligned} \tag{4.1.7}$$

The key feature in the case of each of these variants is that the negative of the modified Hessian matrix is generically positive definite, a property not shared by the true Hessian at a general point. This has the important consequence that the modified methods have good global convergence properties when line search strategies are employed to stabilize the basic iteration. When this is combined with similar transformation invariance properties, with asymptotically fast rates of convergence as a consequence of Lemma 4.1, and with possibly lower computational cost, then the scoring algorithms do look attractive.

Remark 4.1.3 *The transformation invariance properties are readily demonstrated. We consider the standard scoring step but the same argument applies to all the modified algorithms. Let* $\mathbf{u}=\mathbf{u}\left(\mathbf{x}\right)$. *Then*

$$\nabla_{\mathbf{x}}\mathcal{L} = \nabla_{\mathbf{u}}\mathcal{L}T,\ \mathcal{I}_n^{\mathbf{x}} = T^T\mathcal{I}_n^{\mathbf{u}}T,\ T\left(\mathbf{x}\right) = \frac{\partial\mathbf{u}}{\partial\mathbf{x}}.$$

The Fisher scoring steps in the two variables are related by

$$\mathbf{h}_{\mathbf{x}} = \left(T^T\mathcal{I}_n^{\mathbf{u}}T\right)^{-1}\frac{1}{n}T^T\nabla_{\mathbf{u}}\mathcal{L}^T = T^{-1}\left(\mathcal{I}_n^{\mathbf{u}}\right)^{-1}\frac{1}{n}\nabla_{\mathbf{u}}\mathcal{L}^T \tag{4.1.8}$$

so that

$$\mathbf{u} = \mathbf{u}(\mathbf{x}) \Rightarrow \mathbf{h_u} = T\mathbf{h_x}.$$

No derivative of T appears in (4.1.8) so the transformation invariance, in this sense, applies at a general point. This contrasts with Newton's method for function minimization applied to $F(\mathbf{u}(\mathbf{x}))$ where strict transformation invariance applies only in a neighbourhood of a stationary point, while at a general point the invariance is restricted to constant T. The problem occurs because the second differentiation needed to compute the Hessian introduces a term involving derivatives of the transformation matrix multiplying the condition for a stationary point.

4.2 Basic properties of the ascent method

The basic structure of an ascent or maximization algorithm involves two key ingredients:

1. A method for generating a step $\mathbf{h}$ which defines a direction in which the objective function is increasing; and

2. A method for measuring progress in this direction. This second requirement recognizes that a full step (as in the Newton basic iteration (4.1.1)) need not be satisfactory, and that a more detailed local examination can be needed to make progress - this is especially true in the initial stages of the ascent computation. To do this a monitor function $\Phi(\mathbf{x})$ is introduced as a basis for this measurement. This needs to have both the same (local) maximum as $F(\mathbf{x})$, the function to be maximized, and to increase when $F(\mathbf{x})$ is increasing. This requirement is summarised by

$$\nabla F\mathbf{h} \geq 0 \Rightarrow \nabla\Phi\mathbf{h} \geq 0;\ \nabla F\mathbf{h} = 0 \Rightarrow \nabla\Phi\mathbf{h} = 0.$$

   A second desirable property of a monitor function is transformation invariance.

Two classes of approach based on these ideas have proved popular and are considered here. In the first, a single efficient direction $\mathbf{h}$ is computed, and the actual step $\lambda\mathbf{h}$ in this direction is controlled by searching using the monitor function to gauge an effective step. Such methods are useful when a fast rate of ultimate convergence is possible combined with a natural scale determining the length of step. In the second approach the tentative ascent step is required to lie in a trust region surrounding the current point. This

trust region is then modified adaptively as a result of this computation with the step being recomputed if necessary. Typically the ascent step is computed by linearizing the problem, and the trust region serves to define a region in which this linearization satisfies an acceptability condition computed using the monitor.

4.2.1 Ascent methods using line searches

It is a classic result [30] that to prove convergence to a stationary point it is not sufficient to make nominal progress in the direction of ascent. It is required that the step $\lambda\mathbf{h}$ taken from the current point be associated with a sufficiently large λ in the set of allowable values.Typical strategies for choosing λ include:

1. (Goldstein) Let

$$\Psi(\lambda, \mathbf{x}, \mathbf{h}) = \frac{\Phi(\mathbf{x} + \lambda\mathbf{h}) - \Phi(\mathbf{x})}{\lambda \nabla_{\mathbf{x}} \Phi(\mathbf{x})\, \mathbf{h}},\ 0 < \varrho < .5, \tag{4.2.1}$$

   and choose λ to satisfy the inequalities

$$\varrho \leq \Psi(\lambda, \mathbf{x}, \mathbf{h}) \leq 1 - \varrho. \tag{4.2.2}$$

   This can always be done provided Φ decreases for λ large enough causing a violation of the first (left hand) inequality, because the second inequality prevents too small values of λ occuring as

$$\lim_{\lambda \to 0} \Psi(\lambda, \mathbf{x}, \mathbf{h}) = 1$$

   provided $\mathbf{x}$ is not a stationary point. Typically, ϱ is chosen small (say 10^{-4}). One aim here is prevent the test missing a suitable stopping value by flipping from $\Psi < \varrho$ to $1 - \varrho < \Psi$ in one correction to λ.

2. (Simple) Let $0 < \tau < 1$, and choose $\lambda = \tau^k$ where k is the smallest integer such that $k = 1$ if $\Phi(\mathbf{x} + \mathbf{h}) > \Phi(\mathbf{x})$ else k satisfies

$$\Phi\left(\mathbf{x} + \tau^{k-1}\mathbf{h}\right) \leq \Phi(\mathbf{x}) < \Phi\left(\mathbf{x} + \tau^{k}\mathbf{h}\right). \tag{4.2.3}$$

   The value of τ does not seem critical. Values satisfying $.1 \leq \tau \leq .5$ seem satisfactory.

Remark 4.2.1 *Direct convergence results will consider the Goldstein test. There are even counter examples where premature convergence is flagged under conditions similar to the Simple test [73]. This would require* $\inf\{\lambda\} = 0$. *However, if this behaviour occurs in the scoring algorithm then it can be shown that the Hessian of the objective function is unbounded (Theorem 4.2).*

Remark 4.2.2 *There are some useful connections between the two search strategies. Typically, if* $\Phi(\mathbf{x}+\mathbf{h}) > \Phi(\mathbf{x})$ *then* $\lambda = 1$ *is accepted corresponding to the Simple test being satisfied with* $k = 0$*. The argument uses that this step can yield a fast convergence rate for the transformation invariant scoring algorithms under appropriate conditions so that it makes sense to use it to set the search scale. Equation (4.2.2) expresses a somewhat more stringent condition. However, if successive iterates are contained in a bounded region* R *in which* $\widehat{k}$ *is an upper bound to the values of* k *computed in the Simple steps then a three term Taylor series based analysis shows that* $\varrho = \tau^{\widehat{k}}$ *satisfies the left hand inequality in (4.2.2) for each step provided the Hessian of* $\mathcal{L}$ *is positive definite. Also, if* $\lambda = \tau^k$ *is accepted for a Simple step then in (4.2.2)*

$$\Psi\left(\tau^{k-1}, \mathbf{x}, \mathbf{h}\right) < 0 < \varrho. \tag{4.2.4}$$

A method for computing λ *to satisfy (4.2.2) is given in Subsubsection 4.2.1. This can be interpreted as computing a series of Simple multipliers* $\tau_j < .5$ *with* $\lambda = \prod_{j=1}^{k} \tau_j$*.*

Lemma 4.2 *The scoring algorithms (4.1.2), (4.1.4) generate directions of ascent provided* $\mathcal{I}_n$*,* $\mathcal{S}_n$ *are positive definite (they are necessarily at least positive semidefinite). In both cases* $\Phi(\mathbf{x}) = \mathcal{L}(\mathbf{y}; \mathbf{x}, \mathbf{T}_n)$ *provides a suitable monitor.*

Proof. The arguments are identical for both the variants of scoring so only the standard one is considered here. If $\mathbf{x}$ is not a stationary point then $\nabla_{\mathbf{x}}\mathcal{L} \neq 0$ and positive definiteness gives

$$\nabla_{\mathbf{x}}\mathcal{L}\mathbf{h} = \frac{1}{n}\nabla_{\mathbf{x}}\mathcal{L}\mathcal{I}_n^{-1}\nabla_{\mathbf{x}}\mathcal{L}^T > 0.$$

To show transformation invariance in the step determining λ only the first criterion needs to be considered, and here it is necessary to show that $\nabla_{\mathbf{x}}\mathcal{L}\mathbf{h}_x$ is transformation invariant. We have

$$\begin{aligned} n\nabla_{\mathbf{x}}\mathcal{L}\mathbf{h}_x &= \nabla_{\mathbf{x}}\mathcal{L}\mathcal{I}_n^{-1}\nabla_{\mathbf{x}}\mathcal{L}^T, \\ &= \nabla_{\mathbf{u}}\mathcal{L}T\left(T^T\mathcal{I}_nT\right)^{-1}T^T\nabla_{\mathbf{u}}\mathcal{L}^T, \\ &= \nabla_{\mathbf{u}}\mathcal{L}\left(\mathcal{I}_n^u\right)^{-1}\nabla_{\mathbf{u}}\mathcal{L}^T, \\ &= n\nabla_{\mathbf{u}}\mathcal{L}\mathbf{h}_u. \end{aligned} \tag{4.2.5}$$

■

The following Lemma is important for establishing convergence of the scoring iteration. Again there are essentially similar results for both forms of scoring.

Lemma 4.3 *For the basic scoring algorithm*

$$\frac{\nabla_{\mathbf{x}}\mathcal{L}\mathbf{h}}{\|\nabla_{\mathbf{x}}\mathcal{L}\|\,\|\mathbf{h}\|} \geq \frac{1}{\mathrm{cond}\,(\mathcal{I}_n)^{1/2}} \tag{4.2.6}$$

where the condition number is the spectral condition number.

Proof. We have

$$\frac{\nabla_{\mathbf{x}}\mathcal{L}\mathbf{h}}{\|\nabla_{\mathbf{x}}\mathcal{L}\|\,\|\mathbf{h}\|} = \frac{\mathbf{h}^T\mathcal{I}_n\mathbf{h}}{\|\mathcal{I}_n\mathbf{h}\|\,\|\mathbf{h}\|}.$$

Let $\mathbf{v} = \mathcal{I}_n^{1/2}\mathbf{h}$. Then

$$\begin{aligned}\frac{\nabla_{\mathbf{x}}\mathcal{L}\mathbf{h}}{\|\nabla_{\mathbf{x}}\mathcal{L}\|\,\|\mathbf{h}\|} &= \frac{\mathbf{v}^T\mathbf{v}}{\{\mathbf{v}^T\mathcal{I}_n\mathbf{v}\}^{1/2}\{\mathbf{v}^T\mathcal{I}_n^{-1}\mathbf{v}\}^{1/2}},\\ &\geq \frac{2\,\mathrm{cond}\,(\mathcal{I}_n)^{1/2}}{1+\mathrm{cond}\,(\mathcal{I}_n)},\\ &\geq \frac{2}{\mathrm{cond}\,(\mathcal{I}_n)^{1/2}+\mathrm{cond}\,(\mathcal{I}_n)^{-1/2}}.\end{aligned}$$

The key inequality used is the Kantorovich inequality [49], p.83 .

■

Definition 4.1 *If $\mathcal{I}_n$ (similarly $\mathcal{S}_n$) is positive definite throughout a compact region R on which $\mathcal{L}$ is bounded then the scoring algorithm gives an ascent direction at every point which is not a stationary point of $\mathcal{L}$, and the scaled inequality (4.2.6) holds. In this case we say that a uniform ascent condition holds in R.*

Lemma 4.4 *If a uniform ascent condition holds in a compact region R on which $\mathcal{L}$ is bounded then, if $\lambda(\mathbf{x})$, $\mathbf{x}$ not a stationary point, satisfies the inequalities (4.2.2), there is a uniform lower bound λ_R such that $\lambda(\mathbf{x}) \geq \lambda_R > 0$, $\forall \mathbf{x} \in R$.*

Proof. This follows by observing that the two propositions $\lambda(\mathbf{x}_i) \to 0$ and $\Psi(\lambda(\mathbf{x}_i), \mathbf{x}_i, \mathbf{h}(\mathbf{x}_i)) < 1 - \varrho$ are inconsistent. ■

Theorem 4.1 *Let the sequence of iterates $\{\mathbf{x}_i\}$ produced by the ascent algorithm using the line search criterion (4.2.2) be contained in a compact region R in which a uniform ascent condition holds. Then the sequence of values $\{\mathcal{L}(\mathbf{y}; \mathbf{x}_i, \mathbf{T}_n)\}$ converges, and limit points of $\{\mathbf{x}_i\}$ are stationary points of $\mathcal{L}$.*

Proof. Convergence of the bounded, increasing sequence $\{\mathcal{L}(\mathbf{y};\mathbf{x}_i,\mathbf{T}_n)\}$ is a consequence of the uniform ascent condition. It follows from (4.2.1) that

$$\begin{aligned}\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{y};\mathbf{x}_i,\mathbf{T}_n)\,\mathbf{h}_i &\leq \frac{\mathcal{L}(\mathbf{y};\mathbf{x}_i+\lambda_i\mathbf{h}_i,\mathbf{T}_n)-\mathcal{L}(\mathbf{y};\mathbf{x}_i,\mathbf{T}_n)}{\varrho\lambda_R},\\ &\to 0,\ i\to\infty,\end{aligned}$$

as the numerator on the right hand side is the difference between consecutive terms in a convergenct sequence, and $0<\lambda_R\leq\lambda_i$ by the uniform ascent condition. Combining this with (4.2.6) gives

$$\frac{\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{y};\mathbf{x}_i,\mathbf{T}_n)\|\,\|\mathbf{h}_i\|}{\kappa_R}\leq\frac{\mathcal{L}(\mathbf{y};\mathbf{x}_i+\lambda_i\mathbf{h}_i,\mathbf{T}_n)-\mathcal{L}(\mathbf{y};\mathbf{x}_i,\mathbf{T}_n)}{\varrho\lambda_R},$$

where $\kappa_R=\mathrm{cond}\,(\mathcal{I}_n)^{1/2}$. But

$$\|\mathbf{h}\|=\left\|\mathcal{I}_n^{-1}\frac{1}{n}\nabla_{\mathbf{x}}\mathcal{L}^T\right\|\geq\frac{\|\nabla_{\mathbf{x}}\mathcal{L}\|}{n\,\|\mathcal{I}_n\|}.$$

Thus, if μ_R is an upper bound in R for $\|\mathcal{I}_n\|$,

$$\frac{\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{y};\mathbf{x}_i,\mathbf{T}_n)\|^2}{n\kappa_R\mu_R}\leq\frac{\mathcal{L}(\mathbf{y};\mathbf{x}_i+\lambda_i\mathbf{h}_i,\mathbf{T}_n)-\mathcal{L}(\mathbf{y};\mathbf{x}_i,\mathbf{T}_n)}{\varrho\lambda_R}.$$

It follows that $\{\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{y};\mathbf{x}_i,\mathbf{T}_n)\|\}\to 0$.

■

Basically this theorem says that provided $\inf_i\{\lambda_i\}>0$ and starting conditions have been selected appropriately then scoring provides an effective procedure for maximizing the likelihood. The next result addresses what happens if the computed step lengths approach zero so that $\inf_i\{\lambda_i\}=0$. The argument applies to both the Simple and Goldstein tests, but the Simple test is strictly the only one which allows $\inf_i\{\lambda_i\}=0$ as the right hand inequality in the Goldstein test could approach this possibility in a bounded region only if the Hessian is becoming unbounded and would flag this by terminating.

Theorem 4.2 *Let the sequence of iterates $\{\mathbf{x}_i\}$ produced by the scoring algorithm implemented using the Simple step criterion be contained in a compact region R in which $\mathcal{I}_n$ has full rank and $\inf_i\{\lambda_i\}=0$. Then $\frac{1}{n}\nabla_x^2\mathcal{L}(\mathbf{x})$ is unbounded in R.*

Proof. For $\inf_i\{\lambda_i\}=0$ to obtain there must be an infinite sequence of points $\widehat{\mathbf{x}}_i=\mathbf{x}_i+\widehat{\lambda}_i\mathbf{h}_i$ where the Simple test fails so that, by (4.2.4),

$$\varrho>\frac{\mathcal{L}(\widehat{\mathbf{x}}_i)-\mathcal{L}(\mathbf{x}_i)}{\widehat{\lambda}_i\nabla_x\mathcal{L}(\mathbf{x}_i)\mathbf{h}_i},\quad\inf_i\{\widehat{\lambda}_i\}=0.$$

Using the mean value theorem and the property that $\mathbf{h}_i$ is a direction of ascent gives

$$\varrho > \frac{\widehat{\lambda}_i \nabla_x \mathcal{L}(\mathbf{x}_i)\mathbf{h}_i - \frac{\widehat{\lambda}_i^2}{2}|\mathbf{h}_i^T \nabla_x^2 \mathcal{L}(\overline{\mathbf{x}})\mathbf{h}_i|}{\widehat{\lambda}_i \nabla_x \mathcal{L}(\mathbf{x}_i)\mathbf{h}_i},$$

where the bar denotes that a mean value is appropriate. Thus

$$\widehat{\lambda}_i |\mathbf{h}_i^T \frac{1}{n} \nabla_x^2 \mathcal{L}(\overline{\mathbf{x}})\mathbf{h}_i| > 2(1-\varrho)\frac{1}{n}\nabla_x \mathcal{L}(\mathbf{x}_i)\mathbf{h}_i,$$

so that

$$\begin{aligned} \|\frac{1}{n}\nabla_x^2 \mathcal{L}(\overline{\mathbf{x}})\| &> \frac{2(1-\varrho)}{\widehat{\lambda}_i} \frac{\frac{1}{n}\nabla_x \mathcal{L}(\mathbf{x}_i)\mathbf{h}_i}{\|\mathbf{h}_i\|^2} \\ &= \frac{2(1-\varrho)}{\widehat{\lambda}_i} \frac{\mathbf{h}_i^T \mathcal{I}_n \mathbf{h}_i}{\|\mathbf{h}_i\|^2} \\ &> \frac{2(1-\varrho)}{\widehat{\lambda}_i}\, \sigma_{min}(\mathcal{I}_n), \end{aligned} \tag{4.2.7}$$

where σ_{min} is the smallest singular value of $\mathcal{I}_n$. The result now follows from the definition of $\widehat{\lambda}_i$.

■

Remark 4.2.3 *This theorem provides something like a global convergence result for the scoring algorithm . In particular, if $\mathcal{L}$ has bounded second derivatives in any finite region of R^p then unbounded second derivatives can occur only at ∞ so that the case $\inf_i\{\lambda_i\} = 0$ must be associated with an unbounded sequence $\{\mathbf{x}_i\}$ in this case. It is necessary to allow for this behaviour for consider approximation of t by nonlinear combinations of the form $x(1) + x(2)\exp^{-x(3)t}$. Here*

$$t = \lim_{n\to\infty}\{n - n\exp -t/n\}$$

with an error for large n which is $O(1/n)$. Thus large n solutions for exact data must be close to $x(1) = n$, $x(2) = -n$, and $x(3) = 1/n$. This example highlights the structural information that this theorem gives. It would seem to improve on global convergence claims which start from an a priori position that the second derivative matrix is bounded.

Remark 4.2.4 *This example shows that this set of approximants is not closed. Note also that the design matrix must become increasingly singular with n. It is of interest to note that the approximating function here has the form of a separable regression function (3.5.1). At least some of the methods that exploit separability in solving the parameter estimation problem can avoid this specific difficulty.*

Finding a suitable monitor for the quasi-likelihood iteration is a rather different proposition because of the lack of explicit values of $\mathcal{K}$, the incompletely defined function to be maximized. Just for the moment let the correction computed at the current step be denoted by $\mathbf{h}^*$. One possibility is the transformation invariant quantity ("natural criterion function")

$$\begin{aligned}\Phi_{\mathcal{K}}(\mathbf{x}+\lambda\mathbf{h}^*) &= \frac{1}{n}\nabla\mathcal{K}(\mathbf{x}+\lambda\mathbf{h}^*)\left(\mathcal{I}_K^n(\mathbf{x}+\lambda\mathbf{h}^*)\right)^{-1}\frac{1}{n}\nabla\mathcal{K}(\mathbf{x}+\lambda\mathbf{h}^*)^T,\\ &= \mathbf{h}^T(\mathbf{x}+\lambda\mathbf{h}^*)\,\mathcal{I}_K^n(\mathbf{x}+\lambda\mathbf{h}^*)\,\mathbf{h}(\mathbf{x}+\lambda\mathbf{h}^*). \qquad (4.2.8)\end{aligned}$$

This function vanishes at stationary points and is positive whenever $\mathbf{h}^*$ defines a direction of ascent for $\mathcal{K}$. The possible requirement to compute repeatedly $\left(\mathcal{I}_K^n(\mathbf{x}+\lambda\mathbf{h}^*)\right)^{-1}\nabla\mathcal{K}(\mathbf{x}+\lambda\mathbf{h}^*)^T$ (amounting to a full scoring correction) in order to carry out the line search is something of a disadvantage. However, at least the final adjustment in the line search provides also the next scoring step. Another disadvantage is that $\mathbf{h}$ is not guaranteed to be an ascent direction for $\Phi_{\mathcal{K}}(\mathbf{x})$. Here the condition is

$$\nabla\Phi_{\mathcal{K}}\mathbf{h} = \mathbf{h}^T\left\{2\nabla^2\mathcal{K}+\nabla\left\{\mathcal{I}_K^n[\mathbf{h}]\right\}\right\}\mathbf{h} > 0,$$

where the square brackets indicates that $\mathbf{h}$ is held constant in the differentiation. There are two difficult terms here as $\nabla^2\mathcal{K}$ is not guaranteed to be positive definite, nor is the gradient term involving $\mathcal{I}_K^n$. One modification - called affine invariance when applied to the Newton step [23], [25] - considers

$$\begin{aligned}\Phi_{\mathcal{K}}^A(\mathbf{x}+\lambda\mathbf{h}^*) &= \frac{1}{n}\nabla\mathcal{K}(\mathbf{x}+\lambda\mathbf{h}^*)\left(\mathcal{I}_K^n(\mathbf{x})\right)^{-1}\frac{1}{n}\nabla\mathcal{K}(\mathbf{x}+\lambda\mathbf{h}^*)^T,\\ &= \mathbf{h}_A^T(\mathbf{x}+\lambda\mathbf{h}^*)\,\mathcal{I}_K^n(\mathbf{x})\,\mathbf{h}_A(\mathbf{x}+\lambda\mathbf{h}^*), \qquad (4.2.9)\\ \mathbf{h}_A(\mathbf{x}+\lambda\mathbf{h}^*) &= \left(\mathcal{I}_K^n(\mathbf{x})\right)^{-1}\frac{1}{n}\nabla\mathcal{K}(\mathbf{x}+\lambda\mathbf{h}^*)^T.\end{aligned}$$

Here $\Phi_{\mathcal{K}}^A$ has the possible disadvantage that apparently acceptable corrections can lead to cycling with resultant stalling of the iteration in a related application [5] which is discussed in Remark 5.3.1. The condition that $\mathbf{h}$ be a descent direction for this objective also must be verified. The descent condition is a simpler calculation as the gradient of $\mathcal{I}_K^n$ now does not appear. The Simple form of the line search (4.2.3) would be the natural one to apply in the quasilikelihood calculations as it does not require derivatives of the monitor. An alternative ([77]) is to use a relatively robust method such as the secant algorithm to iterate towards a solution of

$$\nabla\mathcal{K}(\mathbf{x}+\lambda\mathbf{h}^*)\,\mathbf{h}^* = 0 \qquad (4.2.10)$$

which characterizes stationary points of $\mathcal{K}$ in the direction defined by $\mathbf{h}^*$. As function values are not available it is necessary to be conservative and ensure quite high accuracy in determining the root of (4.2.10) in order to be confident that $\mathcal{K}$ has increased in the computed step.

Invariance plus the classic result that the basic Newton iteration (4.1.1) has a second order rate of convergence implies that $\lambda = 1$ is a good initial choice in the line search step in this case. This result carries over to the forms of the scoring algorithm. Note that it does not depend on the method used to carry out the linesearch.

Lemma 4.5 *Consider the standard scoring algorithm with monitor Φ chosen equal to $\mathcal{L}$. Let the sequence of iterates $\{\mathbf{x}_i\}$ computed using the Goldstein linesearch converge to $\widehat{\mathbf{x}}_n$. Then*

$$\Psi\left(1, \mathbf{x}_i, \mathbf{h}_i\right) \approx .5,\ n \to \infty,$$

so that $\lambda = 1$ is acceptable eventually almost surely for all n large enough. The size of n is determined by the requirement that $\|\widehat{\mathbf{x}}_n - \mathbf{x}^*\|$ be small (that is by consistency).

Proof. Expanding $\mathcal{L}\left(\mathbf{y}; \mathbf{x}_i + \lambda \mathbf{h}_i, \mathbf{T}_n\right)$ using Taylor series, and assuming $\|\mathbf{h}_i\|$ is small, gives

$$\begin{aligned}
\Psi\left(1, \mathbf{x}_i, \mathbf{h}_i\right) &= \frac{\nabla_{\mathbf{x}}\mathcal{L}\mathbf{h}_i + \frac{1}{2}\mathbf{h}_i^T\nabla_{\mathbf{x}}^2\mathcal{L}\mathbf{h}_i + O\left(\|\mathbf{h}_i\|^3\right)}{\nabla_{\mathbf{x}}\mathcal{L}\mathbf{h}_i}, \\
&= 1 + \frac{1}{2}\frac{\frac{1}{n}\nabla_{\mathbf{x}}\mathcal{L}\mathcal{I}_n^{-1}\nabla_{\mathbf{x}}^2\mathcal{L}\mathcal{I}_n^{-1}\nabla_{\mathbf{x}}\mathcal{L}^T}{\nabla_{\mathbf{x}}\mathcal{L}\mathcal{I}_n^{-1}\nabla_{\mathbf{x}}\mathcal{L}^T} + O\left(\frac{1}{n^2}\|\nabla_{\mathbf{x}}\mathcal{L}\|\right), \\
&= 1 + \frac{1}{2}\frac{\nabla_{\mathbf{x}}\mathcal{L}\mathcal{I}_n^{-1}\left(-\mathcal{I}_n + \frac{1}{n}\nabla_{\mathbf{x}}^2\mathcal{L} + \mathcal{I}_n\right)\mathcal{I}_n^{-1}\nabla_{\mathbf{x}}\mathcal{L}^T}{\nabla_{\mathbf{x}}\mathcal{L}\mathcal{I}_n^{-1}\nabla_{\mathbf{x}}\mathcal{L}^T} + O\left(\frac{1}{n^2}\|\nabla_{\mathbf{x}}\mathcal{L}\|\right), \\
&= \frac{1}{2} + \frac{1}{2}\frac{\nabla_{\mathbf{x}}\mathcal{L}\mathcal{I}_n^{-1}\left(\frac{1}{n}\nabla_{\mathbf{x}}^2\mathcal{L} + \mathcal{I}_n\right)\mathcal{I}_n^{-1}\nabla_{\mathbf{x}}\mathcal{L}^T}{\nabla_{\mathbf{x}}\mathcal{L}\mathcal{I}_n^{-1}\nabla_{\mathbf{x}}\mathcal{L}^T} + O\left(\frac{1}{n^2}\|\nabla_{\mathbf{x}}\mathcal{L}\|\right), \\
&\approx \frac{1}{2},\ i \to \infty,\ n \text{ large enough.}
\end{aligned}$$

Here the law of large numbers and consistency of $\mathbf{x}_n$ are used to show the second term on the right hand side is small almost surely when n is large enough.

■

Subproblems in least squares form

The actual computation of the scoring step often can be reduced to that of solving a linear least squares problem with associated advantages in conditioning and scaling which correspond to those discussed for the linear least squares problem in Chapter 2. The main idea is illustrated in simplest form by the sample form of the scoring algorithm . In this case

$$\begin{aligned}\mathcal{S}_n &= \frac{1}{n}\sum_{\mathbf{t}\in\mathbf{T}_n}\nabla_{\mathbf{x}}L_t^T\nabla_{\mathbf{x}}L_t,\\ &= \frac{1}{n}S_n^TS_n,\end{aligned}$$

where

$$S_n = \begin{bmatrix}\vdots\\ \nabla_{\mathbf{x}}L_t\\ \vdots\end{bmatrix}.$$

Comparison with (4.1.4) shows that the sample scoring step is identical with the linear least squares problem:

$$\min_{\mathbf{h}}\mathbf{r}^T\mathbf{r};\ \mathbf{r} = S_n\mathbf{h}-\mathbf{e}. \tag{4.2.11}$$

The other cases require rather more work. To adapt equation (4.1.7) for the quasilikelihood step we have

$$\begin{aligned}\mathcal{I}_K &= \frac{1}{n}\sum_{\mathbf{t}\in\mathbf{T}_n}\nabla_{\mathbf{x}}\boldsymbol{\mu}_t^TV\left(\boldsymbol{\mu}_t\right)^{-1}\nabla_{\mathbf{x}}\boldsymbol{\mu}_t,\\ &= \frac{1}{n}\left(I_K^n\right)^TI_K^n,\end{aligned}$$

where

$$I_K^n = \begin{bmatrix}\vdots\\ V_t^{-1/2}\nabla_{\mathbf{x}}\boldsymbol{\mu}_t\\ \vdots\end{bmatrix}.$$

The least squares problem equivalent to (4.1.7) is [77]

$$\min_{\mathbf{h}}\mathbf{r}^T\mathbf{r};\ \mathbf{r} = I_K^n\mathbf{h}-\mathbf{b}, \tag{4.2.12}$$

where

$$\mathbf{b} = \begin{bmatrix}\vdots\\ V_t^{-1/2}\left(\mathbf{y}_t-\boldsymbol{\mu}_t\right)\\ \vdots\end{bmatrix}.$$

To set the basic scoring method (4.1.2) in this framework write the component terms as

$$\mathcal{I}_n = \frac{1}{n}\sum_{\mathbf{t}\in\mathbf{T}_n} \nabla_{\mathbf{x}}\boldsymbol{\eta}_t^T \mathcal{E}\left\{\nabla_\eta L_t^T \nabla_\eta L_t\right\} \nabla_{\mathbf{x}}\boldsymbol{\eta}_t, \tag{4.2.13}$$

$$\nabla_x \mathcal{L}^T = \sum_{\mathbf{t}\in\mathbf{T}_n} \nabla_x \boldsymbol{\eta}_t^T \nabla_\eta L_t^T. \tag{4.2.14}$$

For comparison with the quasilikelihood algorithm it is convenient to write

$$V_t^{-1} = \mathcal{E}\left\{\nabla_\eta L_t^T \nabla_\eta L_t\right\}, \tag{4.2.15}$$

$$I_L^n = \begin{bmatrix} \vdots \\ V_t^{-1/2}\nabla_{\mathbf{x}}\boldsymbol{\eta}_t \\ \vdots \end{bmatrix}, \tag{4.2.16}$$

$$\nabla_x \mathcal{L}^T = \sum_t \nabla_x \boldsymbol{\eta}_t^T V_t^{-T/2}\mathbf{b}_t = I_L^{nT}\mathbf{b}. \tag{4.2.17}$$

where

$$\mathbf{b} = \begin{bmatrix} \vdots \\ V_t^{T/2}\nabla_\eta L_t^T \\ \vdots \end{bmatrix}. \tag{4.2.18}$$

The form of $\mathbf{b}$ is chosen so the least squares necessary conditions correspond to the equations for the scoring step. The scoring iteration can now be written as the least squares problem

$$\min_{\mathbf{h}} \mathbf{r}^T\mathbf{r};\ \mathbf{r} = I_L^n\mathbf{h} - \mathbf{b}, \tag{4.2.19}$$

Example 4.2.1 *In the case of the normal distribution a typical situation could have*

$$y_t \backsim \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu(\mathbf{x}^*,t))^2}{2\sigma^2}},$$

and the corresponding terms in the likelihood would be

$$L_t = -\frac{1}{2\sigma^2}\left(y_t - \mu\left(\mathbf{x},t\right)\right)^2 + \ const.$$

Here the maximum likelihood formulation is equivalent to a nonlinear least squares problem. To set up the scoring iteration

$$\nabla_{\mathbf{x}} L_t = \frac{1}{\sigma^2}\left(y_t - \mu\left(\mathbf{x},t\right)\right)\nabla_{\mathbf{x}}\mu\left(\mathbf{x},t\right),$$

$$\nabla_{\mathbf{x}}^2 L_t = -\frac{1}{\sigma^2}\left(\nabla_{\mathbf{x}}\mu\left(\mathbf{x},t\right)^T\nabla_{\mathbf{x}}\mu\left(\mathbf{x},t\right) - \left(y - \mu\left(\mathbf{x},t\right)\right)\nabla_{\mathbf{x}}^2\mu\left(\mathbf{x},t\right)\right),$$

giving (compare Remark 4.1.2)

$$\mathcal{I}_n = \frac{1}{n\sigma^2} \sum_{t \in T_n} \nabla_{\mathbf{x}} \mu (\mathbf{x}, t)^T \nabla_{\mathbf{x}} \mu (\mathbf{x}, t) .$$

The factor $1/n\sigma^2$ *cancels in the equations determining the scoring algorithm correction, and these can be written*

$$\min_{\mathbf{h}} \mathbf{r}^T \mathbf{r};\ \mathbf{r} = I_L^n \mathbf{h} - (\mathbf{y} - \boldsymbol{\mu} (\mathbf{x})) ,$$

where

$$I_L^n = \begin{bmatrix} \vdots \\ \nabla_{\mathbf{x}} \mu (\mathbf{x}, t) \\ \vdots \end{bmatrix} .$$

In this form the basic step of the algorithm is identical with the Gauss Newton algorithm for nonlinear least squares problems. The corresponding sample form sets

$$\mathcal{S}_n = \frac{1}{n} \sum_{\mathbf{t} \in \mathbf{T}_n} \nabla_{\mathbf{x}} L_t^T \nabla_{\mathbf{x}} L_t = \frac{1}{n\sigma^4} \sum_{\mathbf{t} \in \mathbf{T}_n} (y_t - \mu (\mathbf{x}, t))^2 \nabla_{\mathbf{x}} \mu (\mathbf{x}, t)^T \nabla_{\mathbf{x}} \mu (\mathbf{x}, t) .$$

The linear least squares algorithm for the sample scoring correction is

$$\min_{\mathbf{h}} \mathbf{r}^T \mathbf{r};\ \mathbf{r} = S_n \mathbf{h} - \mathbf{e}$$

where

$$S_n = \begin{bmatrix} \vdots \\ \frac{(y_t - \mu(\mathbf{x}, t))}{\sigma^2} \nabla_{\mathbf{x}} \mu (\mathbf{x}, t) \\ \vdots \end{bmatrix} .$$

An interesting feature here is that σ^2 *appears explicitly. It can be absorbed into* $\mathbf{h}$*, but then the scale of the correction depends on* σ^2 *and contains information on its value. However, both the direction of search, and the minimum along the resulting line are independent of* σ*. Here the sampling form of scoring is not attractive in comparison with the Gauss-Newton method. A case where it is useful is considered later (Example (4.7.1)).*

Exercise 4.2.1 *Consider the separable model (3.5.1). Show that the corresponding least squares problem for the Gauss-Newton correction in the case of normally distributed errors is*

$$\begin{bmatrix} A (\boldsymbol{\beta}) & \nabla_{\beta} A (\boldsymbol{\beta}) [\boldsymbol{\alpha}] \end{bmatrix} \begin{bmatrix} \mathbf{h}_{\alpha} \\ \mathbf{h}_{\beta} \end{bmatrix} = \mathbf{r},$$

where

$$\mathbf{r} = \mathbf{y} - A (\boldsymbol{\beta}) \boldsymbol{\alpha}.$$

4.2.2 Some computational details

Consider first the nonlinear least squares problem:

$$\min_{\mathbf{x}\in S} \mathbf{s}^T\mathbf{s}; \quad \mathbf{s} = \mathbf{f}(\mathbf{x}) - \mathbf{z} \tag{4.2.20}$$

where $\mathbf{z} \sim N(\mathbf{f}(\mathbf{x}^*), \sigma^2 I)$. Here $\mathbf{z}$ can be considered the observed data while $\mathbf{f}(\mathbf{x})$ provides the underlying model. The scoring (Gauss-Newton) algorithm predicts a correction step by solving the linear subproblem which corresponds to (4.1.2):

$$\min_{\mathbf{h}} \mathbf{r}^T\mathbf{r}; \quad \mathbf{r} = A\mathbf{h} - \mathbf{b}, \tag{4.2.21}$$

where $A = \nabla\mathbf{f}(\mathbf{x})$, A is assumed to have full column rank p, and $\mathbf{b} = \mathbf{z} - \mathbf{f}(\mathbf{x})$. Note that scoring is a method for maximizing a likelihood while Gauss-Newton is a method for minimizing a sum of squares which in this context is effectively the negative of a likelihood so there is some potential to be confused by "-" signs.

The necessary conditions for a minimum of (4.2.20) are

$$\begin{aligned}\mathbf{s}^T A &= (\mathbf{f} - \mathbf{z})^T A = 0,\\ &= (\mathbf{f} + A\mathbf{h} - \mathbf{z})^T A - \mathbf{h}^T A^T A,\\ &= \mathbf{r}^T A - \mathbf{h}^T A^T A\end{aligned}$$

Thus the linear subproblem (4.2.21) has the solution $\mathbf{h} = 0$ at the maximum likelihood estimate. Taking the scalar product of the linear subproblem with $\mathbf{r}$ gives

$$\|\mathbf{r}\|^2 = -\mathbf{r}^T\mathbf{b} \Rightarrow \|\mathbf{r}\| \leq \|\mathbf{b}.\|.$$

Also, taking the scalar product with $\mathbf{b}$,

$$\begin{aligned}-\mathbf{r}^T\mathbf{b} &= -\mathbf{b}^T A\mathbf{h} + \|\mathbf{b}\|^2,\\ &= -(\mathbf{b} + \mathbf{r})^T A\mathbf{h} + \|\mathbf{b}\|^2,\\ &= -\|\mathbf{b} + \mathbf{r}\|^2 + \|\mathbf{b}\|^2.\end{aligned}$$

From this follows conditions characterizing a stationary point.

Lemma 4.6 *The following chain of implications holds at a stationary point of (4.2.20):*

$$\|\mathbf{r}\| = \|\mathbf{b}\| \Rightarrow \mathbf{r} + \mathbf{b} = 0 \Rightarrow \nabla_x \mathbf{f}^T\mathbf{h} = 0 \Rightarrow \|\mathbf{b}\| = \|\mathbf{r}\|. \tag{4.2.22}$$

Remark 4.2.5 *The size of* $\|\mathbf{b}\|^2 - \|\mathbf{r}\|^2$ *provides a convenient test for convergence involving quantities readily available from the solution of the linear subproblem (4.2.21). The correspondence to quantities computed from the log likelihood is*

$$\left(\|\mathbf{b}\|^2 - \|\mathbf{r}\|^2\right) = \nabla_x \mathcal{L}\mathbf{h}$$

so that it shares the transformation invariance properties (4.2.5).

In the case of nonlinear least squares $L_t = -\frac{1}{2}b_t^2$ so the Goldstein condition becomes

$$\begin{aligned}\Psi(\lambda, \mathbf{x}, \mathbf{h}) &= \frac{\left(\mathbf{b}^T\mathbf{b}\right)(\mathbf{x}+\mathbf{h}) - \left(\mathbf{b}^T\mathbf{b}\right)(\mathbf{x})}{2\lambda\mathbf{b}^T A\mathbf{h}}, \\ &= \frac{\left(\mathbf{b}^T\mathbf{b}\right)(\mathbf{x}+\mathbf{h}) - \left(\mathbf{b}^T\mathbf{b}\right)(\mathbf{x})}{2\left(\|\mathbf{b}\|^2 - \|\mathbf{r}\|^2\right)}. \end{aligned} \tag{4.2.23}$$

Remark 4.2.6 *The linear least squares form of the linear subproblem for the scoring algorithm for maximum likelihood estimation (4.2.19) suggests the use of orthogonal transformation methods as the preferred method for its numerical solution. Here these have the additional advantage that required auxiliary quantities can be computed readily. For example, let*

$$I_L^n = Q \begin{bmatrix} U \\ 0 \end{bmatrix}.$$

then

$$\begin{aligned}\nabla_x \mathcal{L}\mathbf{h} &= \mathbf{b}^T I_L^n \mathbf{h} \\ &= \mathbf{b}^T Q \begin{bmatrix} U \\ 0 \end{bmatrix} U^{-1} Q_1^T \mathbf{b} \\ &= \|Q_1^T \mathbf{b}\|^2. \end{aligned} \tag{4.2.24}$$

Note that this method for computing $\nabla_x \mathcal{L}\mathbf{h} = \|\mathbf{b}\|^2 - \|\mathbf{r}\|^2$ *necessarily gives a nonnegative result.*

A line search method

Two tests (4.2.1,4.2.2) and (4.2.3) have been presented for characterizing the value of λ determining the increment in $\mathbf{x}$ in the direction of $\mathbf{h}$. The value $\lambda = 1$ is preferred as it is associated with fast convergence under appropriate conditions (see section 3). As the algorithms are transformation invariant this suggests the use of $\lambda = 1$ as an initial increment and it is acceptable if

$\mathcal{L}(\mathbf{x}+\mathbf{h}) > \mathcal{L}(\mathbf{x})$. If it is not acceptable then it is necessary to predict an improved value. This is well defined in the Simple test. For the Goldstein test the information available after testing a step $\lambda_1\mathbf{h}$ ($\lambda_1 = 1$ initially) is $\{\mathcal{L}(\mathbf{x}),\ \mathcal{L}(\mathbf{x}+\lambda_1\mathbf{h}),\ \nabla_x\mathcal{L}(\mathbf{x})\mathbf{h}\}$. This is sufficient to permit construction of the quadratic approximation

$$G(\lambda) = \mathcal{L}(\mathbf{x}) + \lambda\nabla_x\mathcal{L}(\mathbf{x})\mathbf{h} + \lambda^2 C,$$

where C is to be determined to satisfy the condition

$$G(\lambda_1) = \mathcal{L}(\mathbf{x}+\lambda_1\mathbf{h}).$$

so that

$$\begin{aligned} C &= \frac{1}{\lambda^2}\left\{\mathcal{L}(\mathbf{x}+\lambda_1\mathbf{h}) - \mathcal{L}(\mathbf{x}) - \lambda_1\nabla_x\mathcal{L}(\mathbf{x})\mathbf{h}\right\}, \\ &= \frac{1}{\lambda_1}\nabla_x\mathcal{L}\mathbf{h}\left\{\Psi(\lambda_1) - 1\right\}. \end{aligned}$$

Note that $C < 0$ if the Goldstein test fails with $\Psi < \sigma$ so that $G(\lambda)$ has a maximum in $0 < \lambda < \lambda_1$. This is given by

$$0 = \frac{dG}{d\lambda} = \nabla_x\mathcal{L}\mathbf{h} + 2\lambda C$$

so that

$$\begin{aligned} \lambda &= -\frac{\nabla_x\mathcal{L}\mathbf{h}}{2C}, \\ &= \frac{\lambda_1}{2\left(1-\Psi(\lambda_1)\right)}. \end{aligned} \tag{4.2.25}$$

This step predicts an increase in $\mathcal{L}$ and so can be used for the new increment for evaluating Ψ. A test that combines the good features of both the Goldstein and Simple tests is :

$$\lambda_2 = \max\left(\rho\lambda_1, \frac{\lambda_1}{2\left(1-\Psi(\lambda_1)\right)}\right). \tag{4.2.26}$$

Setting a maximum net number of reductions in λ serves a similar purpose to the condition $\Psi < 1-\varrho,\ \lambda \to 0$ in the Goldstein test in the sense of flagging the possibility of the exceptional case if $\inf\lambda = 0$.

4.2.3 Trust region methods

The basic trust region method imposes a bound constraint on the solution of the linear subproblem determining the successive corrections of the ascent calculation. The idea is that the bound should be chosen such that the accepted (full) step satisfies an acceptability criterion such as the Goldstein test (4.2.1,4.2.2). In this case the bound is determining a region in which the linear approximation is not too out of kilter with the nonlinear behaviour of the objective in the sense expressed by the Goldstein test. The earliest reports of the application of these methods include ([62],[72], and [67]). They considered their use in the implementation of the Gauss-Newton method for nonlinear least squares. They have proved popular in the development of optimization software, possibly more popular than line search methods in many cases. Here one advantage is that the (modified design) matrix to be inverted at each step is always at least as well conditioned as the information matrix in the basic scoring algorithm. A possible cost is that improved conditioning could be bought by slower convergence.

To develop the basic ideas consider, for example, the linear subproblem associated with the sample information (4.2.11). Let $\|\mathbf{h}\|_D^2 = \mathbf{h}^T D^2 \mathbf{h}$, $D > 0$, diagonal. Incorporating the bound constraint on $\|\mathbf{h}\|_D^2$ leads to the constrained least squares problem

$$\min_{\mathbf{h}, \|\mathbf{h}\|_D^2 \leq \gamma} \mathbf{r}^T \mathbf{r};\ \mathbf{r} = S_n \mathbf{h} - \mathbf{e}. \tag{4.2.27}$$

The Kuhn-Tucker necessary conditions give

$$\left[\ \mathbf{r}^T \quad 0\ \right] = \boldsymbol{\lambda}^T \left[\ I \quad -S_n\ \right] - \pi \left[\ 0 \quad \mathbf{h}^T D^2\ \right],$$

where $\boldsymbol{\lambda}, \pi$ are the multipliers associated with the equality and inequality constraints respectively. It follows that

$$\boldsymbol{\lambda} = \mathbf{r},\ \mathbf{r}^T S_n + \pi \mathbf{h}^T D^2 = 0.$$

Thus $\mathbf{h}$ satisfies the system

$$\left(\mathcal{S}_n + \frac{\pi}{n} D^2\right) \mathbf{h} = \frac{1}{n} S_n^T \mathbf{e} = \frac{1}{n} \nabla_{\mathbf{x}} \mathcal{L}^T \tag{4.2.28}$$

and is an ascent direction if $\pi \geq 0$. If $\pi = 0$ then the inequality constraint is generically inactive and the scoring correction is obtained.

Remark 4.2.7 *It is possible to use the multiplier π rather than the bound γ to control $\|\mathbf{h}\|_D^2$. The key result is that $\|\mathbf{h}\|_D$ is monotonic decreasing as π increases. To show this differentiate (4.2.28) with respect to π. This gives*

$$\left(\mathcal{S}_n + \frac{\pi}{n} D^2\right) \frac{d\mathbf{h}}{d\pi} = -\frac{1}{n} D^2 \mathbf{h}.$$

It follows that

$$\frac{d\mathbf{h}^T}{d\pi}D^2\mathbf{h} = -n\frac{d\mathbf{h}^T}{d\pi}\left(\mathcal{S}_n + \frac{\pi}{n}D^2\right)\frac{d\mathbf{h}}{d\pi} < 0,$$
$$\Rightarrow \frac{d}{d\pi}\left\|\mathbf{h}\right\|_D^2 < 0.$$

Thus a simple and effective strategy for controlling the size of π, which was used in the initial implementations [62] and [67], keeps a multiplier ($\alpha > 1$ say) to increase π if $\mathbf{h}$ fails the acceptability test, and a second $\beta < 1$ to decrease π if the acceptability test is passed easily. Typically $\alpha\beta < 1$ to favour the scoring step with its known favourable convergence rate as the iteration proceeds ($\alpha = 1.5$, $\beta = .5$ is recommended in [76] but the choices do no appear critical), but $\pi = 0$ is not achieved by this approach without separate intervention, and this could be a disadvantage as we know from Lemma 4.5 that $\pi = 0$ is appropriate eventually if n is large enough. This problem can be overcome by controlling γ [71]. With the α, β strategy a successful step will always be taken eventually because

$$\mathbf{h} \to \frac{1}{\pi}D^{-2}\nabla_{\mathbf{x}}\mathcal{L}^T,\ \pi \to \infty,$$

and this is necessarily an ascent direction. Because of this asymptotic relation it follows that if the sequence $\{\pi_i\}$ of multiplier values is bounded then $\exists\, \varrho > 0$ such that the left hand inequality in the Goldstein condition (4.2.2) is always satisfied. Thus the boundedness of the multiplier sequence plays the same role as the condition that the line search step be bounded away from zero.

Remark 4.2.8 *The form of the trust region constraint interferes with the good scaling properties of the scoring algorithm . The best that can be hoped for in practical terms is that the linear subproblem has a useful invariance property with respect to diagonal scaling. Introduce the new variables $\mathbf{u} = T\mathbf{x}$ where T is diagonal. Then, using subscripts to distinguish $\mathbf{x}$, $\mathbf{u}$ variables,*

$$T^{-1}\left(S_x^T S_x + \pi D^2\right)T^{-1}T\mathbf{h}_x = T^{-1}\nabla_{\mathbf{x}}\mathcal{L}^T.$$

This is equivalent to

$$\left(S_u^T S_u + \pi T^{-1}D^2T^{-1}\right)\mathbf{h}_u = \nabla_{\mathbf{u}}\mathcal{L}^T.$$

Thus if D_i transforms with $\frac{\partial}{\partial x_i}$ then $T_i^{-1}D_i$ transforms in the same way with respect to $\frac{\partial}{\partial u_i}$. This requirement is satisfied by

$$D_i = \left\|(S_n)_{*i}\right\|.$$

This transformation effects a rescaling of the least squares problem. We have

$$\begin{aligned}\mathbf{h} &= \left(S_n^T S_n + \pi D^2\right)^{-1} S_n^T \mathbf{e},\\ \Rightarrow D\mathbf{h} &= \left(D^{-1} S_n^T S_n D^{-1} + \pi I\right)^{-1} D^{-1} S_n^T \mathbf{e}.\end{aligned}$$

The effect of this choice is to rescale the columns of S_n to have unit length, a strategy recommended in [46]. It is often sufficient to set $\pi = 1$, and $D = \operatorname{diag}\{\|(S_n)_{*i}\|,\ i = 1, 2, \cdots, p\}$ initially [76]. However, if there are significant fluctuations in the size of the elements of S_n then [71] recommends updating D by

$$D_i = \max\{D_i,\ \|(S_n)_{*i}\|\}.$$

Theorem 4.3 *Let the sequence of iterates $\{\mathbf{x}_i\}$ produced by the trust region algorithm using the Goldstein criterion (4.2.2) be contained in a compact region R in which the sequence of values $\{\pi_i\}$ are bounded ($\leq \pi$). Then the sequence of values $\{\mathcal{L}(\mathbf{y}; \mathbf{x}_i, \mathbf{T}_n)\}$ converges, and limit points of $\{\mathbf{x}_i\}$ are stationary points of $\mathcal{L}$.*

Proof. This is very similar to that in Theorem 4.1. Convergence of the bounded, increasing sequence $\{\mathcal{L}(\mathbf{y}; \mathbf{x}_i, \mathbf{T}_n)\}$ is a consequence of the ascent condition. It follows from (4.2.1) that

$$\begin{aligned}\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{y}; \mathbf{x}_i, \mathbf{T}_n)\,\mathbf{h}_i &\leq \frac{\mathcal{L}(\mathbf{y}; \mathbf{x}_i + \mathbf{h}_i, \mathbf{T}_n) - \mathcal{L}(\mathbf{y}; \mathbf{x}_i, \mathbf{T}_n)}{\varrho},\\ &\to 0,\ i \to \infty,\end{aligned}$$

as the numerator on the right hand side is the difference between consecutive terms in a convergenct sequence. Combining this with (4.2.6) gives

$$\frac{\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{y}; \mathbf{x}_i, \mathbf{T}_n)\|\,\|\mathbf{h}_i\|}{\kappa_R} \leq \frac{\mathcal{L}(\mathbf{y}; \mathbf{x}_i + \mathbf{h}_i, \mathbf{T}_n) - \mathcal{L}(\mathbf{y}; \mathbf{x}_i, \mathbf{T}_n)}{\varrho},$$

where $\kappa_R = \operatorname{cond}\left(\mathcal{S}_n + \frac{\pi}{n}D^2\right)^{1/2}$. Note $\kappa_R \to 1$ as $\pi \to \infty$. But

$$\|\mathbf{h}\| = \left\|\left(\mathcal{S}_n + \frac{\pi}{n}D^2\right)^{-1}\frac{1}{n}S_n^T\mathbf{e}\right\| \geq \frac{\|\nabla_{\mathbf{x}}\mathcal{L}\|}{n\left\|\mathcal{S}_n + \frac{\pi}{n}D^2\right\|}.$$

Thus, if ν_R is an upper bound in R for $\left\|\mathcal{S}_n + \frac{\pi}{n}D^2\right\|$,

$$\frac{\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{y}; \mathbf{x}_i, \mathbf{T}_n)\|^2}{n\kappa_R\nu_R} \leq \frac{\mathcal{L}(\mathbf{y}; \mathbf{x}_i + \mathbf{h}_i, \mathbf{T}_n) - \mathcal{L}(\mathbf{y}; \mathbf{x}_i, \mathbf{T}_n)}{\varrho}.$$

It follows that $\{\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{y}; \mathbf{x}_i, \mathbf{T}_n)\|\} \to 0$ because $\nu_R < \infty$ by assumption.

■

The (α, β) form of the trust region algorithm provides information about the case $\{\pi_i\} \uparrow \infty$. The only additional information assumed is that the elements of the scaling matrix D are bounded away from zero - $D_i \geq \xi > 0$, $i = 1, 2, \cdots, p$. In this case it is convenient to write $\Psi = \Psi(\pi, \mathbf{x}, \mathbf{h})$ in the Goldstein test to emphasise the role played by the multiplier.

Theorem 4.4 *Assume that the sequence $\{\pi_i\}$ determined by the (α, β) form of the trust region algorithm is unbounded. Then the norm of the Hessian $\nabla^2_{\mathbf{x}}\mathcal{L}$ is also unbounded.*

Proof. If $\{\pi_i\}$ is unbounded then there exists an unbounded subsequence $\{\pi_i^*\}$ with the property that $\varrho > \Psi(\pi_i^*/\alpha, \mathbf{x}_i^*, \mathbf{h}_i^*)$ because there must be an infinite sequence of points at which the (α, β) test (hence also the Goldstein test) fails causing π_i to be increased. Thus there exists an unbounded sequence $\{\widehat{\pi}_i\}$ with the property that $\varrho > \Psi\left(\widehat{\pi}_i, \mathbf{x}_i, \widehat{\mathbf{h}}_i\right)$ where $\widehat{\mathbf{h}}_i$ is the tentative step generated by the linear subproblem at $\mathbf{x} = \mathbf{x}_i$ for $\pi = \widehat{\pi}_i$. Denoting mean values by a bar this gives

$$\varrho > \frac{\left|\nabla_{\mathbf{x}}\mathcal{L}\widehat{\mathbf{h}}_i\right| - \frac{1}{2}\left|\widehat{\mathbf{h}}_i^T\overline{\nabla^2_{\mathbf{x}}\mathcal{L}}\widehat{\mathbf{h}}_i\right|}{\left|\nabla_{\mathbf{x}}\mathcal{L}\widehat{\mathbf{h}}_i\right|}$$

so that

$$\left|\widehat{\mathbf{h}}_i^T\overline{\nabla^2_{\mathbf{x}}\mathcal{L}}\widehat{\mathbf{h}}_i\right| > 2(1-\varrho)\left|\nabla_{\mathbf{x}}\mathcal{L}\widehat{\mathbf{h}}_i\right|.$$

Thus

$$\left\|\overline{\nabla^2_{\mathbf{x}}\mathcal{L}}\right\| > 2(1-\varrho)\frac{\left|\nabla_{\mathbf{x}}\mathcal{L}\widehat{\mathbf{h}}_i\right|}{\left\|\widehat{\mathbf{h}}_i\right\|^2}.$$

Now

$$\begin{aligned}\left|\nabla_{\mathbf{x}}\mathcal{L}\widehat{\mathbf{h}}_i\right| &= n\widehat{\mathbf{h}}_i^T\left(\mathcal{S}_n + \frac{\pi}{n}D^2\right)\widehat{\mathbf{h}}_i, \\ &\geq \pi\widehat{\mathbf{h}}_i^T D^2\widehat{\mathbf{h}}_i.\end{aligned}$$

It follows that

$$\left\|\overline{\nabla^2_{\mathbf{x}}\mathcal{L}}\right\| > 2\pi(1-\varrho)\min_i D_i^2.$$

■

Remark 4.2.9 *As in remark 4.2.3 iterations can become unbounded for smooth sets of approximating functions as the closure of these approximating sets cannot be guaranteed in the nonlinear case.*

Some computational details

The linear subproblem at iteration step i has the generic form

$$\min_{\mathbf{h}} \mathbf{r}^T\mathbf{r}; \quad \mathbf{r} = \begin{bmatrix} A_i \\ \pi_i^{1/2} I \end{bmatrix} \mathbf{h} - \begin{bmatrix} \mathbf{b}_i \\ 0 \end{bmatrix}. \tag{4.2.29}$$

It proves convenient to start each iteration by making an orthogonal factorization of A_i : $A_i \rightarrow Q_i \begin{bmatrix} U_i \\ 0 \end{bmatrix}$. It is assumed that this orthogonal factorization is well behaved - here column pivoting, corresponding to diagonal pivoting in the Cholesky factorization, could be used for extra stability but for current purposes this is assumed to be unnecessary.

Remark 4.2.10 *At this point* $\mathbf{h}_i(0)$ *can be determined cheaply. If* $\|\mathbf{h}_i(0)\| \leq \sigma\gamma_{i-1}$, *with* $0 < \sigma < 1$ *and* $\pi_i = \beta * \pi_{i-1}$ *where* σ *signals an appropriate decrease in the trust region radius,* β *is the multiplier for decreasing* π, *and both are preset constants, then the computation could be switched to the standard scoring algorithm with its known good convergence properties. It could be reset to the Levenberg iteration at any stage that a nontrivial linesearch step is required.*

Adjustment of $\mathbf{h}(\pi)$ is now carried out by working with the typically significantly smaller linear least squares problem:

$$\min_{\mathbf{h}} \mathbf{s}^T\mathbf{s}; \quad \mathbf{s} = \begin{bmatrix} U_i \\ \pi^{1/2} I \end{bmatrix} \mathbf{h} - \begin{bmatrix} \mathbf{c}_1 \\ 0 \end{bmatrix}. \tag{4.2.30}$$

This requires a further orthogonal factorization specific to the particular value of π that is current :

$$\begin{bmatrix} U_i \\ \pi^{1/2} I \end{bmatrix} \rightarrow Q' \begin{bmatrix} U_i' \\ 0 \end{bmatrix}, \tag{4.2.31}$$

$$Q'^T \begin{bmatrix} \mathbf{c}_1 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{c}_1' \\ \mathbf{c}_2' \end{bmatrix}. \tag{4.2.32}$$

As in the case of the scoring algorithm the results of the orthogonal factorization can be used in the computation of important auxiliary quantities. For example (compare the derivation of (4.2.24))

$$\begin{aligned} \nabla_x \mathcal{L}\mathbf{h} &= \mathbf{c}_1^T U \mathbf{h}, \\ &= \begin{bmatrix} \mathbf{c}_1^T & 0 \end{bmatrix} \begin{bmatrix} U \\ \pi^{1/2} I \end{bmatrix} \left(U^T U + \pi I\right)^{-1} \begin{bmatrix} U^T & \pi^{1/2} I \end{bmatrix} \begin{bmatrix} \mathbf{c}_1 \\ 0 \end{bmatrix}, \\ &= \|\mathbf{c}_1'\|^2. \end{aligned} \tag{4.2.33}$$

Adjusting the trust region parameter

One catch with the (α, β) form of the Levenberg algorithm has to do with the initial choice of π (say $\pi = \pi_0$). While $\mathbf{h}(\pi_0)$ is guaranteed to generate a descent direction, and while the unit step can be expected ultimately to be satisfactory for small π, there can be a serious requirement to adapt π_0 to ensure the resulting step $\mathbf{h}(\pi_0)$ is satisfactory. One example where problems are possible is provided by models containing exponential terms such as $e^{-x_k t}$ which on physical grounds should have negative exponents ($x_k > 0$) but which can be made positive ($x_k + h_k < 0$) by too large a step. That is by π_0 too small. This could result in the introduction of seriously large (negative) terms into the log likelihood with consequent complications in making subsequent decisions. One simple cure is to repeatedly increase π by α until the resulting $\mathbf{h}(\pi)$ meets the positivity requirements. An alternative approach is to note that there is a relatively straight forward fix which involves introducing a damping factor τ chosen such that the critical components of $\mathbf{x} + \tau\mathbf{h}(\pi_0) \geq 0$, but this is a line search rather than a trust region device. However, $\|\tau\mathbf{h}(\pi_0)\|$ provides a possible choice for a revised trust region bound γ. The associated computing problem is given γ find the corresponding π. This involves soving the equation

$$\|\mathbf{h}(\pi)\| = \gamma.$$

This can be done by the application of Newton's method as $\frac{d\mathbf{h}}{d\pi}$ can be found by solving the equation

$$\left(U^T U + \pi I\right) \frac{d\mathbf{h}}{d\pi} = -\mathbf{h}$$

which is obtained by differentiating the normal equations 4.2.28) determining $\mathbf{h}(\pi)$. It can be solved in tandem with the solution of (4.2.30) when this is written in least squares form:

$$\min_{\mathbf{h}} \mathbf{s}^T\mathbf{s}; \quad \mathbf{s} = \begin{bmatrix} U \\ \pi^{1/2} I \end{bmatrix} \frac{d\mathbf{h}}{d\pi} = \begin{bmatrix} -U^{-T}\mathbf{h} \\ 0 \end{bmatrix}.$$

Newton's method involves estimating the zero by solving a linear approximation to the function. However, a more intuitively satisfactory extrapolation can be found in this case by noting that if

$$A = W\Sigma V^T$$

is the singular value decomposition of the matrix in the least squares formulation, then

$$\mathbf{h} = \sum_{i=1}^{p} \frac{\sigma_i(\mathbf{w}_i^T\mathbf{b})}{\sigma_i^2 + \pi}\mathbf{v}_i$$

is a rational function of π. This suggests that a rational form for the local approximation of $\|\mathbf{h}\|$ could be appropriate. The form which mirrors the singular value decomposition solution and uses the same amount of information about the function as Newton's method is

$$\|\mathbf{h}\| \approx \frac{a}{b + (\pi - \pi_0)}.$$

To identify the parameters a and b given the values $\|\mathbf{h}(\pi_0)\|$ and $\frac{d}{d\pi}\|\mathbf{h}(\pi_0)\|$ we have

$$\begin{aligned} \|\mathbf{h}(\pi_0)\| &= \frac{a}{b}, \\ \frac{d}{d\pi}\|\mathbf{h}(\pi_0)\| &= \frac{\mathbf{h}^T \frac{d\mathbf{h}}{d\pi}}{\|\mathbf{h}\|}(\pi_0) = -\frac{a}{b^2}. \end{aligned}$$

The result is

$$b = -\frac{\|\mathbf{h}(\pi_0)\|}{\frac{d}{d\pi}\|\mathbf{h}(\pi_0)\|}, \quad a = \|\mathbf{h}(\pi_0)\|b,$$

giving the correction

$$\pi = \pi_0 - \frac{\|\mathbf{h}(\pi_0)\| - \gamma}{\frac{d}{d\pi}\|\mathbf{h}(\pi_0)\|}\frac{\|\mathbf{h}(\pi_0)\|}{\gamma}.$$

The effect of the rational extrapolation is just the Newton step modulated by the term $\frac{\|\mathbf{h}(\pi_0)\|}{\gamma}$. It follows immediately that this modified Newton iteration is also second order convergent. Measures to safeguard this iteration are discussed in [71].

Exercise 4.2.2 *Show that the same sequence of iterations is obtained by applying Newton's method to solve*

$$\frac{1}{\|\mathbf{h}(\pi)\|} = \frac{1}{\gamma}.$$

4.3 Estimation of the rate of convergence

The equivalence result, Lemma 4.1, suggests that the asymptotic rate of convergence of the scoring algorithms as $n \to \infty$ should approach that of the Newton iteration itself . It is worthwhile showing this explicitly because further useful points follow. First note that it follows from Lemma 4.5 that $\lambda = 1$ will be accepted ultimately so that the scoring iterations in either line search or trust region form can be written in the form of a fixed point

iteration for the purpose of estimating the rate of convergence. Here we consider the standard form (4.1.2) which becomes

$$\mathbf{x}_{i+1} = F_n(\mathbf{x}_i),$$

where

$$F_n(\mathbf{x}) = \mathbf{x} + \mathcal{I}_n(\mathbf{x})^{-1} \frac{1}{n} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x})^T. \tag{4.3.1}$$

The condition for convergence of the fixed point algorithm is

$$\varpi(F_n'(\widehat{\mathbf{x}}_n)) < 1,$$

where $\varpi(F_n'(\mathbf{x}_n))$ is the spectral radius of the variation $F_n' = \nabla_{\mathbf{x}} F_n$.

Theorem 4.5

$$\varpi(F_n'(\widehat{\mathbf{x}}_n)) \to 0, \ \textit{a.s.}, n \to \infty.$$

Proof. To calculate $\varpi(F_n'(\widehat{\mathbf{x}}_n))$ note that $\nabla_{\mathbf{x}} \mathcal{L}(\widehat{\mathbf{x}}_n) = 0$. Thus

$$F_n'(\widehat{\mathbf{x}}_n) = I + \mathcal{I}_n(\widehat{\mathbf{x}}_n)^{-1} \frac{1}{n} \nabla_{\mathbf{x}}^2 \mathcal{L}(\widehat{\mathbf{x}}_n), \tag{4.3.2}$$

$$= \mathcal{I}_n(\widehat{\mathbf{x}}_n)^{-1} \left(\mathcal{I}_n(\widehat{\mathbf{x}}_n) + \frac{1}{n} \nabla_{\mathbf{x}}^2 \mathcal{L}(\widehat{\mathbf{x}}_n) \right). \tag{4.3.3}$$

If the right hand side were evaluated at $\mathbf{x}^*$ then the result would follow from the strong law of large numbers which shows that the $p \times p$ matrix $F_n'(\mathbf{x}^*)$ gets small (hence ϖ gets small) almost surely as $n \to \infty$. However, by consistency of the estimates, we have

$$\varpi(F_n'(\widehat{\mathbf{x}}_n)) = \varpi(F_n'(\mathbf{x}^*)) + O(\|\widehat{\mathbf{x}}_n - \mathbf{x}^*\|), \text{ a.s., } n \to \infty,$$

and the desired result follows. ■

Remark 4.3.1 *This result shows that both the scoring algorithms have arbitrarily fast rates of first order convergence as $n \to \infty$. In other words, they get closer and closer to the second order convergence associated with the Newton method. This rate is actually achieved for generalised linear models when the link is the canonical link. Let*

$$L_t = y_t \theta_t - b(\theta_t),$$

Then

$$\nabla_x L_t = (y_t - \mu_t) \nabla_x \theta_t,$$
$$\nabla_x^2 L_t = -\nabla_x \theta_t^T \nabla_x^2 b(\theta_t) \nabla_x \theta_t - \mu_t \nabla_x^2 \theta_t.$$

If the link is canonical then $\nabla_x^2 \theta_t = 0$ and $\nabla_x^2 L_t = \mathcal{E}\{\nabla_x^2 L_t\}$.

Remark 4.3.2 *In the trust region form of the algorithm $\mathcal{I}_n$ is replaced by $\mathcal{I}_n + \frac{\pi}{n}D^2$. Thus the aditional requirement for asymptotic second order convergence is $\{\pi_i\} \to 0$.*

Remark 4.3.3 *It is an important fact that $\varpi\left(F_n'\left(\widehat{\mathbf{x}}_n\right)\right)$ is an affine invariant. Making use of the transformation rule $\nabla_x(*) = \nabla_u(*)T$ with constant matrix T in equation (4.3.2) gives*

$$\begin{aligned} F_n'\left(\widehat{\mathbf{x}}_n\right) &= I + T^{-1}\mathcal{I}_n\left(\widehat{\mathbf{x}}_n\right)^{-1}\frac{1}{n}\nabla_u^2\mathcal{L}\left(\widehat{\mathbf{u}}_n\right)T, \\ &= T^{-1}F_n'\left(\widehat{\mathbf{u}}_n\right)T. \end{aligned}$$

Thus F_n' transforms by a similarity transformation so the eigenvalues are invariant.

Remark 4.3.4 *If an incorrect choice of density $f\left(\mathbf{y};\boldsymbol{\theta}_t,\mathbf{t}\right)$ is made in constructing the likelihood then the basic scoring algorithm (4.1.2) uses as the expected Hessian (compare (3.2.15))*

$$\mathcal{I}_f^n\left(\mathbf{x}\right) = -\mathcal{E}_{f(\mathbf{x})}\left\{\mathcal{J}_f^n\left(\mathbf{x}\right)\right\}.$$

The effect of this misidentification gives

$$F_f\left(\mathbf{x}\right) = \mathbf{x} + \mathcal{I}_f^n\left(\mathbf{x}\right)^{-1}\frac{1}{n}\nabla_{\mathbf{x}}\mathcal{L}_f\left(\mathbf{y};\mathbf{x},\mathbf{T}_n\right)^T.$$

The condition for $\mathbf{x}_f^n$ to be a point of attraction is

$$\varpi\left(F_f'\left(\mathbf{x}_f^n\right)\right) = \varpi\left(\mathcal{I}_f^n\left(\mathbf{x}_f^n\right)^{-1}\left(\mathcal{I}_f^n\left(\mathbf{x}_f^n\right) + \frac{1}{n}\nabla_{\mathbf{x}}^2\mathcal{L}_f\left(\mathbf{y};\mathbf{x}_f^n,\mathbf{T}_n\right)\right)\right) < 1.$$

The limiting value of $\varpi\left(F_f'\left(\mathbf{x}_f^n\right)\right)$ as $n \to \infty$ is (compare (3.2.16))

$$\varpi\left\{\begin{array}{c} \left(\int_{S(\mathbf{T})}\mathcal{E}_f\left\{\nabla_{\mathbf{x}}L_f^T\nabla_{\mathbf{x}}L_f\right\}\rho\left(\mathbf{t}\right)d\mathbf{t}\right)^{-1} \\ \left(\int_{S(\mathbf{T})}\left[\int_{\mathrm{range}(\mathbf{y})}\left(g-f\right)\nabla_{\mathbf{x}}^2L_f d\mathbf{y}\right]\rho\left(\mathbf{t}\right)d\mathbf{t}\right) \end{array}\right\}.$$

Thus the condition that $\varpi\left(F_f'\left(\mathbf{x}_f^n\right)\right) < 1$ here is just the condition in Theorem 3.5 for $\mathbf{x}_f$ to be an isolated solution. It follows that there is an intimate connection between the condition for an isolated maximum of the misidentified likelihood function and the condition for the ultimate convergence of the scoring algorithm with a unit linesearch step.

Remark 4.3.5 *An alternative way to look at the above results notes that if* $\varpi\left(F_f'\left(\mathbf{x}_f^n\right)\right)$ *is small then it is likely that the density used in constructing the likelihood is close to the true density. Thus the size of* $\varpi\left(F_f'\left(\mathbf{x}_f^n\right)\right)$ *provides confirmation of the modelling strategy. It is shown in Remark 4.3.3 that* $\varpi\left(F_f'\left(\mathbf{x}_f^n\right)\right)$ *is an invariant of the likelihood surface characterising local nonlinearity. It follows that if it is small then confidence intervals based on linear theory will be adequate for the parameter estimates, and if it is not then they wont and the modelling process must be regarded as suspect. Clearly knowledge of* $\varpi\left(F_f'\left(\mathbf{x}_f^n\right)\right)$ *is of value. Frequently it can be estimated from the ratio* $\|\mathbf{h}_{i+1}\| / \|\mathbf{h}_i\|$ *of successive corrections in the scoring algorithm provided the largest eigenvalue in modulus of* $F_f'\left(\mathbf{x}_f^n\right)$ *is isolated. This follows from*

$$\begin{aligned}\mathbf{h}_{i+1} &= F\left(\mathbf{x}_{i+1}\right) - F\left(\mathbf{x}_i\right) \\ &= F'\left(\mathbf{x}_i\right)\mathbf{h}_i + O\left(\|\mathbf{h}_i\|^2\right) \\ &= F'\left(\mathbf{x}^*\right)\mathbf{h}_i + O\left(\|\mathbf{x}_i - \mathbf{x}^*\|\|\mathbf{h}_i\|\right) + O\left(\|\mathbf{h}_i\|^2\right), \\ &\approx F'\left(\mathbf{x}^*\right)\mathbf{h}_i + \frac{1}{1-\varpi}O\left(\|\mathbf{h}_i\|^2\right).\end{aligned}$$

This shows that successive iterations look like steps of the classical power method for estimating the largest eigenvalue of $F'\left(\mathbf{x}^*\right)$*. Experience suggests it works well when convergence is fairly slow* ($\varpi \sim .1$)*. It will return a small value in the case of fast convergence, and this should be all that is needed in many cases.*

Exercise 4.3.1 *Assume the case of normal errors* $\boldsymbol{\varepsilon} \sim N\left(0, \sigma^2 I\right)$*. Show that the covariance matrix of* $F_n'\left(\widehat{\mathbf{x}}_n\right)$ *is small with* σ^2*. This shows the scoring algorithm converges rapidly in the case of small errors in this case.*

Exercise 4.3.2 *The consistency of nonlinear least squares estimators was considered in Exercise 3.2.1. Show that the Gauss-Newton algorithm provides fast convergence in large samples for the class of estimation problems considered there.*

4.4 Variable projection for separable problems

Separable problems are introduced in Section 3.5. In this section algorithmic possibilities arising from the special model structure (3.5.1) are analyzed. The treatment follows closely [80]. The resulting methods minimize a special sum of squares in a reduced number of variables. However, some of the

simplicity of the scoring algorithm is lost as a result of the variable projection transformation. The basic components of the scoring Gauss-Newton method for minimizing a sum of squares corresponding to the log likelihood associated with a signal in noise model and *additive*, independent normal errors are summarised in Remark 4.1.2. There is a corresponding form for minimizing a sum of squares,

$$S_n\left(\mathbf{x}, \boldsymbol{\varepsilon}^n\right) = \frac{1}{2n}\left\|\mathbf{s}^n\left(\mathbf{x}, \boldsymbol{\varepsilon}^n\right)\right\|_2^2, \tag{4.4.1}$$

which applies when the residuals s_i^n can be made small but the attractive stochastic properties do not hold. This form of the iteration is basically similar:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - H_i^{-1} \nabla S_n^T\left(\mathbf{x}_i\right), \tag{4.4.2}$$

$$H_i = \frac{1}{n}\left\{\nabla \mathbf{s}^n\left(\mathbf{x}_i\right)\right\}^T\left\{\nabla \mathbf{s}^n\left(\mathbf{x}_i\right)\right\}. \tag{4.4.3}$$

In the variable projection formulation $\mathbf{x} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix}$, and it is necessary to distinguish between the roles of the linear parameters $\boldsymbol{\alpha}$ and nonlinear parameters $\boldsymbol{\beta}$. The vector of residuals in the variable projection objective is

$$\mathbf{s}^n\left(\boldsymbol{\beta}, \boldsymbol{\varepsilon}^n\right) = P_n\left(\boldsymbol{\beta}\right)\mathbf{b}^n,$$

where the projection P_n is given by (3.5.5). The Gauss-Newton method applied to minimize the resulting sum of squares is referred to as the RGN algorithm . Several factors complicate the analysis of the RGN algorithm:

1. in this case there is coupling between the parameters and the noise;
2. this has the consequence that H_i does not correspond to the expected Hessian; and
3. rather harder work is needed to justify a similar large sample convergence rate to that of the scoring algorithm applied to the original signal in noise formulation.

Most methods for minimizing the variable projection sum of squares use a modification of the RGN algorithm due to Kaufman —indexKaufman algorithm [56]. This serves to reduce the amount of computation needed in the RGN algorithm. The Kaufman algorithm also shows the favourable large data set rates despite being developed using an explicit small residual (small σ) argument. It also proves to be actually closer to the scoring algorithm applied to the original problem statement than is the RGN algorithm.

Implementation of the RGN algorithm has been discussed in detail in [96]. It uses the approximate Hessian computed from (4.4.3) and this requires derivatives of $P_n(\boldsymbol{\beta})$. The derivative of P in the direction defined by $\mathbf{t} \in R^p$ is

$$\nabla_\beta P[\mathbf{t}] = -P\nabla_\beta \Phi[\mathbf{t}]\,\Phi^+ - \left(\Phi^+\right)^T \nabla_\beta \Phi^T[\mathbf{t}]\,P, \tag{4.4.4}$$

$$= A(\boldsymbol{\beta}, \mathbf{t}) + A^T(\boldsymbol{\beta}, \mathbf{t}), \tag{4.4.5}$$

where $A \in R^n \to R^n$, explicit dependence on both n and $\boldsymbol{\beta}$ is understood, and Φ^+ denotes the generalised inverse of Φ. Note that $\Phi^+ P = \Phi^+ - \Phi^+\Phi\Phi^+ = 0$ so the two components of $\nabla_\beta P[\mathbf{t}]$ in (4.4.5) are orthogonal. Define matrices $K, L : R^p \to R^n$ by

$$A(\boldsymbol{\beta}, \mathbf{t})\,\mathbf{b} = K(\boldsymbol{\beta}, \mathbf{b})\,\mathbf{t}, \tag{4.4.6}$$

$$A^T(\boldsymbol{\beta}, \mathbf{t})\,\mathbf{b} = L(\boldsymbol{\beta}, \mathbf{b})\,\mathbf{t}. \tag{4.4.7}$$

Then the RGN correction solves

$$\min_{\mathbf{t}} \|P\mathbf{b} + (K + L)\,\mathbf{t}\|^2, \tag{4.4.8}$$

where

$$L^T K = 0 \tag{4.4.9}$$

as a consequence of the orthogonality noted above.

Remark 4.4.1 *Kaufman [56] has examined these terms in more detail. We have*

$$\mathbf{t}^T K^T K \mathbf{t} = \mathbf{b}^T A^T A \mathbf{b} = O\left(\|\boldsymbol{\alpha}\|^2\right),$$
$$\mathbf{t}^T L^T L \mathbf{t} = \mathbf{b}^T A A^T \mathbf{b} = O\left(\|P\mathbf{b}\|^2\right).$$

If the orthogonality noted above is used then the second term in the design matrix in (4.4.8) corresponds to a small residual term when $\|P\mathbf{b}\|^2$ is relatively small and can be ignored . The resulting correction solves the linear least squares problem

$$\min_{\mathbf{t}} \|P\mathbf{b} + K\mathbf{t}\|^2. \tag{4.4.10}$$

Equation (4.4.10) is the basis of the modification suggested by Kaufman . It can be implemented with less computational cost, and it is favoured for this reason. Numerical experience is reported to be very satisfactory [38].

It is not possible to repeat exactly the rate of convergence calculation of the previous section because of the coupling between parameters and noise noted above. Here the fixed point form of the iteration is

$$\begin{aligned} \boldsymbol{\beta}_{i+1} &= F_n\left(\boldsymbol{\beta}_i\right), \\ F_n\left(\boldsymbol{\beta}\right) &= \boldsymbol{\beta} - H\left(\boldsymbol{\beta}\right)^{-1} \nabla_\beta S_n\left(\boldsymbol{\beta}\right)^T. \end{aligned} \tag{4.4.11}$$

The condition for $\widehat{\boldsymbol{\beta}}_n$ to be a fixed point is $\varpi\left(F_n'\left(\widehat{\boldsymbol{\beta}}_n\right)\right) < 1$:

$$F_n'\left(\widehat{\boldsymbol{\beta}}_n\right) = -\left(\frac{1}{n}\nabla_\beta \mathbf{s}^T \nabla_\beta \mathbf{s}\right)^{-1} \frac{1}{n}\left\{\sum_{i=1}^{n} s_i \nabla_\beta^2 s_i\right\}, \tag{4.4.12}$$

where the right hand side is evaluated at $\widehat{\boldsymbol{\beta}}_n$. The property of consistency is unchanged so the asymptotic convergence rate is again determined by $\varpi\left(F_n'\left(\boldsymbol{\beta}^*\right)\right)$. This expression is now examined in more detail.

Lemma 4.7 *Consider a regular sequence of experiments. Then*

$$\frac{1}{n}\Phi_n^T \Phi_n \underset{n\to\infty}{\overset{r.e.}{\to}} G, \tag{4.4.13}$$

where

$$G_{ij} = \int_0^1 \phi_i\left(t\right)\phi_j\left(t\right) w\left(t\right) dt,\ 1 \le i,j \le m,$$

and the density $w(t)$ is determined by the asymptotic properties of the sequence of sample points t_i^n, $i = 1, 2, \cdots, n$ for large n. The Gram matrix G is bounded and generically positive definite. Let $T_n = I - P_n$. Then

$$\left(T_n\right)_{ij} = \frac{1}{n}\boldsymbol{\phi}_i^T G^{-1} \boldsymbol{\phi}_j + o\left(\frac{1}{n}\right), \tag{4.4.14}$$

where

$$\boldsymbol{\phi}_i = \left[\ \phi_1\left(t_i\right) \quad \phi_2\left(t_i\right) \quad \cdots \quad \phi_m\left(t_i\right)\ \right]^T$$

This gives an $O\left(\frac{1}{n}\right)$ component-wise estimate which applies also to derivatives of both P_n and T_n with respect to $\boldsymbol{\beta}$.

Proof. The result (4.4.13) is discussed Chapter 1, in particular equation (1.1.7). Positive definiteness is a consequence of the problem rank assumption. To derive (4.4.14) note that

$$\begin{aligned} T_n &= \Phi_n\left(\Phi_n^T\Phi_n\right)^{-1}\Phi_n^T, \\ &= \frac{1}{n}\Phi_n G^{-1}\Phi_n^T + o\left(\frac{1}{n}\right). \end{aligned}$$

■

The starting point for determining the asymptotics of the convergence rate of the RGN algorithm as $n \to \infty$ is the computation of the expectations of the numerator and denominator matrices in (4.4.12). The expectation of the denominator is bounded and generically positive definite. The expectation of the numerator is $O\left(\frac{1}{n}\right)$ as $n \to \infty$. This suggests strongly that $\varpi\left(F_n'\left(\boldsymbol{\beta}^*\right)\right) \to 0,\ n \to \infty$, a result of essentially similar strength to that obtained for the additive error case. To complete the proof requires showing that both numerator and denominator terms converge to their expectations with probability 1.

Consider first the denominator term.

Lemma 4.8 *Let $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, and set*

$$b_i = \mu_i\left(\boldsymbol{\beta}^*\right) + \varepsilon_i,$$

where

$$\mu_j\left(\boldsymbol{\beta}\right) = \mathbf{e}_j^T \Phi \boldsymbol{\alpha}\left(\boldsymbol{\beta}\right).$$

The expectation of the denominator in (4.4.12) is

$$\frac{1}{n}\mathcal{E}\left\{\nabla_\beta \mathbf{s}^T \nabla_\beta \mathbf{s}\right\} = \sigma^2 M_1 + M_2, \tag{4.4.15}$$

where

$$M_1 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\nabla_\beta P_{ij}\right)^T \nabla_\beta P_{ij}, \tag{4.4.16}$$

$$M_2 = \frac{1}{n}\left\{\sum_{j=1}^{n}\nabla_\beta \mu_j^T \nabla_\beta \mu_j - \sum_{j=1}^{n}\sum_{k=1}^{n}\nabla_\beta \mu_j^T \nabla_\beta \mu_k T_{jk}\right\}, \tag{4.4.17}$$

and $M_1 = O\left(\frac{1}{n}\right),\ n \to \infty$ while M_2 tends to a limit which is a bounded, positive definite matrix when the problem rank assumption is satisfied.

Proof. Set

$$\begin{aligned}\nabla_\beta \mathbf{s}^T \nabla_\beta \mathbf{s} &= \sum_{i=1}^{n}\nabla_\beta s_i^T \nabla_\beta s_i, \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\nabla_\beta P_{ij}\right)^T b_j \sum_{k=1}^{n}\nabla_\beta P_{ik} b_k.\end{aligned} \tag{4.4.18}$$

To calculate the expectation note that

$$\mathcal{E}\left\{b_j b_k\right\} = \sigma^2 \delta_{jk} + \mu_j\left(\boldsymbol{\beta}^*\right)\mu_k\left(\boldsymbol{\beta}^*\right). \tag{4.4.19}$$

It follows that

$$\begin{aligned}\frac{1}{n}\mathcal{E}\left\{\nabla_\beta \mathbf{s}^T \nabla_\beta \mathbf{s}\right\} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\sigma^2 \sum_{j=1}^{n}\left(\nabla_\beta P_{ij}\right)^T \nabla_\beta P_{ij} + \sum_{j=1}^{n}\sum_{k=1}^{n}\mu_j\mu_k\left(\nabla_\beta P_{ij}\right)^T \nabla_\beta P_{ik}\right\}, \\ &= \sigma^2 M_1 + M_2\end{aligned}$$

To show $M_1 \to 0$ is a counting exercise. M_1 consists of the sum of n^2 terms each of which is an $p \times p$ matrix of $O(1)$ gradient terms divided by n^3 as a consequence of Lemma 4.7. M_2 can be simplified somewhat by noting that $\sum_{j=1}^{n} P_{ij}\mu_j = \mathbf{e}_i^T P \Phi \boldsymbol{\alpha}(\boldsymbol{\beta}) = 0$ identically in β so that

$$\sum_{j=1}^{n}\mu_j \nabla_\beta P_{ij} = -\sum_{j=1}^{n}\nabla_\beta \mu_j P_{ij}.$$

This gives, using the symmetry of $P = I - T$,

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\mu_j\mu_k\left(\nabla_\beta P_{ij}\right)^T \nabla_\beta P_{ik} &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\nabla_\beta \mu_j^T \nabla_\beta \mu_k P_{ij} P_{ik}, \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n}\nabla_\beta \mu_j^T \nabla_\beta \mu_k P_{jk}, \qquad (4.4.20) \\ &= \sum_{j=1}^{n}\nabla_\beta \mu_j^T \nabla_\beta \mu_j - \sum_{j=1}^{n}\sum_{k=1}^{n}\nabla_\beta \mu_j^T \nabla_\beta \mu_k T_{jk}.\end{aligned}$$

It follows from the estimates for the size of the T_{ij} computed in Lemma 4.7 that M_2 is bounded as $n \to \infty$. To show that M_2 is positive definite note that it follows from (4.4.20) that

$$\mathbf{t}^T M_2 \mathbf{t} = \frac{d\boldsymbol{\mu}}{dt}^T \{I - T\}\frac{d\boldsymbol{\mu}}{dt} \geq 0.$$

As $\left\|T\frac{d\boldsymbol{\mu}}{dt}\right\| \leq \left\|\frac{d\boldsymbol{\mu}}{dt}\right\|$, this expression can vanish only if there is a direction $\mathbf{t} \in R^p$ such that $\frac{d\boldsymbol{\mu}}{dt} = \gamma\boldsymbol{\mu}$ for some $\gamma \neq 0$. This requirement is contrary to the Gauss-Newton rank assumption that $\left[\begin{array}{cc}\Phi & \nabla_\beta \Phi\boldsymbol{\alpha}\end{array}\right]$ has full rank $m + p$. ■

Lemma 4.9 *The numerator in the expression (4.4.12) defining $F_n'(\boldsymbol{\beta}^*)$ is*

$$\sum_{i=1}^{n} s_i \nabla_\beta^2 s_i = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} b_j b_k P_{ij} \nabla_\beta^2 P_{ik}. \qquad (4.4.21)$$

Let $M_3 = \frac{1}{n}\mathcal{E}\left\{\sum_{i=1}^n s_i \nabla_\beta^2 s_i\right\}$ *then*

$$M_3 = \frac{1}{n}\sum_{i=1}^n \sigma^2 \left\{\nabla_\beta^2 P_{ii} - \sum_{j=1}^n T_{ij}\nabla_\beta^2 P_{ij}\right\}, \tag{4.4.22}$$

and $M_3 \to 0,\ n \to \infty$.

Proof. This is similar to that of Lemma 4.8 in using the component wise estimates of the derivatives of the projection matrices given in Lemma 4.7. The new point is that the contribution to M_3 from the signal terms $\mu_j(\boldsymbol{\beta}^*)$ in the expectation (4.4.19) is

$$\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \mu_j \mu_k P_{ij} \nabla_\beta^2 P_{ik} = 0$$

by summing over j keeping i and k fixed. The previous counting argument can be used again to give the estimate $M_3 = O\left(\frac{1}{n}\right),\ n \to \infty$. ■

The final step required is to show that the numerator and denominator terms in (4.4.12) approach their expectations as $n \to \infty$. Only the case of the denominator is considered here.

Lemma 4.10

$$\left(\frac{1}{n}\nabla_\beta \mathbf{s}^T \nabla_\beta \mathbf{s}\right) \underset{n\to\infty}{\overset{a.s.}{\to}} M_2. \tag{4.4.23}$$

Proof. The basic quantities are:

$$\begin{aligned}
\left(\frac{1}{n}\nabla_\beta \mathbf{s}^T \nabla_\beta \mathbf{s}\right) &= \frac{1}{n}\sum_{i=1}^n \nabla_\beta s_i^T \nabla_\beta s_i, \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \left(\nabla_\beta P_{ij}\right)^T b_j \sum_{k=1}^n \nabla_\beta P_{ik} b_k, \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \left\{\mu_j\mu_k + (\mu_j\varepsilon_k + \mu_k\varepsilon_j) + \varepsilon_j\varepsilon_k\right\}\left(\nabla_\beta P_{ij}\right)^T \nabla_\beta P_{ik}
\end{aligned}$$

The first of the three terms in this last expansion is M_2. Thus the result requires showing that the remaining terms tend to 0. Let

$$\boldsymbol{\pi}_i^n = \sum_{j=1}^n \varepsilon_j \left(\nabla_\beta P_{ij}\right)^T, \boldsymbol{\pi}_i^n \in R^p.$$

As, by Lemma 4.7, the components of $\nabla_\beta P_{ij} = O\left(\frac{1}{n}\right)$, it follows by applications of the law of large numbers that

$$\boldsymbol{\pi}_i^n \underset{n\to\infty}{\overset{a.s.}{\to}} 0,$$

componentwise. Specifically, given $\delta > 0$, there is an n_0 such that

$$\forall i,\ \left\|\boldsymbol{\pi}_i^n\right\|_\infty < \delta\ \forall n > n_0 \text{ with probability 1.}$$

Consider the third term. Let

$$\begin{aligned} S_n &= \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \varepsilon_j \varepsilon_k \left(\nabla_\beta P_{ij}\right)^T \nabla_\beta P_{ik}, \\ &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{\pi}_i^n \left(\boldsymbol{\pi}_i^n\right)^T. \end{aligned}$$

Then, in the maximum norm, with probability 1 for $n > n_0$,

$$\left\|S_n\right\|_\infty \leq p\delta^2,$$

showing that the third sum tends to 0, $n \to \infty$ almost surely. A similar argument applies to the second term which proves to be $O(\delta)$. ■

These results can now be put together to give the desired convergence result.

Theorem 4.6

$$F_n'(\boldsymbol{\beta}^*) \underset{n\to\infty}{\overset{a.s.}{\to}} 0. \tag{4.4.24}$$

Proof. The idea is to write each component term Ω in (4.4.12) in the form

$$\Omega = \mathcal{E}\{\Omega\} + \left(\Omega - \mathcal{E}\{\Omega\}\right),$$

and then to appeal to the asymptotic convergence results established in the preceding lemmas. ■

Remark 4.4.2 *This result when combined with consistency suffices to establish the analogue of Theorem 4.5 in this case. The asymptotic convergence rate of the RGN algorithm can be expected to be similar to that of the full Gauss-Newton method. While the numerator expectation in the Gauss-Newton method is* 0*, and that in the RGN algorithm is* $O\left(\frac{1}{n}\right)$ *by Lemma 4.9, these are both smaller than the discrepancies* $\left(\Omega - \mathcal{E}\{\Omega\}\right)$ *between their full expressions and their expectations. Thus it is these discrepancy terms that are critical in determining the convergence rates. Here these correspond to law of large numbers rates for which a scale of* $O\left(n^{-1/2}\right)$ *is appropriate.*

Exercise 4.4.1 *Show that*

$$\frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \left(\mu_j \varepsilon_k + \mu_k \varepsilon_j\right)\left(\nabla_\beta P_{ij}\right)^T \nabla_\beta P_{ik} = O(\delta),\ n \to \infty.$$

4.5 The Kaufman modification

As the RGN algorithm possesses similar convergence rate properties to Gauss-Newton in large sample problems, and, as the Kaufman modification is favoured in implementation, it is of interest to ask if it too shares the same good large sample convergence rate properties . Fortunately the answer is in the affirmative. This result can be proved in the same way as the main lemmas in the previous section. This calculation is similar to the preceding and is considered after first exploring the close connection between the modified algorithm and the full $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ Gauss-Newton method. That both the Kaufman and Gauss-Newton methods can be implemented with the same amount of work is shown in [96].

First note that the linear least squares problem that determines the full Gauss-Newton correction is

$$\min_{\delta\boldsymbol{\alpha},\delta\boldsymbol{\beta}} \left\| \mathbf{b} - \Phi\boldsymbol{\alpha} - \left[\begin{array}{cc} \Phi & \nabla_\beta \left(\Phi\boldsymbol{\alpha}\right) \end{array} \right] \left[\begin{array}{c} \delta\boldsymbol{\alpha} \\ \delta\boldsymbol{\beta} \end{array} \right] \right\|^2 . \tag{4.5.1}$$

Introducing the variable projection matrix P permits this to be written:

$$\min_{\delta\boldsymbol{\beta}} \left\| P\mathbf{b} - P\nabla_\beta \left(\Phi\boldsymbol{\alpha}\right) \delta\boldsymbol{\beta} \right\|^2 + \min_{\delta\boldsymbol{\alpha}} \left\| \left(I - P\right) \left(\mathbf{b} - \nabla_\beta \left(\Phi\boldsymbol{\alpha}\right) \delta\boldsymbol{\beta}\right) - \Phi\left(\boldsymbol{\alpha} + \delta\boldsymbol{\alpha}\right) \right\|^2 , \tag{4.5.2}$$

where the minimization is to be performed first with respect to $\delta\boldsymbol{\beta}$ and then with respect to $\delta\boldsymbol{\alpha}$. Comparison with (4.4.4) shows that the first minimization is just

$$\min_{\delta\boldsymbol{\beta}} \left\| P\mathbf{b} - K\delta\boldsymbol{\beta} \right\| . \tag{4.5.3}$$

Thus, given $\boldsymbol{\alpha}$, the Kaufman search direction computed using (4.4.10) is exactly the Gauss-Newton correction for the nonlinear parameters . However, the two algorithms predict slightly different corrections $\delta\boldsymbol{\alpha}$. If $\boldsymbol{\alpha}$ is set using (3.5.3) then the second minimization gives

$$\begin{aligned} \delta\boldsymbol{\alpha} &= -\Phi^+ \nabla_\beta \left(\Phi\boldsymbol{\alpha}\right) \delta\boldsymbol{\beta}, \\ &= -\Phi^+ \nabla_\beta \Phi \left[\delta\boldsymbol{\beta}\right] \Phi^+ \mathbf{b}, \end{aligned} \tag{4.5.4}$$

while the increment in $\boldsymbol{\alpha}$ arising from the Kaufman correction is

$$\boldsymbol{\alpha}\left(\boldsymbol{\beta} + \delta\boldsymbol{\beta}\right) - \boldsymbol{\alpha}\left(\boldsymbol{\beta}\right) = \left(\nabla_\beta \Phi^+ \mathbf{b}\right) \delta\boldsymbol{\beta} + O\left(\left\|\delta\boldsymbol{\beta}\right\|^2\right) .$$

Note this increment is not computed as part of the algorithm. To examine

(4.5.4) in more detail we have

$$\begin{aligned}\frac{d\Phi^+}{dt} &= -\left(\Phi^T\Phi\right)^{-1}\left(\frac{d\Phi^T}{dt}\Phi + \Phi^T\frac{d\Phi}{dt}\right)\left(\Phi^T\Phi\right)^{-1}\Phi^T + \left(\Phi^T\Phi\right)^{-1}\frac{d\Phi^T}{dt},\\ &= -\left(\Phi^T\Phi\right)^{-1}\frac{d\Phi^T}{dt}T - \Phi^+\frac{d\Phi}{dt}\Phi^+ + \left(\Phi^T\Phi\right)^{-1}\frac{d\Phi^T}{dt},\\ &= \left(\Phi^T\Phi\right)^{-1}\frac{d\Phi^T}{dt}P - \Phi^+\frac{d\Phi}{dt}\Phi^+.\end{aligned}$$

The second term in this last equation occurs in (4.5.4). Thus, setting $\delta\boldsymbol{\beta} = \|\delta\boldsymbol{\beta}\|\,\mathbf{t}$,

$$\begin{aligned}\delta\boldsymbol{\alpha} - \left(\nabla_\beta\Phi^+\mathbf{b}\right)\delta\boldsymbol{\beta} &= -\|\delta\boldsymbol{\beta}\|\left(\Phi^T\Phi\right)^{-1}\frac{d\Phi^T}{dt}P\mathbf{b} + O\left(\|\delta\boldsymbol{\beta}\|^2\right),\\ &= \frac{\|\delta\boldsymbol{\beta}\|}{n}G^{-1}\left(\frac{d\Phi^T}{dt}P\left(\Phi(\boldsymbol{\beta}^*) - \Phi(\boldsymbol{\beta})\right)\boldsymbol{\alpha}^* - \Phi^T\frac{dP}{dt}\boldsymbol{\varepsilon}\right)\\ &\quad + O\left(\|\delta\boldsymbol{\beta}\|^2\right).\end{aligned}$$

The magnitude of this resulting expression can be shown to be small almost surely compared with $\|\delta\boldsymbol{\beta}\|$ when n is large enough using the law of large numbers and consistency as before. The proximity of the increments in the linear parameters plus the identity of the calculation of the nonlinear parameter increments demonstrates the close alignment between the Kaufman and Gauss-Newton algorithms. The small residual result is discussed in [96].

The variational matrix whose spectral radius evaluated at $\widehat{\boldsymbol{\beta}}_n$ determines the convergence rate of the Kaufman iteration is

$$\begin{aligned}F_n' &= I - \left(\frac{1}{n}K^TK\right)^{-1}\nabla_\beta^2 F,\\ &= -\left(\frac{1}{n}K^TK\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n s_i\nabla_\beta^2 s_i + \frac{1}{n}L^TL\right).\end{aligned} \tag{4.5.5}$$

It is possible to draw on work already done to establish the key convergence rate result that $\varpi\left\{F_n'\right\} \overset{a.s.}{\underset{n\to\infty}{\to}} 0$ also in this case. Lemmas 4.8 and 4.10 describe the convergence behaviour of $\mathcal{I}_n = \frac{1}{n}\left\{K^TK + L^TL\right\}$ as $n \to \infty$. Here it proves to be possible to separate out the properties of the individual terms by making use of the orthogonality of K and L once it has been shown that $\frac{1}{n}\mathcal{E}\left\{L\left(\boldsymbol{\beta}^*,\boldsymbol{\varepsilon}\right)^T L\left(\boldsymbol{\beta}^*,\boldsymbol{\varepsilon}\right)\right\} \overset{a.s.}{\underset{n\to\infty}{\to}} 0$. This calculation can proceed as follows.

Let $\mathbf{t} \in R^p$. Then

$$\begin{aligned}
\mathcal{E}\left\{\frac{1}{n}\mathbf{t}^T L^T L\mathbf{t}\right\} &= \frac{1}{n}\mathcal{E}\left\{\boldsymbol{\varepsilon}^T P\nabla_\beta\Phi\left[\mathbf{t}\right]\Phi^+\left(\Phi^+\right)^T\nabla_\beta\Phi\left[\mathbf{t}\right]^T P\boldsymbol{\varepsilon}\right\},\\
&= \frac{1}{n}\mathcal{E}\left\{\boldsymbol{\varepsilon}^T P\nabla_\beta\Phi\left[\mathbf{t}\right]\left(\Phi^T\Phi\right)^{-1}\nabla_\beta\Phi\left[\mathbf{t}\right]^T P\boldsymbol{\varepsilon}\right\},\\
&= \frac{1}{n^2}\,\mathrm{trace}\left\{\nabla_\beta\Phi\left[\mathbf{t}\right]G^{-1}\nabla_\beta\Phi\left[\mathbf{t}\right]^T P\mathcal{E}\left\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right\}P\right\} + \text{ smaller terms},\\
&= \frac{\sigma^2}{n^2}\,\mathrm{trace}\left\{\nabla_\beta\Phi\left[\mathbf{t}\right]G^{-1}\nabla_\beta\Phi\left[\mathbf{t}\right]^T\left(I-T\right)\right\} + \text{ smaller terms}.
\end{aligned}$$

This last expression breaks into two terms, one involving the unit matrix and the other involving the projection T. Both lead to terms of the same order. The unit matrix term gives

$$\mathrm{trace}\left\{\nabla_\beta\Phi\left[\mathbf{t}\right]G^{-1}\nabla_\beta\Phi\left[\mathbf{t}\right]^T\right\} = \mathbf{t}^T\left\{\sum_{i=1}^n \Psi_i G^{-1}\Psi_i^T\right\}\mathbf{t},$$

where

$$\left(\Psi_i\right)_{jk} = \frac{\partial\phi_{ij}}{\partial\beta_k},\ \Psi_i : R^m \to R^p.$$

It follows that

$$\frac{\sigma^2}{n^2}\sum_{i=1}^n \Psi_i G^{-1}\Psi_i^T = O\left(\frac{1}{n}\right),\ n\to\infty.$$

To complete the story note that the conclusion of Lemma 4.10 can be written

$$\frac{1}{n}\left(K^TK + L^TL\right) \underset{n\to\infty}{\overset{a.s.}{\to}} \mathcal{E}\left\{\frac{1}{n}K^TK + \frac{1}{n}L^TL\right\}.$$

If $\frac{1}{n}K^TK$ is bounded, positive definite then, using the orthogonality (4.4.9),

$$\frac{1}{n}K^TK\left(\frac{1}{n}K^TK - \mathcal{E}\left\{\frac{1}{n}K^TK\right\}\right) \underset{n\to\infty}{\overset{a.s.}{\to}} \frac{1}{n}K^TK\mathcal{E}\left\{\frac{1}{n}L^TL\right\}.$$

This shows that $\frac{1}{n}K^TK$ tends almost surely to its expectation provided it is bounded, positive definite for n large enough and so can be cancelled on both sides in the above expression. Note first that the linear parameters cannot upset boundedness.

$$\begin{aligned}
\boldsymbol{\alpha}\left(\boldsymbol{\beta}\right) &= \left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{b},\\
&= \frac{1}{n}\left(G^{-1} + O\left(\frac{1}{n}\right)\right)\Phi^T\left(\Phi\boldsymbol{\alpha}^* + \left(\Phi^* - \Phi\right)\boldsymbol{\alpha}^* + \boldsymbol{\varepsilon}\right),\\
&= \boldsymbol{\alpha}^* + \frac{1}{n}\left(G^{-1} + O\left(\frac{1}{n}\right)\right)\Phi^T\left(\left(\Phi^* - \Phi\right)\boldsymbol{\alpha}^* + \boldsymbol{\varepsilon}\right),\\
&= \boldsymbol{\alpha}^* + \boldsymbol{\delta},\ \left\|\boldsymbol{\delta}\right\|_\infty = o\left(1\right), \qquad (4.5.6)
\end{aligned}$$

where $\boldsymbol{\alpha}^*$ is the true vector of linear parameters and $\Phi^* = \Phi(\boldsymbol{\beta}^*)$. Positive definiteness follows from

$$\begin{aligned}\mathbf{t}K^TK\mathbf{t} &= \boldsymbol{\alpha}(\boldsymbol{\beta})^T \frac{d\Phi}{dt}^T P\frac{d\Phi}{dt}\boldsymbol{\alpha}(\boldsymbol{\beta}), \\ &= \left\|\frac{d\Phi}{dt}\boldsymbol{\alpha}(\boldsymbol{\beta})\right\|^2 - \left\|T\frac{d\Phi}{dt}\boldsymbol{\alpha}(\boldsymbol{\beta})\right\|^2 \geq 0.\end{aligned}$$

Equality can hold only if there is $\mathbf{t}$ such that $\frac{d\Phi}{dt}\boldsymbol{\alpha}(\boldsymbol{\beta}) = \gamma\Phi\boldsymbol{\alpha}(\boldsymbol{\beta})$. This condition was met also in Lemma 4.8.

Exercise 4.5.1 *Show*

$$\frac{\sigma^2}{n^2}\operatorname{trace}\left\{\nabla_\beta\Phi[\mathbf{t}]\,G^{-1}\nabla_\beta\Phi[\mathbf{t}]^T\,T\right\} \to 0,\ n\to\infty.$$

4.6 Numerical examples

Scoring extends the Gauss-Newton algorithm to log likelihoods based on distributions other than normal, both continuous and discrete. A range of such possibilities is considered in this section.

4.6.1 Simple exponential model

Here the model used is

$$\mu(t,\mathbf{x}) = x(1) + x(2)\exp(-x(3)t). \tag{4.6.1}$$

The values chosen for the parameters are $x(1) = 1$, $x(2) = 5$, and $x(3) = 10$. This should give a well determined model with all values contributing significantly to major features of the graph of $\mu(t,\mathbf{x})$ on the interval $[0,1]$. Thus numerical results should provide insight into the complicating effects of random perturbations. Initial values are generated using

$$x(i)_0 = x(i) + (1 + x(i))(.5 - Rnd)$$

where Rnd indicates a call to a uniform random number generator giving values in $[0,1]$. Two types of random numbers are used to simulate the experimental data.

- The data is generated by evaluating $\mu(t,\mathbf{x})$ on a uniform grid with spacing $\Delta = 1/(n+1)$ and then perturbing these values using normally distributed random numbers to give values

$$b_i = \mu(i\Delta,\mathbf{x}) + \varepsilon_i, \quad \varepsilon_i \sim N(0,2), \quad i = 1,2,\cdots,n.$$

The log likelihood is taken as

$$\mathcal{L}(\mathbf{x}) = -\frac{1}{2}\sum_{i=1}^{n}\left(b_i - \mu(i\Delta, \mathbf{x})\right)^2 .$$

Although the scale does not appear here it is significant in the random number generation where it is important in controlling the estimation problem difficulty. Easy problems correspond to small σ with Newton's method having second order convergence in the limiting case $\sigma = 0$. Here the choice of standard deviation was made so that small sample problems ($n = 32$) are relatively difficult. Also the algorithms used explicitly scale the columns of the design matrix in the normal case while for distributions without the auxiliary scale parameter the scaling matrix is set to the unit matrix.

- A Poisson random number generator is used to generate random counts z_i corresponding to $\mu(i\Delta, \mathbf{x})$ as the mean model. Here

$$P((X = i) = \frac{e^{-\lambda}\lambda^i}{i!} = \frac{1}{i!}e^{-\lambda}e^{i\log\lambda}$$

so that the distribution is a member of the exponential family with $\theta = \log\lambda$, $b(\theta) = e^{\theta}$, and $\mu = \frac{db}{d\theta} = \lambda$. The log likelihood used is

$$\mathcal{L}(\mathbf{x}) = \sum_{i=1}^{n} z_i \log\left(\frac{\mu(i\Delta, \mathbf{x})}{z_i}\right) + \left(z_i - \mu(i\Delta, \mathbf{x})\right).$$

Here constant terms $-z_i\log(z_i) + z_i$, corresponding to the so called saturated model, have been added to the log likelihood. Also, note that if $z_i = 0$ then the contribution from the logarithm term to the log likelihood is taken as zero. The rows of the least squares problem design matrix (4.2.19) are given by

$$\mathbf{e}_i^T A = \frac{1}{s_i}, \frac{\exp(-x(3)t_i)}{s_i}, \frac{-x(2)t_i\exp(-x(3)t_i)}{s_i}, \quad i = 1, 2, \cdots, n,$$

where $s_i = \sqrt{\mu(i\Delta, \mathbf{x})}$. The corresponding components of the right hand side are

$$b_i = \frac{z_i - \mu(i\Delta, \mathbf{x})}{s_i}.$$

Numerical experiments comparing the performance of the line search (LS) and trust region (TR) methods are summarised in Table 4.2, and the final

	Normal		Poisson		
n	LS	TR	LS	TR	S
32	12.1*	12*	10.8	12.3	21
128	11.7	11.9	7.6	7.9	16
512	8.4	7.3	7.1	6.9	13
2048	6.7	6.1	6.3	5.8	9

Table 4.2: Algorithm performance, mean of 10 runs

column gives results for the linesearch version of the sample method. For each n the computations were initiated with 10 different seeds for the basic random number generator, and the average number of iterations is reported as a guide to algorithm performance. The parameter settings used are $\alpha = 2.5,\ \beta = .1$ for the trust region method and $\rho = .25$ for the simple parameter used in the line search (4.2.26). Experimenting with these values (for example, the choice $\alpha = 1.5,\ \beta = .5$) made very little difference in the trust region results. The value returned by the local quadratic fit (4.2.25) used in the line search computations was favoured over the simple parameter about half the time in the relatively few cases when step reduction proved necessary. However, this observation must be qualified. An important cause of step reduction corresponded to cases in which the initial unit step over-corrected the rate constant in the exponential term turning it into an exponentially increasing one. Such a step could not be profitable in maximizing the likelihood and it was especially in these cases that the quadratic interpolation proved less satisfactory. However, it is straight forward to guard against this problem for the simple situation here by reducing the initial step. If this is done then the quadratic interpolation procedure becomes much more competitive. Convergence is assumed if $\nabla_x \mathcal{L}\mathbf{h} < 1.0e^{-8}$. This corresponds to final values of $\|\mathbf{h}\|$ in the range $1.e^{-4}$ to $1.e^{-6}$ while estimates of ϖ between .5 and $1.e^{-2}$ characterize the difference between the slower and faster convergence rates observed. In this range the estimates of ϖ prove stable enough. In almost all cases the values produced by the TR and LS procedures shadow each other closely, more closely than the table entries indicate because of slight differences in the program logic. Also, as expected, the problems become easier as n increases with the consequence that the initial step is almost always accepted for the larger values of n in the normal distribution problems. For the Poisson distribution it is always accepted indicating that initial values with up to 50% relative error componentwise are satisfactory here. The sample algorithm applied to the Poisson data behaves in a similar fashion but takes noticeably more iterations for each n.

x(1)	x(2)	x(3)
-51.198	53.911	3.6269-02
-51.537	54.250	3.6039-02
-51.880	54.593	3.5808-02
-52.228	54.940	3.5578-02
-52.579	55.292	3.5348-02

Table 4.3: Iterations to a non-closure point

The starred entries in table 4.2 indicate two cases of nonconvergence , the same cases causing trouble for both linesearch and trust region algorithms. The phenomenon observed is illustrated in table 4.3. Here the entries are consecutive values of $\mathbf{x}$ from iteration 246 to iteration 250 for the linesearch algorithm for one of the problems but behaviour both of the trust region algorithm and of both algorithms on the other problem are similar. They should be compared with the limiting process in Remark 4.2.3. Clearly the entries in table 4.3 are consistent with an instance of the failure of the approximating set to be closed. Here this means that the perturbation of the model signal with random normal error leads to a data set that is better fitted by a straight line than by the actual exponential model used to generate the data corresponding to the unboundedness of the set of approximations. This point is reinforced in Figure 4.1 which shows the resulting fit (red curve) after 50 iterations of the line search algorithm. The green curve gives a plot of the data, while the blue curve is the fit given by the initial conditions. It has been argued that variable projection would be more satisfactory in situations like this [38]. Here it leads to the solution of a single nonlinear equation, but this solution would still have to be interpreted. In the case of non-compactness in more general separable models there is the likelihood of the coalescing of terms involving the nonlinear parameters with consequent singularity problems in the variable projection formulation. Here this closure problem goes away with more observations, but this means that $n = 32$ corresponds to a dangerously small data set. This problem does not occur in the experiments based on the Poisson distribution, a case of a discrete distribution, and this observation may have wider validity.

4.6.2 Gaussian peaks plus exponential background

The conventional orthodoxy is that well separated gaussian peaks are easy to resolve while close peaks tend to merge into one composite. Extracting information then requires special techniques. One such tool is numerical differentiation of the data. This has the potential to exaggerate changes in

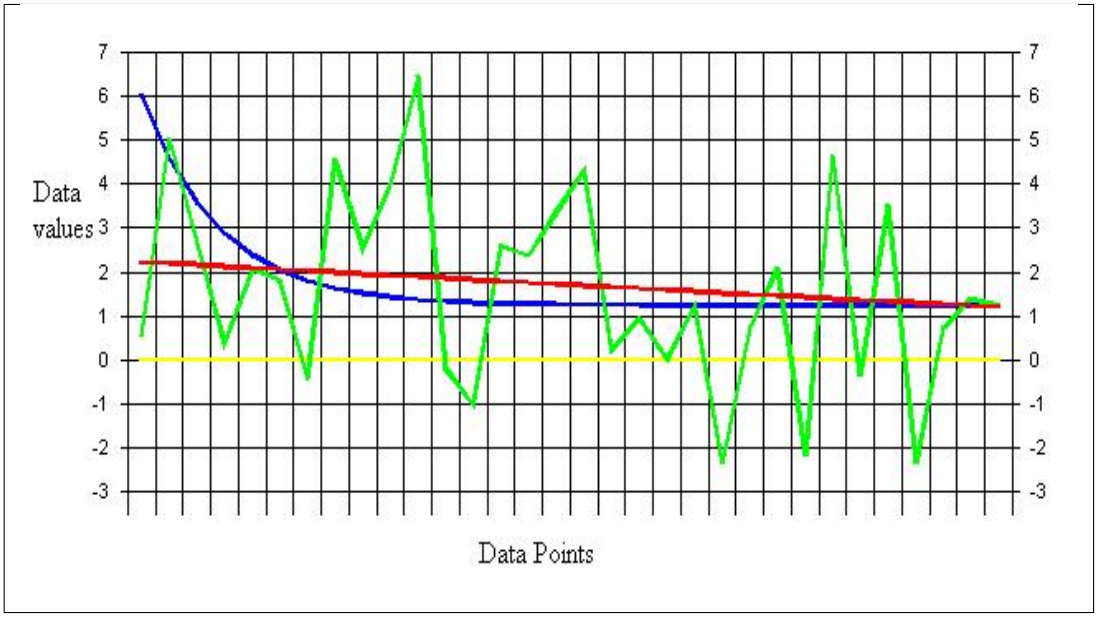

Figure 4.1: Result shows a straight line fit

slope and curvature, but discussion of its properties is beyond the present scope. As in the previous example, difficulties occur as consequence of the effects of random perturbations to the prescribed signal and poorly chosen initial parameter guesses. The model has the form

$$\mu(\mathbf{x}, t) = x(1)e^{-x(2)t} + x(3)e^{-\frac{(t-x(4))^2}{x(5)}} + x(6)e^{-\frac{(t-x(7))^2}{x(8)}}.$$

where typical values used were $x(1) = 5.$, $x(2) = 10.$ $x(3) = 18.$, $x(4) = .3333$, $x(5) = .01$, $x(6) = 15.$, $x(7) = .6667$, $x(8) = .01$. These values are used to evaluate the simulated data which is then perturbed using random numbers drawn from a normal distribution with standard deviation 1.0. Initial values were given by

$$x(i) = x^*(i) + .5 * x^*(i) * RND$$

where RND is the standard uniform distribution. Note that this biases the initial conditions to the high side of the correct values. Numerical results are patchy for both the line search and trust region algorithms. For the satisfactory results the behaviour is basically similar to that observed for the exponential fitting problem. The reason for the problem results can be seen in Figure 4.2 which shows the computed approximation after 50 iterations for one case corresponding to a data set with 128 points. Here the effect of the positive bias in the initial conditions results in the starting values missing

n	$\sigma = 1$	$\sigma = 2$	$\sigma = 4$
64	7	16	nc
256	11	21	50
1024	7	17	18
4096	6	6	7
16384	6	6	7

Table 4.4: Iteration counts for peak fitting with exponential background

the second peak. While the iteration picks up the exponential background and the first peak it does no more than fit to the noise for larger values of t. The estimate of $x(7)$ is actually increasing in this case, and this is further removing the second peak from consideration. The result is that the corresponding columns of the design are getting exponentially small, and the iteration actually fails eventually with an error in the scaling step. Typically, with well spaced peaks in the data such as is the case here, good estimates of both peak location and peak width would be available so this particular phenomenon would not be observed. This is illustrated in Table 4.4. In these calculations

$$\mu(\mathbf{x}, t) = 5e^{-10t} + 18e^{-\frac{(t-.25)^2}{.015}} + 15e^{-\frac{(t-.5)^2}{.03}} + 10e^{-\frac{(t-.75)^2}{.015}}.$$

Initial conditions are chosen such that there are random errors of up to 50% in the background parameters and peak heights, 12.5% in peak locations, and 25% in peak width parameters. Numbers of iterations are reported for $\sigma = 1,\ 2,\ 4$ and $n = 64,\ 256,\ 1024,\ 4096,\ 16384$. The most sensitive parameters prove to be those determining the exponential background, and they trigger the lack of convergence when $\sigma = 4,\ n = 64$. The apparent superior convergence behaviour in the $n = 64$ case over the $n = 256$ case for the smaller σ values can be explained by the sequence of random numbers generated producing smaller residuals. The sequence used here corresponds to the first quarter of the sequence for $n = 256$. Plots for the fits obtained for $\sigma = 4,\ n = 64$ and $\sigma = 4,\ n = 256$ are given in Figure 4.3 and Figure 4.4 respectively. The difficulty with the background estimation in the former shows up in the sharp kink in the fitted (red) curve near $t = 0$. This figure gives the result after 50 iterations when $x(1) = 269$ and $x(2) = 327$. The green curve gives the fit obtained using the initial values. This manages to hide the middle peak fairly well.

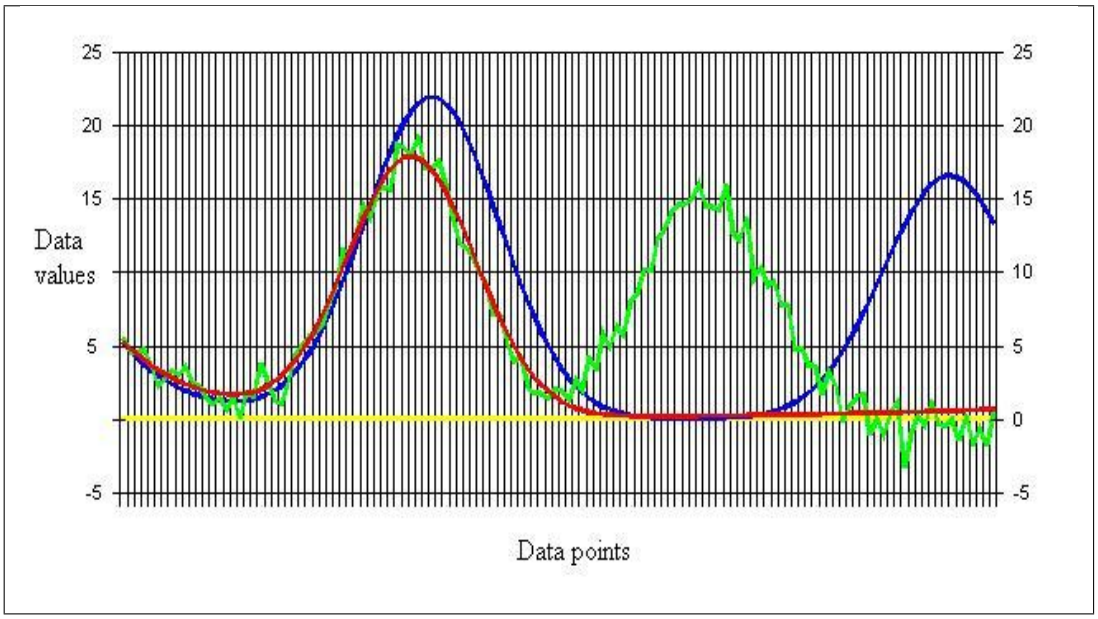

Figure 4.2: Initial conditions miss the second peak

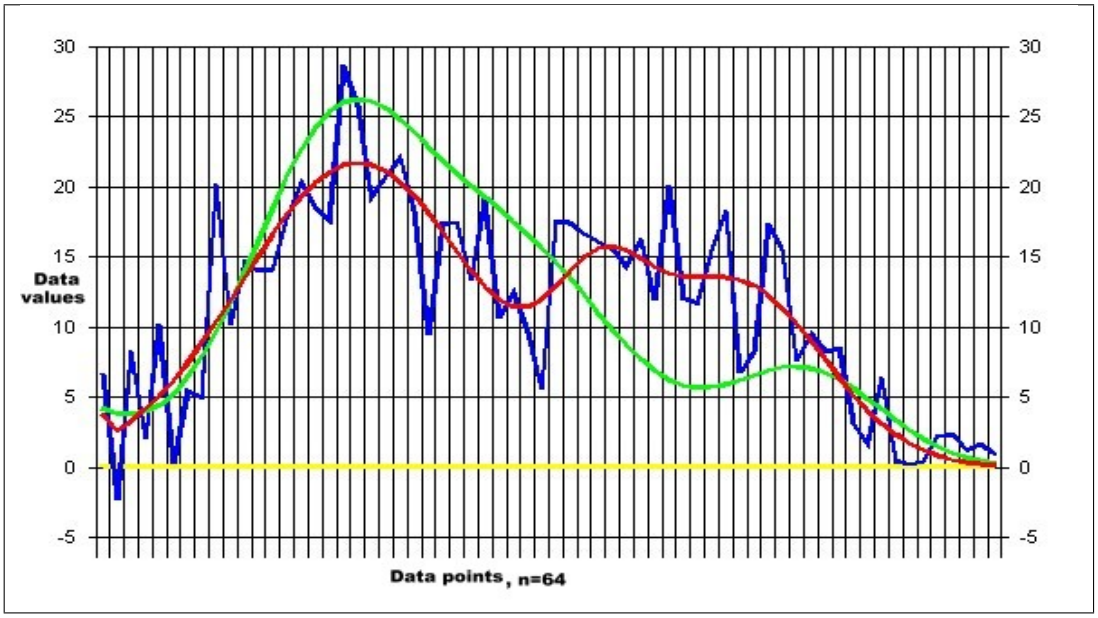

Figure 4.3: No convergence: fit after 50 iterations case $\sigma = 4,\ n = 64$

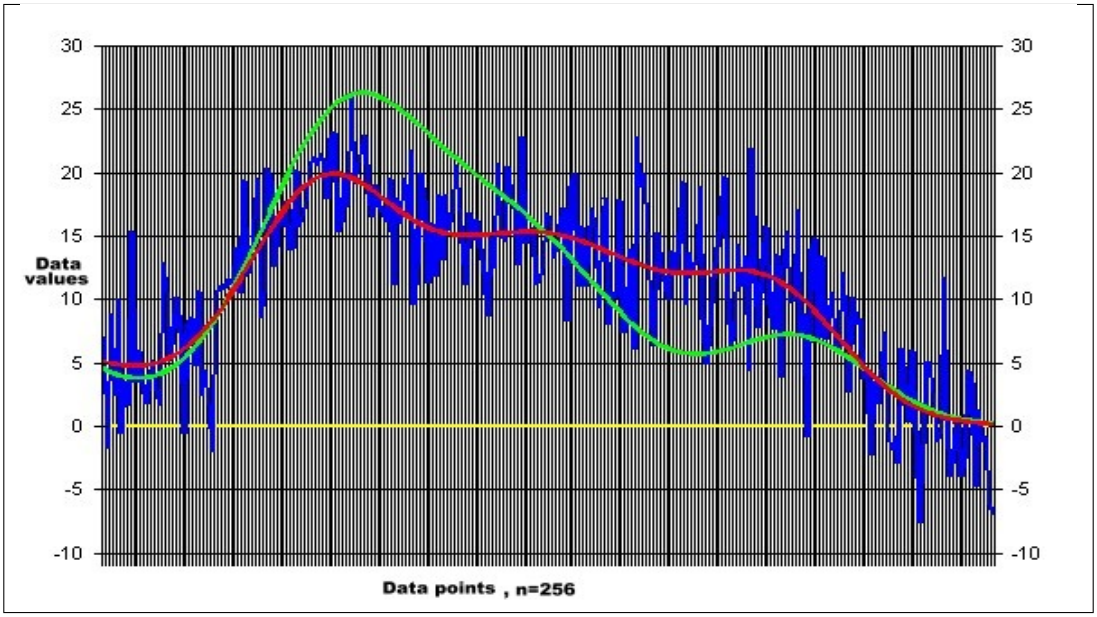

Figure 4.4: Fit obtained: case $\sigma = 4,\ n = 256$

$\log_{10}(titre)$	dead	normal	deformed
-0.42	0	18	0
0.58	1	13	2
1.58	5	4	6
2.58	12	1	6
3.58	18	0	1
4.58	16	0	0

Table 4.5: Cattle virus data

4.6.3 A multinomial example

Data for a trinomial example ($m = 3$ in (3.1.6)) is given in Table 4.5 [77]. It is derived from a study of the effects of a cattle virus on chicken embryos. In this case the log likelihood up to constant terms is

$$L_t = \sum_{j=1}^{m} y_j(t) \log\left(\pi_j(t)\right),$$

where the $y_j(t)$ are the observed counts, and the $\pi_j(t)$ are the frequencies to be modelled. A feature here is that each observation yields a vector of data.

The model fitted to this data is:

$$\begin{aligned}\pi_1 &= \frac{1}{1+\exp\left(-\beta_1-\beta_3\log\left(t\right)\right)},\\ 1-\pi_2 &= \frac{1}{1+\exp\left(-\beta_2-\beta_3\log\left(t\right)\right)},\\ \pi_3 &= 1-\pi_1-\pi_2.\end{aligned}$$

Differentiating gives

$$\begin{aligned}\frac{\partial L_t}{\partial \pi_i} &= \frac{y_i\left(t\right)}{\pi_i\left(t\right)}-\frac{y_m\left(t\right)}{\pi_m\left(t\right)}, i=1,2,\cdots,m-1,\\ \frac{\partial^2 L_t}{\partial \pi_i \partial \pi_j} &= -\frac{y_i\left(t\right)}{\pi_i\left(t\right)^2}\delta_{ij}-\frac{y_m\left(t\right)}{\pi_m\left(t\right)^2}, i,j=1,2,\cdots,m-1.\end{aligned}$$

Means and variances are computed most easily from the exponential family form for the distribution ((3.1.7)-(3.1.9)). This gives:

$$\begin{aligned}\mathcal{E}\left\{y_i\right\} &= n\pi_i,\\ \mathcal{E}\left\{\frac{\partial^2 L_t}{\partial \pi_i \partial \pi_j}\right\} &= -\frac{n\left(t\right)}{\pi_i\left(t\right)}\delta_{ij}-\frac{n\left(t\right)}{\pi_m\left(t\right)}, i,j=1,2,\cdots,m-1.\end{aligned}$$

Thus V_t^{-1} in (4.2.15) is given by

$$\begin{aligned}V_t^{-1} &= n(t)\left\{\operatorname{diag}\left(\frac{1}{\pi_i\left(t\right)}\right)+\frac{1}{\pi_m\left(t\right)}\mathbf{e}\mathbf{e}^T\right\},\\ &= n\left(t\right)D_t\left(I+\frac{1}{\pi_m\left(t\right)}\mathbf{v}_t\mathbf{v}_t^T\right)D_t,\end{aligned}$$

where

$$D_t = \operatorname{diag}\left(\frac{1}{\sqrt{\pi_i(t)}}; i=1,2,\cdots,m-i\right), \quad \mathbf{v}_t = D_t^{-1}\mathbf{e}.$$

An appropriate form to use for $V_t^{-1/2}$ in evaluating (4.2.16) is

$$V_t^{-1/2} = \sqrt{n(t)}\left(I+\rho_t\mathbf{v}_t\mathbf{v}_t^T\right)D_t,$$

where

$$\rho(t) = \frac{1}{\pi_m(t)+\sqrt{\pi_m(t)}}.$$

iteration	$\mathcal{L}$	$\nabla\mathcal{L}\mathbf{h}$	β_1	β_2	β_3
0	-54.86		-4.597	-3.145	.7405
1	-47.70	.1401+2	-3.737	-2.200	.7555
2	-47.01	.1277+1	-4.373	-2.551	.8803
3	-46.99	.3829-1	-4.403	-2.618	.9056
4	-46.99	.1234-5	-4.505	-2.619	.9061
5	-46.99	.3085-8	-4.405	-2.619	.9061

Table 4.6: Results of computations for the trinomial data

The corresponding form for $\left(V_t^{1/2}\right)^T$ which is used in evaluating (4.2.18) is

$$\left(V_t^{1/2}\right)^T = \frac{1}{\sqrt{n(t)}}\left(I - \sqrt{\pi_m(t)}\rho(t)\mathbf{v}_t\mathbf{v}_t^T\right)D_t^{-1}$$

The contributions to the least squares form of the linear subproblem (4.2.19) can now be evaluated. This gives for the contribution from the t'th observation

$$\begin{aligned}
(I_L^n)_t &= \sqrt{n(t)}\left(I + \rho(t)\mathbf{v}_t\mathbf{v}_t^T\right)D_t\nabla_\beta\boldsymbol{\pi}_t, \\
&= \sqrt{n(t)}\left\{D_t\nabla_\beta\boldsymbol{\pi}_t - \rho(t)\mathbf{v}_t\nabla_\beta\pi_m(t)\right\}, \qquad (4.6.2) \\
\mathbf{b}_t &= \frac{1}{\sqrt{n(t)}}\left(I - \sqrt{\pi_m(t)}\rho(t)\mathbf{v}_t\mathbf{v}_t^T\right)D_t^{-1}\left\{D_t^2\mathbf{y}_t - \frac{y_m(t)}{\pi_m(t)}\mathbf{e}\right\}, \\
&= \frac{1}{\sqrt{n(t)}}\left\{D_t\mathbf{y}_t - \left(\frac{\mathbf{e}^T\mathbf{y}_t}{1+\sqrt{\pi_m(t)}} + \frac{y_m(t)}{\sqrt{\pi_m(t)}}\right)\mathbf{v}_t\right\}, \\
&= \frac{1}{\sqrt{n(t)}}\left\{D_t\mathbf{y}_t - \rho(t)\left(\sqrt{\pi_m(t)}n(t) + y_m(t)\right)\mathbf{v}_t\right\}, \qquad (4.6.3)
\end{aligned}$$

where $\mathbf{y}(t)$ is the vector with components $y_1(t), y_2(t), \cdots, y_{m-1}(t)$. Note that the components of $\mathbf{b}_t$ have a scale of $\sqrt{n(t)\pi_i(t)} = \sqrt{\mathcal{E}\left\{y_i(t)\right\}}$ while the untransformed quantities $\frac{\partial L_t}{\partial \pi_i}$ have a corresponding scale of n.

Numerical results are given in Table 4.6. An interesting feature is the very satisfactory rate of convergence despite the fact that the data set could hardly be described as large. This is another example of good convergence behaviour being observed in a problem with a discrete probability model. However, the sample algorithm is not effective in this case, and the most likely explanation would seem to be the size of the data set.

Exercise 4.6.1 *Verify the derivations of equations (4.6.2) and (4.6.3).*

4.7 Constrained likelihood problems

The equality constrained problem presents a new source of difficulty because now steps taken must not only contribute to maximizing the likelihood but also move closer to the constraints. Thus general methods for these problems tend to focus sequentially on these aspects in each overall iteration step. This would seem to favour methods within a trust region framework because then adjustments are carried out within a region which controls the validity of local approximations. Line search algorithms, in contrast, put this requirement onto the choice of monitor function. In many applications involving equality constraints the basic algorithmic approach is based on applying Newton's method to the necessary conditions. To avoid the calculation of second derivatives scoring-like simplifications are often made with reported success. One example is provided by the simultaneous approach which is considered in the next chapter.

Two methods are considered here. The first is the Powell-Hestenes form of the augmented Lagrangian method . This form is attractive both because of its conceptual simplicity and because it suggests a convenient extension of the scoring algorithm. It has an important disadvantage in terms of cost because it has an inner-outer iteration structure in which each outer iteration involves a correction step for the Lagrange multiplier estimates while each inner iteration requires a full unconstrained minimization computed by the scoring algorithm. This latter need not be particularly effective in the initial minimizations even if the conditions for fast convergence are satisfied eventually. An alternative is provided by a sequential quadratic programming algorithm (SQP) which potentially is made more efficient by avoiding expensive inner iterations, but which is also significantly more complicated.

4.7.1 The Powell-Hestenes method

One method which suggests that it should lend itself to the effective use of nonlinear least squares techniques in an unconstrained minimization as a starting point is the Powell-Hestenes method (also known as the augmented Lagrangian method in a slightly modified guise) [31]. This method separates each iteration into a minimization step which considers the objective function

$$\mathcal{P}_n\left(\mathbf{x}, \gamma, \boldsymbol{\theta}\right) = -\frac{1}{n}\mathcal{L}\left(\mathbf{y}; \mathbf{x}, \mathbf{T}_n\right) + \gamma \sum_{i=1}^{m} \left(c_i\left(\mathbf{x}\right) + \theta_i\right)^2,$$

followed by an adjustment step which notes that at a minimum

$$\nabla_{\mathbf{x}}\mathcal{P}_n = -\frac{1}{n}\nabla_{\mathbf{x}}\mathcal{L} + 2\gamma\sum_{i=1}^{m}\left(c_i+\theta_i\right)\nabla_{\mathbf{x}}c_i = 0$$

so that the current $\mathbf{x}$ solves a perturbed constrained problem with multipliers $2\gamma\left(c_i+\theta_i\right)$. Here the idea is to adjust $\boldsymbol{\theta}$ to force constraint satisfaction, while γ, which has the role of a penalty parameter, controls the rate of convergence. The appealing feature of this method is its simplicity because the update step

$$\theta \leftarrow \mathbf{c}\left(\mathbf{x}\left(\boldsymbol{\theta}\right)\right) + \theta;\, \gamma \leftarrow \gamma \tag{4.7.1}$$

proves distinctly effective. This equation is derived in the Appendix under the assumption that γ is large enough. However, if the rate of convergence appears slow then this assumption on γ is challenged. In this case the adjustment step is be replaced by

$$\theta \leftarrow \theta/\alpha;\, \gamma \leftarrow \alpha\gamma,\ \alpha > 1. \tag{4.7.2}$$

Remark 4.7.1 *If this second step (4.7.2) is applied repeatedly then the result is essentially a penalty method for the equality constrained problem. This gives guaranteed convergence with error tending to* 0 *like* $O\left(1/\alpha^k\right)$ *where* k *is the number of minimization/update steps. On the other hand, if the repeated application of the first step (4.7.1) is successful then the convergence rate is given by (4.8.3) and is first order with multiplier* $O\left(1/\gamma\right)$*. An important distinction between this approaches and that based on SQP notes that here the condition number of* $\nabla_{\mathbf{x}}^2\mathcal{P}_n$ *is* $O\left(\gamma\right)$ *. This follows because the norm of* $\nabla_{\mathbf{x}}^2\mathcal{P}_n$ *is* $O\left(\gamma\right)$ *as a consequence of the* γ *dependence in (4.7.3) below, while the norm of* $\nabla_{\mathbf{x}}^2\mathcal{P}_n^{-1}$ *is* $O\left(1\right)$*. This estimate is given in equation (4.8.1) in the appendix. This must impact the minimization calculation. It is bounded if (4.7.1) is successful but tends to* ∞ *like* α^k *if the pure penalty approach (4.7.2) is used.*

A variant of scoring can provide an effective algorithm for minimizing $\mathcal{P}_n$. Starting with the Hessian

$$\nabla_{\mathbf{x}}^2\mathcal{P}_n = -\frac{1}{n}\nabla_{\mathbf{x}}^2\mathcal{L} + 2\gamma\sum_{i=1}^{m}\left(\nabla_{\mathbf{x}}^T c_i\nabla_{\mathbf{x}}c_i + \left(c_i+\theta_i\right)\nabla_{\mathbf{x}}^2 c_i\right), \tag{4.7.3}$$

it proves convenient not only to replace $\nabla_{\mathbf{x}}^2\mathcal{L}$ by its expectation, but also to ignore the second derivative terms $\nabla_{\mathbf{x}}^2 c_i,\ i = 1, 2, \cdots, m$. This leads to the calculation of the correction term

$$\mathbf{h}_P = -H_n^{-1}\nabla_{\mathbf{x}}\mathcal{P}_n^T$$

where

$$H_n = \mathcal{I}_n + 2\gamma C^T C,\ C = \nabla_{\mathbf{x}}\mathbf{c}.$$

Here $\mathcal{P}_n$ provides a suitable monitor as H_n is generically positive definite so that

$$\nabla_{\mathbf{x}}\mathcal{P}_n\mathbf{h}_P = -\nabla_{\mathbf{x}}\mathcal{P}_n H_n^{-1}\nabla_{\mathbf{x}}\mathcal{P}_n^T < 0.$$

The associated fixed point iteration is

$$\mathbf{x} \leftarrow \mathbf{x} - H_n^{-1}\nabla_{\mathbf{x}}\mathcal{P}_n^T.$$

The ultimate rate of convergence at step k depends on the variational matrix evaluated at $\mathbf{x}_P^k$, the minimum of the current iteration step. This is

$$\begin{aligned} I - H_n^{-1}\nabla_{\mathbf{x}}^2\mathcal{P}_n &= H_n^{-1}\left(H_n - \nabla_{\mathbf{x}}^2\mathcal{P}_n\right), \\ &= H_n^{-1}\left(\mathcal{J}_n + \mathcal{I}_n - 2\gamma\sum_{i=1}^{m}\left(c_i + \theta_i\right)\nabla_{\mathbf{x}}^2 c_i\right). \end{aligned}$$

If the constraints are linear then $\nabla_{\mathbf{x}}^2 c_i = 0,\ i = 1, 2, \cdots, m$, so that the spectral radius ϖ of the term $\mathcal{J}_n + \mathcal{I}_n$ determines the ultimate rate of convergence. Here $\varpi = O\left(\left\|\mathbf{x}^* - \mathbf{x}_P^k\right\|\right) + o(1)$ for n large enough as a consequence of consistency and the law of large numbers by the argument used in the unconstrained case. Under these conditions the scoring based minimization procedure will be effective eventually. However, if the constraints are nonlinear then $-\gamma\left(c_i + \theta_i\right)$ tends to the corresponding Lagrange multiplier and the term in $\nabla_{\mathbf{x}}^2 c_i$ cannot be so easily ignored. In the following example, in which the initial parameter choices are $\gamma = \sqrt{n}$, $\boldsymbol{\theta} = 0$, evidence is developed that the scoring algorithm can behave in very much the same fashion as in the unconstrained likelihood case.

Example 4.7.1 *Consider the mixture density*

$$\begin{aligned} f_R(y|\mu_1, \mu_2, \sigma_1, \sigma_2) = &\frac{1}{\sqrt{2\pi}\sigma_1}\frac{\mu_1}{\mu_1 + \mu_2}\exp -\frac{(y-\mu_1)^2}{2\sigma_1^2} \\ &+ \frac{1}{\sqrt{2\pi}\sigma_2}\frac{\mu_2}{\mu_1 + \mu_2}\exp -\frac{(y-\mu_2)^2}{2\sigma_2^2}. \quad (4.7.4) \end{aligned}$$

Random numbers generated according to a realisation of this density can be considered also to be generated according to the density

$$f(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma_1}x_1\exp -\frac{(y-\mu_1)^2}{2\sigma_1^2} + \frac{1}{\sqrt{2\pi}\sigma_2}x_2\exp -\frac{(y-\mu_2)^2}{2\sigma_2^2},$$

where

$$\mathbf{x}^T = [x_1, \mu_1, \sigma_1, x_2, \mu_2, \sigma_2],$$

subject to the constraints

$$c_1(\mathbf{x}) = x_1 - \frac{\mu_1}{\mu_1 + \mu_2} = 0,$$
$$c_2(\mathbf{x}) = x_2 - \frac{\mu_2}{\mu_1 + \mu_2} = 0.$$

Thus it should be possible to recover $\mu_1, \mu_2, \sigma_1, \sigma_2$ from data generated according to f_R by considering the likelihood defined by f subject to the above constraints. Let

$$e_i(y) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp -\frac{(y-\mu_i)^2}{2\sigma_i^2}, \quad i = 1, 2.$$

Then

$$f(y|\mathbf{x}) = x_1 e_1(y) + x_2 e_2(y).$$

We have:

$$\mathcal{L}_n(\mathbf{x}) = \sum_{i=1}^{n} \log f(y_i|\mathbf{x}),$$
$$\nabla_{\mathbf{x}} \mathcal{L}_n(\mathbf{x}) = \sum_{i=1}^{n} \frac{1}{f(y_i|\mathbf{x})} \mathbf{v}_i^T, \tag{4.7.5}$$

where

$$\mathbf{v}_i^T = [\left(1, x_1 \frac{y_i - \mu_1}{\sigma_1^2}, x_1 (\frac{(y_i - \mu_1)^2}{\sigma_1^3} - \frac{1}{\sigma_1})\right) e_1,$$
$$\left(1, x_2 \frac{y_i - \mu_2}{\sigma_2^2}, x_2 (\frac{(y_i - \mu_2)^2}{\sigma_2^3} - \frac{1}{\sigma_2})\right) e_2]. \tag{4.7.6}$$

This is an example where the use of the sample information (4.1.4) proves convenient. Here this gives

$$\mathcal{S}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{f(y_i|\mathbf{a})^2} \mathbf{v}_i \mathbf{v}_i^T. \tag{4.7.7}$$

It follows from (4.2.11), (4.7.5), and (4.7.7) that the scoring method using the sample information gives a set of equations for the correction $\mathbf{h}_S$ which

can be written as the linear least squares problem:

$$\min_{\mathbf{h}} \mathbf{r}^T\mathbf{r};$$

$$\mathbf{r} = \begin{bmatrix} \sqrt{2\gamma}\nabla_{\mathbf{x}} c_1 \\ \sqrt{2\gamma}\nabla_{\mathbf{x}} c_2 \\ \hline \mathbf{v}_1^T/(\sqrt{n} f(y_1|\mathbf{x})) \\ \ldots \\ \mathbf{v}_i^T/(\sqrt{n} f(y_i|\mathbf{x})) \\ \ldots \\ \mathbf{v}_n^T/(\sqrt{n} f(y_n|\mathbf{x})) \end{bmatrix} \mathbf{h}_S + \begin{bmatrix} \sqrt{2\gamma}(c_1+\theta_1) \\ \sqrt{2\gamma}(c_2+\theta_2) \\ \hline -\frac{1}{\sqrt{n}}\mathbf{e} \end{bmatrix}. \tag{4.7.8}$$

Note that the penalised constraint contributions appear first. This is because they typically have a larger scale than the likelihood contributions, and this ordering is advisable for numerical stability when a QR factorization is used to solve the least squares problem (compare (2.3.42)).

Numerical results are presented for computations carried out using $\mu_1 = 1.$, $\mu_2 = 2.$ *for two cases:*

- $\sigma_1 = \sigma_2 = .5$, *and*
- $\sigma_1 = \sigma_2 = .7$.

A composition algorithm [95] is used to produce random numbers to provide data on the mixture density for $n = 100,\ 1000,\ 10000$. *Results obtained using two different seeds for a uniform generator are displayed in the tables given below. In both the estimated values and a summary of the iteration progress are given. The latter is given in the column headed 'P–H steps' which summarises the number of scoring iterations in each Powell–Hestenes step. The final column gives the computed multiplier estimates. In each case the exact values of the parameters were taken as starting values and appear to provide a fair test while the initial multiplier estimates were set to* 0. *The estimate computed by the 'inner' scoring algorithm is accepted when the magnitude of the directional derivative computed from (4.2.24) is less than* 10^{-6}. *The initial step taken in the linesearch is checked and modified appropriately to ensure that the argument of the log in evaluating the log likelihood is positive. The 'outer' iteration is terminated when* $\|\mathbf{c}\| < 10^{-4}$. *Note that the performance of the algorithm improves significantly as* n *is increased. Note also that the sample replacement (4.7.7) for the expected Hessian does not appear to have caused any deleterious effects in the scoring iteration in this application.*

n	μ_1	μ_2	σ_1	σ_2	P–H steps	multipliers
100	1.151	2.097	.4691	.5456	(5,3,2,2)	1.008, .9994
1000	.9907	1.997	.4770	.5151	(3,2,2)	1.004, .9980
10000	1.030	2.001	.5220	.5063	(3,2)	.9982, 1.006
100	1.072	2.131	.5749	.7568	(8,5,2,2)	1.0064, .9966
1000	.9933	1.993	.6605	.7274	(4,2,2)	1.002, .9992
10000	1.081	1.990	.7422	.7156	(3,2)	.9996, 1.000

Table 4.7: Results for first seed.

n	μ_1	μ_2	σ_1	σ_2	P–H steps	multipliers
100	.9390	1.997	.5602	.5586	(6,6,2,2)	1.019, .9911
1000	1.103	2.030	.5851	.5016	(3,2,2)	1.000, .9998
10000	1.028	1.999	.5261	.5133	(3,2)	1.002, .9988
100	.7886	1.982	.7808	..7536	(6,4,2,2)	1.005, .9980
1000	1.231	2.002	.8534	.7120	(4,2,2)	.9996, 1.000
10000	1.110	1.986	.7652	.7270	(3,2)	1.000, .9998

Table 4.8: Results for second seed.

Exercise 4.7.1 *The numerical results suggest that the Lagrange multipliers have asymptotic limit* 1. *Show that this result holds for the mixture problem as formulated here using constraints.*

4.7.2 A trust region method

This subsection is based on [61] which gives a careful discussion of the implementation of ideas due to Byrd and Omojokun for developing a trust region method for equality constrained optimization problems . This algorithm aims to improve on the Powell-Hestenes method by combining the two processes of reducing the objective function and satisfying the equality constraints. By avoiding the need to do a complete minimization of the objective function at each iteration it is expected that the total work necessary will be reduced considerably, but at the cost of developing a significantly more complicated algorithm. The key to this combined approach is to notice that to a significant extent the two processes of minimization and multiplier update are independent activities as locally the constraints are approached most rapidly in the direction of the constraint normals while the constrained optimization has to be completed in the tangent space of the constraints. The importance of a compromise can be seen by considering what might be considered a first

try at generating a descent direction in a trust region context. Let

$$l(\mathbf{x}, \boldsymbol{\lambda}) = \mathcal{L} - \boldsymbol{\lambda}^T \mathbf{c},\ \mathbf{g} = \nabla_{\mathbf{x}} \mathcal{L}^T.$$

Then a possible system to be considered based on a sequential quadratic programming (SQP) approach could be

$$\begin{aligned} \min_{\mathbf{h}} \mathbf{h}^T \mathbf{g} + \frac{1}{2} \mathbf{h}^T \nabla_{\mathbf{x}}^2 l \mathbf{h}, \\ C\mathbf{h} + \mathbf{c} = 0, \\ \|\mathbf{h}\| \leq \Delta. \end{aligned}$$

However, this system will not be consistent in general because the norm constraint on $\mathbf{h}$ will prevent satisfaction of the linear constraints if the bound Δ is too small. To compromise let $\zeta \in (0, 1)$ be a relaxation factor, and consider a (vertical or normal) step toward constraint satisfaction defined by

$$\min_{\mathbf{v}} \|C\mathbf{v} + \mathbf{c}\|,\ \|\mathbf{v}\| \leq \zeta\Delta.$$

The uncoupling idea is employed by setting $\mathbf{v} = C^T \mathbf{w}$ which leads to a simple constrained least squares problem for $\mathbf{v}$. Now, to reduce the function value, consider the variant on the original proposal given by

$$\begin{aligned} \min_{\mathbf{h}} \mathbf{h}^T \mathbf{g} + \frac{1}{2} \mathbf{h}^T \nabla_{\mathbf{x}}^2 l \mathbf{h}, \\ C\mathbf{h} = C\mathbf{v}, \\ \|\mathbf{h}\| \leq \Delta. \end{aligned}$$

This problem has a nonempty feasible region because it already contains $\mathbf{v}$. An improvement can be sought in an approximation to the tangent space by setting

$$\mathbf{h} = \mathbf{v} + Z\mathbf{d},$$

where Z is a basis for this tangent space satisfying

$$CZ = 0.$$

With this choice the linear constraints are automatically satisfied so the new problem, after dropping constant terms, reduces to

$$\begin{aligned} \min_{\mathbf{d}} \left(\mathbf{g} + \nabla_{\mathbf{x}}^2 l \mathbf{v}\right)^T Z\mathbf{d} + \frac{1}{2} \mathbf{d}^T Z^T \nabla_{\mathbf{x}}^2 l Z\mathbf{d}, \\ \|Z\mathbf{d}\| \leq \sqrt{\left\{\Delta^2 - \|\mathbf{v}\|^2\right\}}. \end{aligned}$$

The idea now is to update $\mathbf{x}$

$$\mathbf{x} \leftarrow \mathbf{x} + \mathbf{h}$$

provided there is a suitable reduction in the monitor function. The form suggested here has the exact penalty form

$$\psi = L + \mu \|\mathbf{c}\| .$$

A suitable reduction is decided by comparing the actual reduction in ψ (ared) with the linear prediction

$$\text{pred} = -\mathbf{h}^T\mathbf{g} - \frac{1}{2}\mathbf{h}^T\nabla^2_{\mathbf{x}}l\mathbf{h} + \mu\left(\|\mathbf{c}\| - \|\mathbf{c} + C\mathbf{h}\|\right).$$

Otherwise Δ is reduced and the above process repeated. If the step is successful then it is necessary to consider the possibility of increasing Δ, and the estimate of the Lagrange multipliers $\boldsymbol{\lambda}$ must be upated. This can be done by solving

$$CC^T\boldsymbol{\lambda} = C\mathbf{g}$$

at the new point. The devil is, of course, in the detail, and [61] should be consulted, especially on the implementation techniques needed to cope with large scale problems.

It appears that typically this trust region approach outperforms the relatively simple and robust Powell-Hestenes type of method which suffers from the disadvantage of having to restart the computation at each update step. This is illustrated in the following table taken from [64]in which the mixture density estimation problem is solved by two methods. The first is a new algorithm which uses the Bird and Omojokum trust region approach and includes a scoring option (as in Powell-Hestenes) or a quasi-Newton update option, selecting between them on the basis of progress. The second is a fairly standard, well performed SQP implementation due to Schittkowski (KSc85).

4.8 Appendix 1: The Powell-Hestenes method

The derivation of the Powell–Hestenes correction is carried out for the case of linear constraints ($c_i(\mathbf{x}) = \mathbf{c}_i^T\mathbf{x} - d_i$, $i = 1, 2, \cdots, m$) for simplicity. The necessary conditions for a minimum of $\mathcal{P}_n$ give

$$-\frac{1}{n}\nabla_{\mathbf{x}}\mathcal{L}_n + 2\gamma\sum_{i=1}^{m}(\mathbf{c}_i^T\mathbf{x} - d_i + \theta_i)\mathbf{c}_i^T = 0,$$

case 1		starting from true parameters				starting from hypothesised values			
		MLESOL		NLQPL		MLESOL		NLQPL	
	n	n_{iter}	n_f	n_{iter}	n_f	n_{iter}	n_f	n_{iter}	n_f
	10	9	10	15	23	9	10	17	21
	10^2	11	12	9	12	16	17	11	25
	10^3	4	5	13	22	8	9	13	16
	10^4	3	4	9	18	6	7	15	19
case 2		starting from true parameters				starting from hypothesised values			
		MLESOL		NLQPL		MLESOL		NLQPL	
	n	n_{iter}	n_f	n_{iter}	n_f	n_{iter}	n_f	n_{iter}	n_f
	10	16	17	15	17	11	12	19	20
	10^2	6	7	21	37	8	9	29	35
	10^3	6	7	9	21	4	5	14	26
	10^4	4	5	16	23	6	7	25	36

Table 4.9: Comparison results of MLESOL and NLQPL

and these determine $\mathbf{x}$ as a function of $\boldsymbol{\theta}$. The aim is to adjust $\boldsymbol{\theta}$ so that

$$C\mathbf{x}(\boldsymbol{\theta}) - \mathbf{d} = 0,$$

where $C : R^p \to R^m$, $C_{i*} = \mathbf{c}_i^T$. If a Newton iteration is used to solve this equation then a correction $\boldsymbol{\delta}_\theta$ to the current $\boldsymbol{\theta}$ is given by

$$C\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}\boldsymbol{\delta}_\theta = -(C\mathbf{x}(\boldsymbol{\theta}) - \mathbf{d}).$$

To calculate $\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}$ differentiate the necessary conditions to obtain the equation

$$\left\{-\frac{1}{n}\nabla^2_{\mathbf{x}}\mathcal{L}_n + 2\gamma C^T C\right\}\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}} = -2\gamma C^T.$$

The special form of the right hand side should be noted. Transforming this equation using the orthogonal factorization $C^T = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}\begin{bmatrix} U \\ 0 \end{bmatrix}$ gives

$$Q^T\left\{-\frac{1}{n}\nabla^2_{\mathbf{x}}\mathcal{L}_n + 2\gamma C^T C\right\}QQ^T\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}} = -2\gamma\begin{bmatrix} U \\ 0 \end{bmatrix}.$$

To compute the inverse of $Q^T\nabla^2_{\mathbf{x}}H_nQ$ when γ is large let

$$Q^T\left\{-\frac{1}{n}\nabla^2_{\mathbf{x}}\mathcal{L}_n\right\}Q = \begin{bmatrix} A & B \\ B^T & D \end{bmatrix}$$

Then a straightforward computation gives the estimate

$$\begin{bmatrix} \frac{1}{2\gamma}U^{-T}U^{-1} + \mathcal{O}(\frac{1}{\gamma^2}) & -\frac{1}{2\gamma}U^{-T}U^{-1}BD^{-1} + \mathcal{O}(\frac{1}{\gamma^2}) \\ -\frac{1}{2\gamma}D^{-1}B^TU^{-T}U^{-1} + \mathcal{O}(\frac{1}{\gamma^2}) & D^{-1} + \mathcal{O}(\frac{1}{\gamma}) \end{bmatrix} \tag{4.8.1}$$

Thus

$$Q_1^T\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}} = -U^{-T} + \mathcal{O}(\frac{1}{\gamma}),$$

so that

$$C\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}} = -I + \mathcal{O}(\frac{1}{\gamma}).$$

Substituting in the Newton step gives

$$\boldsymbol{\delta}_\theta = [I + O(1/\gamma)]\,[C\mathbf{x} - \mathbf{d}] \tag{4.8.2}$$

The advantage of a good estimate for the Lagrange multipliers can be seen by arguing in similar fashion. Let $2\gamma\theta_i = \lambda_i + \epsilon_i$ where λ_i is the exact multiplier and $\hat{\mathbf{x}}$ the solution of the constrained problem. Then

$$-\frac{1}{n}\nabla_{\mathbf{x}}\mathcal{L}_n(\mathbf{x}) + 2\gamma(C\mathbf{x} - \mathbf{d} + \boldsymbol{\theta})^TC = 0,$$
$$-\frac{1}{n}\nabla_{\mathbf{x}}\mathcal{L}_n(\widehat{\mathbf{x}}) + \boldsymbol{\lambda}^TC = 0.$$

Subtracting gives

$$\left(-\frac{1}{n}\overline{\nabla^2_{\mathbf{x}}\mathcal{L}_n} + 2\gamma C^TC\right)(\mathbf{x} - \hat{\mathbf{x}}) = -2C^T\boldsymbol{\epsilon},$$

where the bar denotes a mean value is appropriate. Arguing as above gives

$$\|\mathbf{x} - \hat{\mathbf{x}}\| = \mathcal{O}(\|\boldsymbol{\epsilon}\|/\gamma). \tag{4.8.3}$$

The above development of the Powell-Hestenes algorithm follows the original derivation given by Powell. The feature of the method in this form is the nonlinear least squares formulation which suggests it should allow for easy extension of the scoring method, but this requires more work. To connect with a more general mathematical programming formulation note that the necessary conditions for a minimum give the equations

$$-\frac{1}{n}\nabla_x\mathcal{L}_n + \sum_{i=1}^{k} 2\sigma\theta_i\nabla_x c_i = 0, \quad \mathbf{c} = 0. \tag{4.8.4}$$

This gives

$$2\sigma\theta_i^* = \lambda_i^* \tag{4.8.5}$$

where λ_i^* is the Lagrange multiplier for the i'th constraint. This suggests an equivalent formulation in augmented Lagrangian form:

$$\widetilde{\mathcal{H}}_n = -\frac{1}{n}\mathcal{L}_n + \lambda^T \mathbf{c} + \sigma \mathbf{c}^T \mathbf{c} \tag{4.8.6}$$

which differs from $\mathcal{H}_n$ by $\sigma \boldsymbol{\theta}^T \boldsymbol{\theta}$. As this term is constant in each sequential minimization it does not affect the successive iterates. Let

$$\mathcal{D}_\sigma(\boldsymbol{\lambda}) = \min_{\mathbf{x}} \widetilde{\mathcal{H}}_n(\mathbf{x}, \boldsymbol{\lambda}). \tag{4.8.7}$$

Then $\boldsymbol{\lambda}^*$ maximizes $\mathcal{D}_\sigma(\boldsymbol{\lambda})$ under suitable conditions. It follows that the Powell-Hestenes correction can be interpreted as a correction step in computing the maximum of $\mathcal{D}_\sigma$. In this context it can be derived readily as a consequence of the identities

$$\nabla_\lambda \mathcal{D}_\sigma = \mathbf{c}, \tag{4.8.8}$$

$$\nabla_\lambda^2 \mathcal{D}_\sigma^{-1} = -\sigma I - \nabla_\lambda^2 \mathcal{D}_0^{-1}, \tag{4.8.9}$$

assuming the necessary inverses exist. Thus the Newton step for maximizing $\mathcal{D}_\sigma$ can be considered as made up from

1. the Powell-Hestenes step, and
2. a Newton step for maximizing $\mathcal{D}_0$.

Methods for estimating this second component of the Newton step in order to accelerate the Powell-Hestenes step have been considered by several authors. The simplest is due to Jittorntrum [53] who suggests a one parameter correction of the form

$$\boldsymbol{\lambda}_{j+1} = \boldsymbol{\Lambda}^j + w_j \left(\boldsymbol{\Lambda}^j - \boldsymbol{\Lambda}^{j-1} \right),$$

where $\boldsymbol{\Lambda}^j$ is the Powell-Hestenes corrected estimate at step j. He shows that $\{\boldsymbol{\lambda}_j\}$ is a valid sequence of multipliers provided $\{w_j\} \subset [\alpha, \beta]$, a finite interval, and σ is large enough. The Jittorntrum correction at step j is

$$w_j = -\frac{\mathbf{c}_j^T \left(\boldsymbol{\Lambda}^j - \boldsymbol{\Lambda}^{j-1} \right)}{\left(\mathbf{c}_j - \mathbf{c}_{j-1} \right)^T \left(\boldsymbol{\Lambda}^j - \boldsymbol{\Lambda}^{j-1} \right)}. \tag{4.8.10}$$

Fletcher suggests using a quasi-Newton approach to estimate the correction which proves effective for general problems but does not fit in with the scoring philosophy adopted here.

Chapter 5

Parameter estimation in ordinary differential equations

5.1 Introduction

In its basic form the estimation problem seeks to determine information on a parameter vector $\boldsymbol{\beta} \in R^p$ characterizing a particular implementation of a process modelled by the system of ordinary differential equations

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}\left(t, \mathbf{x}, \boldsymbol{\beta}\right), \tag{5.1.1}$$

where $\mathbf{f} \in R \times R^m \times R^p \to R^m$ is assumed to be at least twice continuously differentiable, from observations on the state variable $\mathbf{x}\left(t\right) \in R \to R^m$ given by

$$\mathbf{y}_i = O_i \mathbf{x}\left(t_i\right) + \boldsymbol{\varepsilon}_i, \ i = 1, 2, \cdots, n, \tag{5.1.2}$$

where $\mathbf{y}_i \in R^k$, $O_i \in R^m \to R^k$ defines the observation process , and $\boldsymbol{\varepsilon}_i \in R^k$ is a vector of random variables corresponding to measurement errors in the observations. Note that the exact state variable values $\mathbf{x}\left(t_i\right)$ are unobserved. Hence their estimation is part of the problem solution. Because they enter the model in a different capacity to the actual parameters these latter are distinguished by referring to them as $\boldsymbol{\beta}$.

Superficially, this estimation problem is not a standard parametric modelling problem of the kind considered in the previous chapters because the manifold of system model responses depends not only on the parameters occurring explicitly in the differential equation, but also implicitly on parameters required to take account of intrinsic degrees of freedom equal in number to the order of the differential equation system. The implicit parameters are certainly not avoidable in the sense that the various means

used to specify them effectively classify the different algorithms applied to the estimation problem. However, they do have some properties in common with so-called ignorable or nuisance parameters as the knowledge acquired in determining an implicit parametrization is not of direct relevance to the desired answer to the original question asked. The manner of specifying the implicit parameters can directly influence problem solution computation properties such as the domain of attraction of the Gauss-Newton iteration. Two classes of method are highlighted. If additional conditions are adjoined to make explicit the specification of the implicit parameters then we refer to an *embedding method* . If the cost function is chosen appropriately and the differential system plus auxilliary conditions is integrated exactly to generate the state variable values needed to compare with the observations then the embedding method is an example of a maximum likelihood method . Thus previous discussion applies. New points of interest which occur here include choosing the form of embedding and the role of approximate methods in estimating the likelihood function. This leads naturally to the question of consistency of estimates determined from the approximate likelihood. In the alternative class of methods the differential equation is treated as an explicit constraint on the likelihood objective. This requires the use of techniques of constrained optimisation in estimating the explicit parameters. This approach is referred to as the class of *simultaeous methods* . An immediate question is "do these two classes of method produce identical results"?

Remark 5.1.1 *If the vector of parameters $\boldsymbol{\beta}$ is null then the problem reduces to that of finding a smoothed approximation to the state variable trajectory generating the noisy data.The general estimation problem can be formulated as a smoothing problem by adjoining the differential equations $\frac{d\boldsymbol{\beta}}{dt} = 0$ to (5.1.1) and augmenting the state vector $\begin{bmatrix} \mathbf{x} \\ \boldsymbol{\beta} \end{bmatrix} \rightarrow \mathbf{x}$. This proves convenient in describing the simultaneous class of methods in which the state variables and parameter values are estimated together. However, it may not be the most efficient way to implement this class of methods. It is convenient also in Example 5.3.1 where a simple case is considered in the solution of a nonlinear boundary value problem. In the embedding methods on the other hand the state variables are determined in each iteration by explicit integration of the differential system and then applied in a correction step to update the current parameter values. One point to note is that if the differential equation is linear in the state variables then this property will most likely be lost in the smoothing problem formulation .*

Here, for ease of presentation, the following assumptions are made:

1. The random variables $\boldsymbol{\varepsilon}_i$, $i = 1, 2, \cdots, n$, are independent and can be well replicated by samples from a normal distribution with mean 0 and covariance matrix $\sigma^2 I$.

2. The sampling interval is bounded so that there is no restriction in assuming $0 \leq t_1 < t_2 < \cdots < t_n \leq 1$.

3. The model is assumed to be a true model in the sense that $\exists \boldsymbol{\beta}^* \ni \mathbf{x}^*(t) = \mathbf{x}(t, \boldsymbol{\beta}^*)$ both satifies (5.1.1) and generates the observed data through (5.1.2).

In this case both the method of maximum likelihood and nonlinear least squares lead to the same estimation principle:

$$\min_{\mathbf{x}(t_i),\, i=1,2,\cdots,n,\, \boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^{n} \|\mathbf{r}_i\|_2^2,\ \mathbf{r}_i = \mathbf{y}_i - O_i \mathbf{x}(t_i),\ \frac{d\mathbf{x}}{dt} = \mathbf{f}(t, \mathbf{x}, \boldsymbol{\beta}). \quad (5.1.3)$$

This is a constrained optimization problem in which the values of the state variable $\mathbf{x}(t_i)$ are constrained by the differential equation. It is assumed that this optimization problem has a unique solution $\widehat{\mathbf{x}}\left(t, \widehat{\boldsymbol{\beta}}\right)$ which satisfies appropriate first order necessary and second order sufficiency conditions [73].

Remark 5.1.2 *This last condition is distinctly nontrivial and begs some important questions. For example, let $k = 1$ so that $O_i = \mathbf{o}_i^T$. What restrictions if any does the form of $\mathbf{o}_i$ place on solvability of the estimation problem. Some insight can be gained from the following development. Assume a fixed representer $\mathbf{o}$ for the observation functional , let $\mathbf{x}^*(t)$ correspond to the hypothesized "true" model solution which generated the deterministic component of the data, and set $\eta(t) = \mathbf{o}^T \mathbf{x}^*(t)$. Recovery of $\eta(t)$ from the observed data only could be attempted by a smoothing calculation provided n is large enough (at least hypothetically). Repeated differentiation of $\eta(t)$ at $t = t_0$ gives*

$$\mathbf{o}^T \frac{d^s \mathbf{x}}{dt^s} = \mathbf{o}^T \boldsymbol{\phi}_s(t_0, \mathbf{x}, \boldsymbol{\beta}^*) = \frac{d^s \eta(t_0)}{dt^s}, \quad s = 0, 1, \cdots,$$

where $\boldsymbol{\phi}_s$ depends on its arguments through products of mixed partial derivatives of $\mathbf{f}$ of order up to $s - 1$. For example,

$$\begin{aligned} \boldsymbol{\phi}_0 &= \mathbf{x}(t_0)), \\ \boldsymbol{\phi}_1 &= \mathbf{f}(t_0, \mathbf{x}, \boldsymbol{\beta}^*), \\ \boldsymbol{\phi}_2 &= \nabla_x \mathbf{f}\, \mathbf{f} + \frac{\partial \mathbf{f}}{\partial t}. \end{aligned}$$

Taking $s = m-1$ gives a system of (nonlinear) equations for $\mathbf{x}^*(t_0)$ given the values of the derivatives of $\eta(t)$ at $t = t_0$. If the Jacobian of this system is nonsingular then it can be solved by Newton's method and the model solution can then be reconstructed in consistent fashion by using the differential equation to develop successive terms in the Taylor series for $\mathbf{x}$ about $\mathbf{x}^*(t_0)$. The nonsingularity condition on the Jacobian can be interpreted as a condition on $\mathbf{o}$. There is a connection here to mathematical systems theory with $\mathbf{o}^T\mathbf{x}$ corresponding to the systems output, and the Jacobian rank condition being the requirement of observability as in Remark 1.4.1 in Chapter 1.

If the differential equation is linear in the state variable $\mathbf{x}$ and has the particular form

$$\mathbf{f}(t, \mathbf{x}, \boldsymbol{\beta}) = M(t, \boldsymbol{\beta})\,\mathbf{x} + \mathbf{q}(t)$$

then $\nabla_x \mathbf{f} = M$. Here the Jacobian can be computed relatively easily for small values of m. For example, if $m = 3$ then

$$J_3 = \nabla_x \begin{bmatrix} \mathbf{o}^T\boldsymbol{\phi}_0 \\ \mathbf{o}^T\boldsymbol{\phi}_1 \\ \mathbf{o}^T\boldsymbol{\phi}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{o}^T \\ \mathbf{o}^T M \\ \mathbf{o}^T\left(M^2 + \frac{dM}{dt}\right) \end{bmatrix}.$$

Example 5.1.1 *Consider the simple chemical reaction $A \to B \to C$ with rate constants β_1 and β_2 respectively. The corresponding differential system is*

$$\frac{dA}{dt} = -\beta_1 A, \tag{5.1.4}$$

$$\frac{dB}{dt} = \beta_1 A - \beta_2 B, \tag{5.1.5}$$

$$\frac{dC}{dt} = \beta_2 B. \tag{5.1.6}$$

Here

$$M = \begin{bmatrix} -\beta_1 & 0 & 0 \\ \beta_1 & -\beta_2 & 0 \\ 0 & \beta_2 & 0 \end{bmatrix},$$

$$J_3 = \begin{bmatrix} o_1 & o_2 & o_3 \\ \beta_1(o_2 - o_1) & \beta_2(o_3 - o_2) & 0 \\ \beta_1^2(o_2 - o_1) + \beta_1\beta_2(o_3 - o_2) & -\beta_2^2(o_3 - o_2) & 0 \end{bmatrix}.$$

Note that J_3 has a column of zeros if $o_3 = 0$. This indicates that the observation functional should include a component of the reaction product if the

data is to be estimable. The eigen decomposition $V \operatorname{diag}\{-\beta_1, -\beta_2, 0\} V^{-1}$ *of* M *is*

$$\begin{bmatrix} \beta_1 - \beta_2 & 0 & 0 \\ -\beta_1 & 1 & 0 \\ \beta_2 & -1 & 1 \end{bmatrix} \begin{bmatrix} -\beta_1 & & \\ & -\beta_2 & \\ & & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\beta_1 - \beta_2} & 0 & 0 \\ \frac{\beta_1}{\beta_1 - \beta_2} & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}.$$

It follows that the fundamental matrix (5.2.5) of the differential equation is $X(t,0) = V \exp(\operatorname{diag}\{-\beta_1, -\beta_2, 0\} t) V^{-1}$. *The row of the Gauss-Newton design matrix corresponding to the residual* $r_i = y_i - \mathbf{o}^T X(t_i, 0)\,\mathbf{z}$ *for the problem of estimating* β_1, β_2, *and* $\mathbf{z}$ *is*

$$\mathbf{o}^T \begin{bmatrix} V \exp(\operatorname{diag}\{-\beta_1, -\beta_2, 0\} t_i) & \nabla_\beta (X(t_i, 0)\,\mathbf{z}) \end{bmatrix} \begin{bmatrix} V^{-1} & \\ & I \end{bmatrix}.$$

This shows that because V *is lower triangular the design matrix cannot have full column rank unless* $o_3 \neq 0$ *in agreement with the above prediction. This problem is not encountered directly if the estimation problem uses the information that the differential equation solution is a sum of exponentials as in subsection 4.6.1 and considers the residuals*

$$r_i = y_i - w_1 \exp(-\beta_1 t_i) - w_2 \exp(-\beta_2 t_i) - w_3.$$

However, successful estimation still requires that the observations contain significant contributions from both exponentials.

Remark 5.1.3 *There is a considerable literature on identifiability problems. Typically this considers the output function as given and the identifiability problem as associated with the parametrization so the above emphasis on the choice of a suitable representer* $\mathbf{o}$ *for the output function is somewhat different. Use of a Taylor series expansion to tackle the observability problem for nonlinear differential equations is considered explicitly in [89]. Recent work includes the development of polynomial time, semi-numerical algorithms for analysing the identification problems for systems of differential equations with rational function right hand sides [97].*

There are two main approaches to the estimation problem:

The embedding method (subsection 5.4.2) The embedding approach leads to an unconstrained optimization problem which can be solved by standard methods such as the Gauss-Newton form of the scoring algorithm. However, it removes the differential equation constraint (5.1.1) on the state variable $\mathbf{x}(t)$ by embedding the differential equation into a

parametrised family of boundary value problems which must be solved explicitly in order to generate trial values. This has the apparent disadvantage of requiring information on the structure of the differential equation at set-up time in order to impose stable boundary conditions

$$B_1\mathbf{x}(0) + B_2\mathbf{x}(1) = \mathbf{b} \tag{5.1.7}$$

on (5.1.1). Here $B_1, B_2 \in R^m \to R^m$ are assumed known while $\mathbf{b}$ is a vector of additional parameters which has to be determined as part of the estimation process. The key requirement of the embedding method is that the resulting system (5.1.7), (5.1.1) has a numerically well determined solution $\mathbf{x}(t, \boldsymbol{\beta}, \mathbf{b})$ for all $\boldsymbol{\beta}, \mathbf{b}$ in a large enough neighborhood of $\boldsymbol{\beta}^*, \mathbf{b}^* = B_1\mathbf{x}^*(0) + B_2\mathbf{x}^*(1)$. The qualification that the solution be capable of being stably computed is important here. A sufficient condition discussed in subsection 5.2.5 is that the boundary conditions (5.1.7) be compatible with the dichotomy structure of the model equations (5.1.1) [6] when such a structure is available. Note that this might be a stronger condition than that the estimation problem has a well determined solution. Typically the embedding approach also requires additional information. For example, if it is known that the initial value problem for the differential equation is stable then the initial value choice $B_1 = I$, $B_2 = 0$ is allowable. If such a priori knowledge is not available then it is shown in the next section (Remark 5.2.1) that suitable conditions can be computed when the differential equation is linear. If the differential equation is nonlinear then the device available in the linear case can be applied with the aim of ensuring that the linear differential equations for the Newton correction have well determined solutions. Examples where this approach can be successful include unstable systems whose initial value trajectories are chaotic. Note nonlinearity could require that the selected boundary matrices need to be revised in order that the iteration steps of the estimation procedure are well determined.

The embedding approach has the advantage that it leads to algorithms that are relatively simple to formulate. Let

$$\nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathbf{x} = \left[\frac{\partial \mathbf{x}}{\partial \boldsymbol{\beta}}, \frac{\partial \mathbf{x}}{\partial \mathbf{b}}\right]. \tag{5.1.8}$$

Then the gradient of the objective function (5.1.3) is

$$\nabla_{(\boldsymbol{\beta},\mathbf{b})}F = -\frac{1}{n}\sum_{i=1}^{n} \mathbf{r}_i^T O_i^T \nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathbf{x}(t_i), \tag{5.1.9}$$

where the gradient term components can be evaluated by solving the linear boundary value problems

$$B_1 \frac{\partial \mathbf{x}}{\partial \boldsymbol{\beta}}(0) + B_2 \frac{\partial \mathbf{x}}{\partial \boldsymbol{\beta}}(1) = 0, \tag{5.1.10}$$

$$\frac{d}{dt}\frac{\partial \mathbf{x}}{\partial \boldsymbol{\beta}} = \nabla_{\mathbf{x}}\mathbf{f}\frac{\partial \mathbf{x}}{\partial \boldsymbol{\beta}} + \nabla_{\boldsymbol{\beta}}\mathbf{f}, \tag{5.1.11}$$

and

$$B_1 \frac{\partial \mathbf{x}}{\partial \mathbf{b}}(0) + B_2 \frac{\partial \mathbf{x}}{\partial \mathbf{b}}(1) = I, \tag{5.1.12}$$

$$\frac{d}{dt}\frac{\partial \mathbf{x}}{\partial \mathbf{b}} = \nabla_{\mathbf{x}}\mathbf{f}\frac{\partial \mathbf{x}}{\partial \mathbf{b}}. \tag{5.1.13}$$

Given this information then the scoring (Gauss-Newton) algorithm is applied readily.

The simultaneous method (subsection 5.4.3) This second class of methods has been called the simultaneous approach in [105]. The idea here is to use a discretization of the differential equation to impose constraints on the objective function (5.1.3). For example, if the discretization is based on the trapezoidal rule then the resulting constraint set is

$$\mathbf{c}_i(\mathbf{x}, \boldsymbol{\beta}) = 0,\ i = 1, 2, \cdots, n-1, \tag{5.1.14}$$

where

$$\mathbf{c}_i(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}_{i+1} - \mathbf{x}_i - \frac{\Delta t_i}{2}\left(\mathbf{f}(t_{i+1}, \mathbf{x}_{i+1}, \boldsymbol{\beta}) + \mathbf{f}(t_i, \mathbf{x}_i, \boldsymbol{\beta})\right). \tag{5.1.15}$$

This has the form of a first order recurrence for the state variable values $\mathbf{x}_i$. If the differential equation is linear then the resulting matrix representation of the linear system is *block bi-diagonal* . Here $\Delta t_i = t_{i+1} - t_i$. The simultaneous problem is

$$\min_{\mathbf{x}(t_i),\, i=1,2,\cdots,n,\, \boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n} \|\mathbf{r}_i\|_2^2; \quad \mathbf{c}(\mathbf{x}, \boldsymbol{\beta}) = 0, \tag{5.1.16}$$

where $\mathbf{c}$ is the composite vector with block components $\mathbf{c}_i$ given by equation (5.1.14). The obvious disadvantage of this approach is that the number of constraints tends to ∞ with n. However, the number of effective degrees of freedom is just m which is the number of independent pieces of information which must be added to (5.1.1) in order to

specify the solution uniquely. These additional degrees of freedom are absorbed by the Lagrange multipliers in the simultaneous approach. Here the Bock method [12] has a similar position to the Gauss-Newton method in the embedding method. *The common idea is to drop the second derivative terms which in this case are derived from the multiplier coefficients in the augmented matrix in Newton's method applied to the necessary conditions*. It turns out that the Bock method in the simultaneous approach has a similar convergence rate estimate to the Gauss-Newton iteration in the embedding method, but the justification required is not only quite different, it also makes explicit use of the property that the observational errors are normally distributed . Here this requirement is bound directly to the form of the objective function and the assumption that the observational errors are normally distributed.

A necessary step in both the embedding and simultaneous methods involves the discretization of the system of differential equations. Much of the development presented here centres on the trapezoidal rule (5.1.15). This particular discretization has advantages both in mimicking the stability properties of the differential equation and in leading to compact algebraic problem representations. This latter point proves to have advantages in analysing the more detailed properties of the estimation methods. The trapezoidal rule is the simplest case of compact, symmetric collocation methods [6] . Use of other members of this family may be appropriate if achieving required accuracy in the integration of the differential equation proves difficult. The basic idea used is that of building up interpolation polynomials of increasing order by requiring them to satisfy the differential equation at a sequence of interpolation (collocation) points in $[t_i, t_{i+1}]$ as well as fitting to the solution and its derivative values at t_i, t_{i+1}. If there are k interpolation points then a polynomial of degree $k+3$ is suggested. If a polynomial of degree $k+2$ is specified then the equations defining the interpolation polynomial can be solved only if a compatibility condition on the data is satisfied. This condition is then interpreted as the desired discretization. Here compact refers to the restriction of the interpolation data to the interval $[t_i, t_{i+1}]$. The method is symmetric if the interpolation points are distributed symmetrically in the interval. Such a choice tends to have truncation error advantages, but unsymmetric formulae can have interesting properties. This approach was suggested originally as a means for computer generation of differential equation discretizations in [74]. A single collocation point at the mid point of the interval picks up a form of

Simpson's rule :

$$\mathbf{x}(t_{i+1}) - \mathbf{x}(t_i) = \frac{\Delta t_i}{6}\{\frac{d\mathbf{x}(t_i)}{dt} + 4\frac{d\mathbf{x}(t_{i+1/2})}{dt} + \frac{d\mathbf{x}(t_{i+1})}{dt}\}, \tag{5.1.17}$$

$$\mathbf{x}(t_{i+1/2}) = \frac{1}{2}\left(\mathbf{x}(t_{i+1}) + \mathbf{x}(t_i)\right) - \frac{\Delta t_i}{8}\left(\frac{d\mathbf{x}(t_{i+1})}{dt} - \frac{d\mathbf{x}(t_i)}{dt}\right). \tag{5.1.18}$$

The desired discretization can be obtained by eliminating $\mathbf{x}(t_{i+1/2})$ between these equations using the differential equation. Smaller truncation errors can be obtained by using additional collocation points, and use of Gauss-Lobatto points proves particularly favourable. One possibility in the simultaneous method is to use equations (5.1.17, 5.1.18) directly as constraints on the objective function.

An alternative approach to the integration of the boundary value problem is the method of multiple shooting [75] . This makes use of the *exact discretization* available for linear differential equations

$$\mathbf{x}_{i+1} - X\left(t_{i+1}, t_i\right)\mathbf{x}_i = \mathbf{v}_i, \tag{5.1.19}$$

where the fundamental matrix $X\left(t_{i+1}, t_i\right)$ is defined in (5.2.5), and the particular integral $\mathbf{v}_i$ in (5.2.7). This exact form also has a block bi-diagonal matrix representation. It is used here only for reference purposes. However, it has been applied in the very successful software package PARFIT based on [12]. Traditionally initial value methods have been used to estimate the fundamental matrices. This is a possible cause for concern in the situation where the differential equation can support rapidly varying increasing and/or decreasing solutions while the actual signal is more slowly varying and satisfactorily represented on a relatively sparse set of mesh points $t_i \in \mathbf{K},\ i = 1, 2, \cdots, k = |\mathbf{K}|$. Such rapidly increasing solutions would be a problem for multiple shooting implemented using an initial value solver to evaluate the component fundamental matrices. The reason for this is that if the shooting points $\mathbf{K}$ are chosen using a bound on the corresponding fundamental solution matrix norms as in [75] then the grid spacing is determined by the requirement to follow the more rapidly increasing members of the set of solutions rather than to follow the more slowly changing target solution. This could be thought of as a characteristic example of the use of a non-stiff solution procedure on a stiff problem . However, the choice of solution method corresponding to this situation is potentially much more interesting, and in Example 5.2.4 a problem which fits this category is discussed. Here the differential equation is discretized on a grid on which the rapidly growing solutions are poorly represented by the trapezoidal rule discretization, but the slowly varying problem solution is well represented and

is still found satisfactorily using the trapezoidal rule discretization *when the boundary conditions are chosen appropriately*.

Approximations to values of the state variables are going to be needed at each data point $t \in \mathbf{T}$ in order to evaluate the log likelihood, and this is going to put constraints on the choice of the discretization grid $\mathbf{K}$. Unless otherwise stated it will be a convenient simplification to assume that points of both grids are equispaced in $0 \leq t \leq 1$. The possibilities for choosing the discretization grid are the following:

1. Integration of the differential equation is easy and a grid $\mathbf{K}$ coarser than that corresponding to the set of observation points $\mathbf{T}$ is adequate. In this case interpolation methods are used to estimate solution values for points of $\mathbf{T}$ not in $\mathbf{K}$. Linear interpolation provides a method with an accuracy compatible with the trapezoidal rule approximation.

2. It proves convenient to work with the solution grid $\mathbf{K}$ and the observation set $\mathbf{T}$ identical. This requires that $\mathbf{T}$ provide a suitable mesh for the differential equation integration in addition to being part of the refinement process of a regular sampling scheme.

3. Sufficiently accurate integration of the differential equation needs a finer solution grid $\mathbf{K}$ than the given observation set $\mathbf{T}$. If $\mathbf{T} \nsubseteq \mathbf{K}$ then interpolation is required to provide the necessary solution values at the observation points.

Typically cases 1 and 2 above prove suitable strategies when the integration of the differential equation is relatively easy. The reason is that the parameter estimates reflect the stochastic properties of the errors in the observations and so the attainable accuracy in the case of regular experiments is restricted by the generic $O(n^{-1/2})$ convergence rate associated with maximum likelihood estimation, while it amounts to carelessness if the error from the discretization contributes more than the $O(k^{-2})$ rate associated with, for example, the trapezoidal rule in all but cases of severely unequal mesh grading. A similar error rate is achieved when linear interpolation is used to fill in intermediate solution values in case 1. It is a reasonable, minimum expectation that the sample points $t \in \mathbf{T}$ should be capable of well representing the general behaviour of the signal. The slow almost sure convergence rate attainable even in the large data set case makes very important the achievement of a small value for the standard deviation σ of the experimental error. Some improvement in the parameter estimates can be obtained by careful consideration of the experimental design [15]. Faster convergence rates are achievable in the closely related problem of frequency estimation , but rather different algorithms are favoured [92].

The most difficult problems correspond to nonlinear problems which do not meet the dichotomy criterion when linearised about the solution trajectory, but do have locally well determined solutions which typically change rapidly, often in narrow transition regions. In these circumstances adaptive mesh selection techniques can be expected to be important, as can continuation methods for developing suitable solution approximations systematically as the computation proceeds [6]. A classic example is provided by limit cycle behaviour in the Van der Pol equation , example 5.3.2 . What to observe in order to well determine the estimation problem in cases where the solution changes dramatically becomes a good question.

5.2 Linear differential equations

Equation (5.1.1) is specialised to

$$\frac{d\mathbf{x}}{dt} = M(t, \boldsymbol{\beta})\,\mathbf{x} + \mathbf{q}(t) \tag{5.2.1}$$

in this section. Note that this equation is linear in the state variables but not necessarily linear in the parameters.

Example 5.2.1 *One class of problems that leads to linear differential equations is obtained by considering separable regression models (3.5.1)*

$$\Phi(t) = \sum_{i=1}^{p} \alpha_i \phi_i(t, \beta),\ \boldsymbol{\beta} \in R^m. \tag{5.2.2}$$

Here $\Phi(t)$ satisfies the p'th order differential equation computed from the relation of linear dependence

$$\begin{vmatrix} \Phi & \phi_1 & \cdots & \phi_p \\ \Phi^{(1)} & \phi_1^{(1)} & \cdots & \phi_p^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ \Phi^{(p)} & \phi_1^{(p)} & \cdots & \phi_p^{(p)} \end{vmatrix} = 0.$$

Let

$$\Delta_i = (-1)^i \begin{vmatrix} \phi_1 & \cdots & \phi_p \\ \vdots & \vdots & \vdots \\ \phi_1^{(i-1)} & \cdots & \phi_p^{(i-1)} \\ \phi_1^{(\bar{i}+1)} & \cdots & \phi_p^{(i+1)} \\ \vdots & \vdots & \vdots \\ \phi_1^{(p)} & \cdots & \phi_p^{(p)} \end{vmatrix},$$

then this differential equation is

$$\Phi^{(p)} + \frac{\Delta_{p-1}}{\Delta_p}\Phi^{(p-1)} + \cdots + \frac{\Delta_0}{\Delta_p}\Phi = 0. \tag{5.2.3}$$

The condition for this equation to be nonsingular is $\Delta_p \neq 0$. *Note that* Δ_p *is the Wronskian of* $\phi_1, \phi_2, \cdots, \phi_p$ *so this is just the standard linear independence condition. Equation (5.2.3)can be converted into a first order system by setting*

$$x_i = \Phi^{(i-1)},\ i = 1, 2, \cdots, p.$$

The result is

$$\frac{d\mathbf{x}}{dt} = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -\frac{\Delta_0}{\Delta_p} & \cdots & \cdots & -\frac{\Delta_{p-1}}{\Delta_p} \end{bmatrix} \mathbf{x}. \tag{5.2.4}$$

Familiar examples include:

1. *exponential fitting :* $\phi_i = e^{\beta_i t},\ i = 1, 2, \cdots, p,$ *and*
2. *rational fitting :* $\phi_i = \frac{t^{i-1}}{1+\beta_1 t+\cdots+\beta_m t^p},\ i = 1, 2, \cdots, p.$

In general, if $L(\phi_1, \cdots, \phi_k)\,\phi = 0$ *is the k'th order differential equation satisfied by* $\phi_1, \cdots, \phi_k$, *then*

$$L_{k+1}(\phi_1, \cdots, \phi_{k+1})\,\phi_{k+1} = \left(\frac{d}{dt} - b_{k+1}\right) L_k(\phi_1, \cdots, \phi_k)\,\phi_{k+1} = 0$$

provided

$$b_{k+1} = \frac{\frac{d}{dt}L_k(\phi_1, \cdots, \phi_k)\,\phi_{k+1}}{L_k(\phi_1, \cdots, \phi_k)\,\phi_{k+1}}.$$

Thus, in principle, a sequence of differential equations of increasing complexity can be generated recursively provided $L_k(\phi_1, \cdots, \phi_k)\,\phi_{k+1} \neq 0$ *corresponding to linear independence of the ordered subsets of the* ϕ_i.

Exercise 5.2.1 *Show that the fitting problem is correctly posed for equation (5.2.4) given data in the form*

$$\eta_i = \mathbf{e}_1^T \mathbf{x}(t_i, \boldsymbol{\beta}) + \varepsilon_i,\ i = 1, 2, \cdots, n.$$

Associated with the first order system (5.2.1) is the family of fundamental matrices $X(t, \xi)$ defined by

$$\frac{dX}{dt} = M(t, \boldsymbol{\beta})\,X;\ X(\xi, \xi) = I. \tag{5.2.5}$$

Fundamental matrices have the properties :

1. $X(t,\eta) = X(t,\xi)\,X(\xi,\eta)\,;$
2. $X(t,\eta)^{-1} = X(\eta,t)\,;$
3. $\frac{dX^{-1}}{dt} = -X^{-1}M(t,\boldsymbol{\beta})\,;$
4. $\frac{dX^{T}(s,t)}{dt} = -M^{T}X^{T}(s,t)\,.$

Any solution of (5.2.1) can be written using the fundamental matrix as

$$\mathbf{x}(t) = X(t,\xi)\,\mathbf{x}(\xi) + \mathbf{v}(t,\xi)\,, \tag{5.2.6}$$

where the fixed vector $\mathbf{x}(\xi)$ can be thought of as fixing the degrees of freedom in the general solution , and the particular integral $\mathbf{v}(t,\xi)$ is given by

$$\mathbf{v}(t,\xi) = \int_{\xi}^{t} X(t,u)\,\mathbf{q}(u)\,du. \tag{5.2.7}$$

The trapezoidal rule discretization (5.1.15) provides an approximation to (5.2.6). The exact equation evaluates on the given discretization grid $\mathbf{K}$ to give

$$\mathbf{c}_i(\mathbf{x},\boldsymbol{\beta}) = \mathbf{x}_{i+1} - X(t_{i+1},t_i)\,\mathbf{x}_i - \mathbf{v}(t_{i+1},t_i)\,,\ i = 1,2,\cdots,k-1. \tag{5.2.8}$$

This corresponds to the multiple shooting or exact form of the constraint equations in contrast to the trapezoidal rule discretization or approximate form.

The embedding method requires the explicit solution of a system of boundary value problems at each step of the parameter estimation process in order to evaluate the state variables and their derivatives with respect to the parameters. As the differential and boundary operators change only as a consequence of changes to the parameter values $\boldsymbol{\beta}$, $\mathbf{b}$ when the boundary value problem is linear, this works if and only if the equation determining the state variables has a well determined solution. From (5.2.6) with the specialization $\xi = 0$ the general solution satisfies the boundary conditions (5.1.7) provided

$$(B_1 + B_2X(1,0))\,\mathbf{x}(0) + B_2\int_0^1 X(1,u)\mathbf{q}(u)du = \mathbf{b} \tag{5.2.9}$$

can be solved for $\mathbf{x}(0)$ for each set of parameter values $\boldsymbol{\beta}$. This is only possible for general $\mathbf{b}$ if $(B_1 + B_2X(1,0))$ has a bounded inverse for the range of $\boldsymbol{\beta}$ of interest. This condition is appropriate here because $\mathbf{b}$ has the role of the extra vector of parameters to be estimated and so must be regarded as a general vector unless additional information is avaiable.

5.2.1 Constraint elimination by cyclic reduction

Cyclic reduction in the form considered here is an elimination procedure applied recursively to a block bi-diagonal matrix system. This provides a useful generalisation to earlier discussions which considered a more restrictive case involving tri-diagonal matrices. In its simplest form each step expresses the current even indexed variables in terms of the immediately adjacent variables of odd index. It has been applied to the solution of boundary value problems by Wright [117] who was interested both in parallel implementation of the solution process and in questions relating to its stability. However, the memory access stride of the cyclic reduction process considered is generically a power of two, and this tends to cause computer memory contention problems which reduce the computational efficiencey [44]. But the flip side is that the process offers significant insight into the structure of the solutions of the system of differential equations, and it is in this sense that it is considered here.

It simplifies addressing in developing the cyclic reduction procedure to assume that $n = 2^k + 1$ - but it should be noted that this is not a necessary assumption in this context and a general formulation of the procedure under the name of wrap-around partitioning has been developed in [44]. Associated with the basic difference scheme (5.2.8) is the stencil for the initial step which expresses $\mathbf{x}_i$ in terms of $\mathbf{x}_{i-1},\ \mathbf{x}_{i+1}$:

$$\left[\begin{array}{cccc} -X_{i-1} & I & 0 & -\mathbf{v}_{i-1} \\ 0 & -X_i & I & -\mathbf{v}_i \end{array}\right],\ i = 2(2)2^k,$$

where $X_i = X\left(t_{i+1}, t_i\right),\ \mathbf{v}_i = \mathbf{v}\left(t_{i+1}, t_i\right)$. Operators C_i are introduced to produce the transformation

$$C_i^1 \left[\begin{array}{cccc} -X_{i-1} & I & 0 & -\mathbf{v}_{i-1} \\ 0 & -X_i & I & -\mathbf{v}_i \end{array}\right] \rightarrow \left[\begin{array}{cccc} V_i^1 & -I & W_i^1 & \mathbf{w}_i^1 \\ H_i^1 & 0 & G_i^1 & -\mathbf{d}_i^1 \end{array}\right]. \qquad (5.2.10)$$

Typically this transformation can be based on methods such as orthogonal reduction or partial pivoting with row interchanges. The right hand stencil in (5.2.10) has the following important interpretation. The first row gives the *interpolation equation* at stride 2^1,

$$\mathbf{x}_i = V_i^1 \mathbf{x}_{i-1} + W_i^1 \mathbf{x}_{i+1} + \mathbf{w}_i^1,$$

while the second row gives the *constraint equation* at stride 2^1,

$$H_i^1 \mathbf{x}_{i-1} + G_i^1 \mathbf{x}_{i+1} - \mathbf{d}_i^1 = 0.$$

This is the first step in a recursive procedure in which the constraint equations are successively reduced, and the interpolation equations updated. Let $\mathbf{x}_s$

be a state variable to be eliminated at step j corresponding to a stride of 2^j in the resulting constraint equation. The current adjacent values are written $\mathbf{x}_-$, $\mathbf{x}_+$, and these will form the support values for the interpolation equation for $\mathbf{x}_s$ after the elimination step. Let $\mathbf{x}_q$ be a value adjacent to $\mathbf{x}_s$ which has been eliminated already. At this stage it will have supports given either by $\mathbf{x}_-$, $\mathbf{x}_s$, with corresponding interpolation equation

$$\mathbf{x}_q = V_q^j \mathbf{x}_- + W_q^j \mathbf{x}_s + \mathbf{w}_q^j,$$

or by $\mathbf{x}_s$, $\mathbf{x}_+$, with corresponding interpolation equation

$$\mathbf{x}_q = V_q^j \mathbf{x}_s + W_q^j \mathbf{x}_+ + \mathbf{w}_q^j.$$

The next step of cyclic reduction applies to the constraint equations relating the active solution values (those not already eliminated) at the current stride. This gives

$$C_s^{j+1} \begin{bmatrix} H_-^j & G_-^j & 0 & -\mathbf{d}_-^j \\ 0 & H_+^j & G_+^j & -\mathbf{d}_+^j \end{bmatrix} \rightarrow \begin{bmatrix} V_s^{j+1} & -I & W_s^{j+1} & \mathbf{w}_s^{j+1} \\ H_s^{j+1} & 0 & G_s^{j+1} & -\mathbf{d}_s^{j+1} \end{bmatrix}.$$

The interpolation equations must now be updated by replacing $\mathbf{x}_s$ by its support using the new interpolation equation. This gives

$$\mathbf{x}_q = \left\{ V_q^j + W_q^j V_s^{j+1} \right\} \mathbf{x}_- + W_q^j W_s^{j+1} \mathbf{x}_+ + \mathbf{w}_q^j + W_q^j \mathbf{w}_s^{j+1},$$

for state variables with support $\mathbf{x}_-$, $\mathbf{x}_s$, or

$$\mathbf{x}_q = V_q^j V_s^{j+1} \mathbf{x}_- + \left\{ W_q^j + V_q^j W_s^{j+1} \right\} \mathbf{x}_+ + \mathbf{w}_q^j + V_q^j \mathbf{w}_s^{j+1},$$

for those with support $\mathbf{x}_s$, $\mathbf{x}_+$, while the constraint equation introduced is

$$H_s^{j+1} \mathbf{x}_- + G_s^{j+1} \mathbf{x}_+ - \mathbf{d}_s^{j+1} = 0.$$

After k elimination sweeps the result is a system of interpolation equations

$$\mathbf{x}_i = V(t_i)\, \mathbf{x}_1 + W(t_i)\, \mathbf{x}_n + \mathbf{w}(t_i), i = 1, 2, \cdots, n, \tag{5.2.11}$$

and a constraint equation

$$H\mathbf{x}_1 + G\mathbf{x}_n = \mathbf{d}. \tag{5.2.12}$$

Setting $i = 1$, n in the interpolation equations gives the boundary conditions

$$\begin{aligned} V(t_1) = I,\ V(t_n) = 0, \\ W(t_1) = 0,\ W(t_n) = I, \\ \mathbf{w}(t_1) = 0,\ \mathbf{w}(t_n) = 0. \end{aligned} \tag{5.2.13}$$

The estimation problem can now be written

$$\min_{\mathbf{x}_1,\mathbf{x}_n,\boldsymbol{\beta}} \sum_{i=1}^{n} \|\mathbf{y}_i - O_i\left(V\left(t_i\right)\mathbf{x}_1 + W\left(t_i\right)\mathbf{x}_n + \mathbf{w}\left(t_i\right)\right)\|_2^2 \tag{5.2.14}$$

subject to the m constraints (5.2.12). Here the Lagrange multipliers for this reduced problem pick up exactly the m degrees of freedom needed to specify the solution of the differential equation (5.1.1).

Remark 5.2.1 *The above development permits something to be said about the choice of appropriate boundary conditions for the embedding algorithm for the smoothing problem for a linear system of differential equations. These are characterized by the triple* $(B_1, B_2, \mathbf{b})$. *What is required is that* (B_1, B_2) *be chosen such that the system*

$$\begin{bmatrix} H & G \\ B_1 & B_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} \mathbf{d} \\ \mathbf{b} \end{bmatrix}$$

is computationally as well behaved as possible. Consider, for example, the orthogonal factorization

$$\begin{bmatrix} H & G \end{bmatrix} = \begin{bmatrix} U^T & 0 \end{bmatrix} Q^T.$$

Then suitable (B_1, B_2) *, possibly up to a suitable scale factor* ρ*, can be found from*

$$\begin{bmatrix} H & G \\ B_1 & B_2 \end{bmatrix} = \begin{bmatrix} U^T & 0 \\ 0 & \rho I \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix}. \tag{5.2.15}$$

This gives

$$\begin{bmatrix} B_1 & B_2 \end{bmatrix} = \rho Q_2^T. \tag{5.2.16}$$

This choice of boundary conditions depends soley on the differential equation and in this sense is natural . It can be considered best possible as it does not affect the spectral conditioning of (5.2.15) provided ρ *is in the range of the singular values of* U*. This provides a sense in which this choice has an essentially passive role in the solution of the boundary value problem. Here the conditioning of the matrix* U *in (5.2.15) provides a measure of the inherent sensitivity of solutions of the differential equation system to two point boundary conditions of the form (5.1.7) .*

The above discussion is becomes more complicated if there is a nontrivial parametric dependence for then a range of parameter values must be allowed for in the estimation problem, and this may militate against a single fixed choice of embedding boundary conditions. It is conceivable that the nonlinear behaviour in parameter space could require some adaptivity to be introduced into the boundary condition selection process.

The order in which the intermediate variables are eliminated is

$$\begin{array}{ccccc} 1+1 & 3+1 & 5+1 & 7+1 & \cdots \\ 1.2+1 & 3.2+1 & 5.2+1 & 7.2+1 & \cdots \\ 1.2^2 & 3.2^2+1 & 5.2^2+1 & 7.2^2+1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{array}.$$

If the individual frontal matrices C_i^j are orthogonal then the result of the processing of the constraint stencils in the case $n = 2^2 + 1$ is a factorization of the form $\widehat{Q}\widehat{R}$ where $\widehat{Q}$ is orthogonal and

$$\widehat{R} = \begin{bmatrix} V_2^1 & -I & W_2^1 & & \\ V_3^2 & & -I & & W_3^2 \\ & & V_4^1 & -I & W_4^1 \\ H & & & & G \end{bmatrix}. \tag{5.2.17}$$

Introducing permutations

$$p_R : 2, 4, 3, 1, 5 \to 1, 2, 3, 4, 5,$$
$$p_L : 1, 3, 2, 4 \to 1, 2, 3, 4,$$

and associated permutation matrices P_R, P_L then

$$R = P_L \widehat{R} P_R^T = \begin{bmatrix} -I & W_2 & & V_2 & \\ & -I & V_4 & & W_4 \\ & & -I & V_3 & W_3 \\ & & & H & G \end{bmatrix} \tag{5.2.18}$$

is basically an upper triangular matrix while $Q = \widehat{Q} P_L^T$ involves the transformation of the constraint equation stencils. In this context, the updating of the interpolation equations corresponds to steps which are intermediate steps in a backsubstitution based on $\widehat{R}$. Note that if the C_i are based on orthogonal transformations then the computations at each stage develop an orthogonal transformation taking the matrix of the (permuted) constraint equations to upper triangular form. The numerical stability of this case is considered in [118].

Exercise 5.2.2 *For the case $n = 2^2 + 1$ considered above write out the orthogonal matrix Q in terms of the C_i^j frontal matrices.*

5.2.2 Properties of the reduced system

There are two main points to make about the equations that result from the cyclic reduction process applied to (5.2.8):

1. The interpolation equation (5.2.11) is generic in the sense that it must hold for values $\mathbf{x}(t_1), \mathbf{x}(t_n)$ corresponding to any solution of the differential system.

2. The quantities $V_t, W_t, \mathbf{w}_t$ which define the interpolation equation, and the quantities $G, H, \mathbf{d}$ which define the constraint equation are by no means uniquely defined as premultiplying the transformation matrix C_i by any block upper triangular, nonsingular matrix

$$C \leftarrow \begin{bmatrix} R_1 & R_{12} \\ & R_2 \end{bmatrix} C \tag{5.2.19}$$

leads to another recurrence of the same form as it preserves the zero submatrix produced by the original transformation (5.2.10). It follows from (5.2.11) that the freedom in the interpolation equation resides in $R_1^{-1}R_{12}$ (the unit matrix must be preserved in the interpolation equations for different schemes to be comparable), and from (5.2.12) that the freedom in the constraint equation resides in R_2 which serves to scale this equation.

Note that, from (5.2.6), the final constraint equation for the multiple shooting form of the recurrence (5.2.8) requires that

$$G^{-1}H = -X(t_n, t_1),\ G^{-1}\mathbf{w}_1^k = \mathbf{v}(t_n, t_1) = \int_{t_1}^{t_n} X(t_n, u)\mathbf{q}(u)du$$

independent of R_2. These quantities will be approximated if discretized forms of (5.2.8) such as the trapezoidal rule are used.

Next assume that $V(t)$, $W(t)$, $\mathbf{w}(t)$ satisfying (5.2.11) are differentiable functions of their arguments. Differentiating and eliminating using the differential equation (5.2.1) gives

$$\begin{aligned} 0 &= \left\{ \begin{array}{c} \left(\frac{dV(t)}{dt} - MV(t)\right)\mathbf{x}(t_1) + \left(\frac{dW(t)}{dt} - MW(t)\right)\mathbf{x}(t_n) \\ +\frac{d\mathbf{w}(t)}{dt} - M\mathbf{w}(t) - \mathbf{q} \end{array} \right\}, \\ &= \left\{ \begin{array}{c} \left(\frac{d}{dt} - M\right)\{V(t) + W(t)X(t_n, t_1)\}\mathbf{x}(t_1) + \\ \left(\frac{d}{dt} - M\right)\{\mathbf{w}(t) + W(t)\mathbf{v}(t_n, t_1)\} - \mathbf{q} \end{array} \right\}, \end{aligned}$$

on substituting for $\mathbf{x}(t_n)$ using (5.2.6), (5.2.7). Thus $V(t) + W(t)X(t_n, t_1)$ satisfies the homogeneous differential equation as $\mathbf{x}(t_1)$ can be prescribed arbitrarily. However, neither $V(t)$ nor $W(t)$ can do so separately in general as the pair of boundary conditions (5.2.13) would overspecify the solution of the first order system. A similar comment applies to $\mathbf{w}(t)$.

In considering the generation of the multiple shooting formulation of the recurrence it is convenient to start with C_i which satisfies

$$C_i \begin{bmatrix} I \\ -X_i \end{bmatrix} = \begin{bmatrix} D_i \\ 0 \end{bmatrix} \tag{5.2.20}$$

where $X_i = X(t_{i+1}, t_i)$, and the exact form of D_i is determined by the transformation . A further scaling which takes D_i to the unit matrix is necessary to make the transformation correspond exactly to (5.2.10). However, it follows from (5.2.22) below that this rescaling, which involves premultiplying the transformed tableau by a nonsingular matrix, does not affect the determination of V, W, $\mathbf{w}$. Nor does it affect the freedom in the interpolation equation. The simplest choice satisfying (5.2.20) is given by

$$C_i = \begin{bmatrix} -X_i & I \\ X_i & I \end{bmatrix} \tag{5.2.21}$$

Assume equispaced points t_i for simplicity. Then $t_i = t_1 + (i-1)\Delta t$, $\Delta t = (t_n - t_1)/(n-1)$. Postmultiplying (5.2.10) by

$$\begin{bmatrix} \mathbf{x}(t_{i-1}) \\ \mathbf{x}(t_i) \\ \mathbf{x}(t_{i+1}) \\ 1 \end{bmatrix} = \begin{bmatrix} V(t_{i-1})\mathbf{x}(t_1) + W(t_{i-1})\mathbf{x}(t_n) + \mathbf{w}(t_{i-1}) \\ V(t_i)\mathbf{x}(t_1) + W(t_i)\mathbf{x}(t_n) + \mathbf{w}(t_i) \\ V(t_{i+1})\mathbf{x}(t_1) + W(t_{i+1})\mathbf{x}(t_n) + \mathbf{w}(t_{i+1}) \\ 1 \end{bmatrix}$$

gives

$$0 = \left\{ \begin{array}{l} \begin{bmatrix} -X_i & I \\ X_i & I \end{bmatrix} \begin{bmatrix} -X_{i-1} & I & 0 & -\mathbf{v}_{i-1} \\ 0 & -X_i & I & -\mathbf{v}_i \end{bmatrix} \\ \begin{bmatrix} V(t_{i-1})\mathbf{x}(t_1) + W(t_{i-1})\mathbf{x}(t_n) + \mathbf{w}(t_{i-1}) \\ V(t_i)\mathbf{x}(t_1) + W(t_i)\mathbf{x}(t_n) + \mathbf{w}(t_i) \\ V(t_{i+1})\mathbf{x}(t_1) + W(t_{i+1})\mathbf{x}(t_n) + \mathbf{w}(t_{i+1}) \\ 1 \end{bmatrix} \end{array} \right\}. \tag{5.2.22}$$

The second component equation that results from(5.2.22) gives no new information. Substituting the values for the terms in the recurrence relations in the first equation and collecting terms gives

$$\begin{aligned} &\{V(t_{i+1}) - 2X(t_{i+1}, t_i)V(t_i) + X(t_{i+1}, t_{i-1})V(t_{i-1})\}\mathbf{x}(t_1) + \\ &\{W(t_{i+1}) - 2X(t_{i+1}, t_i)W(t_i) + X(t_{i+1}, t_{i-1})W(t_{i-1})\}\mathbf{x}(t_n) + \\ &\mathbf{w}(t_{i+1}) - 2X(t_{i+1}, t_i)\mathbf{w}(t_i) + X(t_{i+1}, t_{i-1})\mathbf{w}(t_{i-1}) = \mathbf{v}_i - X(t_{i+1}, t_i)\mathbf{v}_{i-1}. \end{aligned} \tag{5.2.23}$$

Expanding the coefficient of $\mathbf{x}(t_1)$ in this equation up to terms of $O(\Delta t^2)$, and using

$$X(t,u) = I + (t-u)M(u) + \frac{(t-u)^2}{2}\left(M^2(u) + \frac{dM}{dt}(u)\right) + O(t-u)^3,$$

gives

$$\begin{aligned}
&V(t_{i+1}) - 2V(t_i) + V(t_{i-1}) - 2\left(\Delta t M(t_i) + \frac{\Delta t^2}{2}(M(t_i)^2 + \frac{dM}{dt}(t_i))\right)V(t_i) \\
&+ \left(2\Delta t(M(t_i) - \Delta t\frac{dM}{dt}(t_i)) + 2\Delta t^2\left(M(t_i)^2 + \frac{dM}{dt}(t_i)\right)\right)V(t_{i-1}) + O(\Delta t^3) \\
&= \Delta t^2\left(\frac{d^2V}{dt^2} - 2M(t_i)\frac{dV}{dt} + (M(t_i)^2 - \frac{dM}{dt}(t_i))V\right) + O(\Delta t^3).
\end{aligned}$$

There is an identical expression for $W(t)$. The corresponding expression for $\mathbf{w}(t)$ must take account of the inhomogeneous term. This requires expanding the term $\mathbf{v}_i - X(t_{i+1}, t_i)\mathbf{v}_{i-1}$ up to terms of $O(\Delta t^2)$ (the trapezoidal rule proves convenient), and gives

$$\begin{aligned}
&\mathbf{w}(t_{i+1}) - 2X(t_{i+1}, t_i)\mathbf{w}(t_i) + X(t_{i+1}, t_{i-1})\mathbf{w}(t_{i-1}) = \\
&\Delta t^2\left(\frac{d^2\mathbf{w}}{dt^2} - 2M(t_i)\frac{d\mathbf{w}}{dt} + (M(t_i)^2 - \frac{dM}{dt}(t_i))\mathbf{w} - \frac{d\mathbf{q}}{dt} + M(t_i)\mathbf{q}\right) + O(\Delta t^3).
\end{aligned}$$

The $O(\Delta t^2)$ terms correspond to second order differential systems, and these can be equated to zero as the boundary conditions (5.2.13) can now be satisfied. The equation satisfied by V at $t = t_i,\ i = 2, 3, \cdots, n-1$ is

$$\frac{d^2V}{dt^2} - 2M(t)\frac{dV}{dt} + (M(t)^2 - \frac{dM}{dt}(t))V = 0. \tag{5.2.24}$$

If the substitution $V(t) = X(t, \xi)\Phi(t)$ is made then (5.2.24) reduces to

$$X(t, \xi)\frac{d^2\Phi}{dt^2} = 0.$$

This is equivalent to the equation

$$\frac{d^2}{dt^2}\left(X^{-1}(t, \xi)V\right) = 0$$

so the solution satisfying the boundary conditions is

$$V = X(t, 0)(1 - t). \tag{5.2.25}$$

A similar calculation gives

$$W = X(t,1)t. \tag{5.2.26}$$

The corresponding equation for the particular integral is

$$\frac{d^2}{dt^2}\left(X^{-1}(t,\xi)\mathbf{w}\right) = X^{-1}\left(\frac{d\mathbf{q}}{dt} - M(t)\mathbf{q}\right),$$

and this has the solution satisfying the homogeneous boundary conditions

$$\mathbf{w} = \int_0^1 X(t,0)\mathcal{G}(t,u)X^{-1}(u,0)\left(\frac{d\mathbf{q}}{du} - M(u)\mathbf{q}\right)du \tag{5.2.27}$$

where $\mathcal{G}$ is the Green's function for the second order operator $\frac{d^2}{dt^2}$ with zero function boundary conditions.

$$\begin{aligned}\mathcal{G}(t,u) &= -(1-u)t,\ t<u,\\ &= -u(1-t),\ u<t.\end{aligned}$$

The constraint equation correponding to this choice of C has

$$G_i^1 = I,\ H_i^1 = -X(t_{i+1},t_{i-1}),\ \mathbf{d}_i^1 = \int_{t_{i-1}}^{t_{i+1}} X(t_{i+1},u)\mathbf{q}(u)du. \tag{5.2.28}$$

Note that V and W show the same growth as the fast (respectively) slow solutions of linear differential equation (equations (5.2.25) and (5.2.26)). This possibility of very large elements occurring is unattractive numerically. This realisation of the cyclic reduction process will be called the 'compactification' case [6] .

5.2.3 The orthogonal reduction

Growth in the computed $G_i^1, H_i^1, \mathbf{d}_i^1$ can be prevented by using orthogonal transformations in the cyclic reduction process as this preserves column lengths in the transformed matrix. Orthogonal matrices in the transformation family must satisfy

$$C^T R^T R C = I.$$

where R is given by (5.2.19). Thus

$$\begin{aligned}R^T R &= C^{-T}C^{-1},\\ &= \frac{1}{4}\begin{bmatrix} -X^{-T} & I\\ X^{-T} & I\end{bmatrix}\begin{bmatrix} -X^{-1} & X^{-1}\\ I & I\end{bmatrix},\\ &= \frac{1}{4}\begin{bmatrix} I+X^{-T}X^{-1} & I-X^{-T}X^{-1}\\ I-X^{-T}X^{-1} & I+X^{-T}X^{-1}\end{bmatrix}.\end{aligned}$$

It is convenient to consider the general form of transformation by revealing the freedom in the transformation explicitly . Let $R_1^{-1}R_{12} = S_1$. Then the $\mathbf{x}(t_1)$ component of the transformed system (5.2.23) corresponding to the transformation given by RC is

$$(I + S_1) V(t_{i+1}) - 2X(t_{i+1}, t_i)V(t_i) + (I - S_1)X(t_{i+1}, t_{i-1})V(t_{i-1})$$

This term can be $O(\Delta t^2)$ for smooth enough data only if $S_1 = O(\Delta t) = \Delta t S$. Also

$$\Delta t S\left(V(t_{i+1}) - X(t_{i+1}, t_{i-1})V(t_i)\right) = \Delta t S\left(2\Delta t \frac{dV}{dt}(t_i) - \Delta t M(t_i)V(t_i)\right) + O(\Delta t^3)$$

so $O(\Delta t)$ terms in expanding $S(t + \Delta t)$ can be ignored in determining the differential equations defining the interpolation operation. For example, the differential equation satisfied by V is

$$\frac{d^2V}{dt^2} + 2(S - M))\frac{dV}{dt} + \left(M^2 - 2SM - \frac{dM}{dt}\right)V = 0. \tag{5.2.29}$$

The inhomogeneous terms transform to

$$\begin{aligned}\mathbf{v}_i - X(t_{i+1}, t_i)\mathbf{v}_{i-1} + \Delta t S\left(\mathbf{v}_i + X(t_{i+1}, t_i)\mathbf{v}_{i-1}\right) \\ = \Delta t^2\left(\frac{d\mathbf{q}}{dt} - M\mathbf{q} + 2S\mathbf{q}\right) + O(\Delta t^3).\end{aligned}$$

To compute $S(t)$ corresponding to orthogonal transformation note that

$$\begin{aligned}\left(X(t_{i+1}, t_i)X^T(t_{i+1}, t_i)\right)^{-1} &= \left(I + \Delta t\left(M(t_i) + M^T(t_i)\right) + O(\Delta t^2)\right)^{-1}, \\ &= I - \Delta t\left(M(t_i) + M^T(t_i)\right) + O(\Delta t^2).\end{aligned}$$

Thus

$$R^T R = \frac{1}{4}\begin{bmatrix} 2I - \Delta t\left(M(t_i) + M^T(t_i)\right) & \Delta t\left(M(t_i) + M^T(t_i)\right) \\ \Delta t\left(M(t_i) + M^T(t_i)\right) & 2I - \Delta t\left(M(t_i) + M^T(t_i)\right)\end{bmatrix} + O(\Delta t^2).$$

Let

$$\frac{M(t_i) + M^T(t_i)}{2} = U^T + 2D + U,$$

where U is upper triangular with zero diagonal, and D is diagonal. Then

$$R = \frac{1}{\sqrt{2}}\begin{bmatrix} I - \Delta t\left(\frac{D+U}{2}\right) & \frac{\Delta t\left(M(t_i) + M^T(t_i)\right)}{2} \\ & I - \Delta t\left(\frac{D+U}{2}\right)\end{bmatrix} + O(\Delta t^2),$$

so that, as $S_1 = R_1^{-1} R_{12}$,

$$S(t_i) = \frac{M(t_i) + M^T(t_i)}{2}. \tag{5.2.30}$$

Substituting in (5.2.29) gives

$$\frac{d^2V}{dt^2} + \left(M^T - M\right)\frac{dV}{dt} - \left(M^T M + \frac{dM}{dt}\right)V = 0. \tag{5.2.31}$$

Again there is a corresponding equation for W. The equation for the particular integral term is

$$\frac{d^2\mathbf{w}}{dt^2} + \left(M^T - M\right)\frac{d\mathbf{w}}{dt} - \left(M^T M + \frac{dM}{dt}\right)\mathbf{w} = \frac{d\mathbf{q}}{dt} + M^T\mathbf{q}. \tag{5.2.32}$$

To solve for V note that (5.2.31) factorizes (the order must be respected) to give

$$\left(\frac{d}{dt} + M^T\right)\left(\frac{d}{dt} - M\right)V = 0.$$

Making the standard variation of parameters substitution $V = XY$ gives

$$\left(\frac{d}{dt} + M^T\right)X\frac{dY}{dt} = 0.$$

It is straightforward to verify that this equation has a fundamental matrix given by X^{-T}. It follows that

$$\frac{dY}{dt} = X^{-1}(t,0)\,X^{-T}(t,0)\,\frac{dY}{dt}(0), \text{ and}$$

$$V = \left\{\int_t^1 X(t,u)\,X^{-T}(u,0)\,du\right\}\left\{\int_0^1 X^{-1}(u,0)\,X^{-T}(u,0)\,du\right\}^{-1}.$$

Remark 5.2.2 *The use of a limiting argument in which $\Delta t \to 0$ in deriving the equations for V, W, $\mathbf{w}$ could suggest that the dependence on the two parameters t, Δt should be acknowledged explicitly by the adoption of a notation such as $V(t, \Delta t)$. However, this is not necessary here. This follows because the discretization (5.2.8) is satisfied exactly by (5.2.11) for all Δt, and not just in a limiting sense. The notation acknowledging dependence on both t and Δt is required for solutions of approximate discretizations such as that which would be obtained by using the trapezoidal rule to integrate (5.2.1). In such cases solutions given as linear combinations of $V(t, \Delta t)$, $W(t, \Delta t)$, $\mathbf{w}(t, \Delta t)$ can be analyzed using the methods of difference (defect) correction [32]. The solution given here then corresponds to $V(t, 0)$, $W(t, 0)$, $\mathbf{w}(t, 0)$.*

5.2.4 Interpretation of the constraint equation

As the combination of the influence matrices $Z = V(t) + W(t)\, X(t_n, t_1)$ must satisfy the homogeneous differential equation (5.2.1), while the individual components must satisfy boundary conditions appropriate to a second order system, it follows that the simplest form for the determining equations is

$$L_1(K)\left(\frac{d}{dt} - M\right)\left\{\begin{array}{c} V(t) \\ W(t) \end{array}\right. = 0, \text{ and} \tag{5.2.33}$$

$$L_1(K)\left(\frac{d}{dt} - M\right)\mathbf{w}(t) = L_1(K)\mathbf{q}. \tag{5.2.34}$$

where L_1 is the differential operator

$$L_1(K) = \frac{d}{dt} + K(t)$$

and the correspondence with the transforming matrix (5.2.19) is given by

$$K = 2S - M,\ \Delta t S = R_1^{-1} R_{12}.$$

The compactification case corresponds to $S = 0$, while orthogonal reduction corresponds to $S = \left(M + M^T\right)/2$.

Some insight into the different stability characteristics of the compactification and orthogonal reduction cases can be obtained by considering the first order systems associated with (5.2.33), (5.2.34). The first order system here can be written

$$\begin{aligned} \frac{d}{dt}\left[\begin{array}{c} Y \\ Z \end{array}\right] &= N\left[\begin{array}{c} Y \\ Z \end{array}\right] \\ \frac{d}{dt}\left[\begin{array}{c} \mathbf{w} \\ \mathbf{z} \end{array}\right] &= N\left[\begin{array}{c} \mathbf{w} \\ \mathbf{z} \end{array}\right] + \left[\begin{array}{c} \mathbf{q} \\ 0 \end{array}\right], \end{aligned} \tag{5.2.35}$$

with matrix

$$N = \left[\begin{array}{cc} M & I \\ & -(2S - M) \end{array}\right].$$

Here Y can be either V or W with corresponding terms Z_V, Z_W completing the solution vector.

The role of the constraint equation can now be identified. The solution $\mathbf{x}(t)$ constructed using the cyclic reduction derived quantities V, W, $\mathbf{w}$ satisfies the higher order equation (5.2.35) and so potentially depends on a larger set of fundamental solutions. The function of the constraint equations is to remove this unwanted generality. Let

$$\mathbf{x} = V\mathbf{x}(t_1) + W\mathbf{x}(t_n) + \mathbf{w}.$$

Then the requirement that $\mathbf{x}$ satisfy the lower order system (5.2.1) is, using (5.2.35),

$$\begin{aligned} 0 &= \frac{d\mathbf{x}}{dt} - M\mathbf{x} - \mathbf{q}, \\ &= \left(\frac{dV}{dt} - MV\right)\mathbf{x}(t_1) + \left(\frac{dW}{dt} - MW\right)\mathbf{x}(t_n) + \frac{d\mathbf{w}}{dt} - M\mathbf{w} - \mathbf{q}, \\ &= Z_V(t)\mathbf{x}(t_1) + Z_W(t)\mathbf{x}(t_n) + \mathbf{z}(t). \end{aligned} \tag{5.2.36}$$

This form of constraint depends on t, but this dependence does not contain independent information as a change in t dependence (for example, from t_i to t_j) is achieved by premultiplying each term by $Z(t_j, t_i)$, a fundamental matrix for the Z dependence

$$L_1(K)\, Z(t, \xi) = 0,\ Z(\xi, \xi) = I.$$

The point to note here is that the terms Z_V, Z_W, $\mathbf{z}$ satisfy the same homogeneous equation. A form independent of t can be obtained by integration, for example. This gives

$$\left\{\int_{t_1}^{t_n} Z_V dt\right\}\mathbf{x}(t_1) + \left\{\int_{t_1}^{t_n} Z_W dt\right\}\mathbf{x}(t_n) = -\int_{t_1}^{t_n} \mathbf{z} dt.$$

Exercise 5.2.3 *The constraint equation (5.2.36) relates* $\mathbf{x}(t_1)$ *and* $\mathbf{x}(t_n)$. *Show that this equation is equivalent to (5.2.6).*

Exercise 5.2.4 *Verify that the linear system (5.2.35) is equivalent to (5.2.34).*

5.2.5 Dichotomy and stability

A generic starting point in our analysis of estimation problems is the assumption that the problem under consideration has a well determined solution, at least when correctly formulated. Aspects of these considerations which are relevant here include:

1. The estimation problem may have a well determined solution but is this true also of the boundary value embedding formulation ?

2. An explicit differential equation solution procedure is not an explicit part of the simultaneous method problem statement. Does this indicate possibly more robust solution properties?

3. Let the boundary value formulation of the differential equation problem have a well determined solution. Does it follow that the ($\mathbf{w}_i$, V_i, W_i) determined by the cyclic reduction algorithms is also well determined. This is suggested in the orthogonal case by the connection with the stable orthogonal factorization and back substitution algorithm for solving linear equations. Note that these quantities are independent of the boundary conditions on the original differential equation. It follows that stability or otherwise of these augmented equations is a generic property of the original differential equation only and does not depend on its boundary conditions.

The stability analysis of the initial value problem

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(t, \mathbf{x}), \quad \mathbf{x}(0) = \mathbf{b}.$$

is classic. It is required here that solutions with close initial conditions $\mathbf{x}_1(0)$, $\mathbf{x}_2(0)$ remain close in an appropriate sense for large t. For example:

1. Strong (initial value) stability requires $\|\mathbf{x}_1(t) - \mathbf{x}_2(t)\| \to 0$, $t \to \infty$.

2. Weak (initial value) stability requires $\|\mathbf{x}_1(t) - \mathbf{x}_2(t)\|$ remain bounded as $t \to \infty$. This property is most useful for linear systems. Chaos provides an example of bounded families possessing unbounded first variations. In most circumstances the initial value behaviour of such families must be considered as unstable from a computational point of view.

Numerical considerations introduce the concept of stiff discretizations. These are discretizations which inherit at least a form of the weak stability characteristics of the original problem. In the linear case this corresponds to a mapping of nonincreasing solutions of the differential equation onto nonincreasing solutions of the difference approximation. Note that very rapidly decreasing solutions of the differential equation cannot carry significant solution information forward, but would require a fine mesh to follow them accurately. Thus it is desirable that a mesh better adapted to follow the actual solution effectively should be used. In this context use of stiffly stable discretizations is appropriate [18]. However, initial value computations are not limited to initial value stable problems. For example, computing fundamental matrices is important in multiple shooting calculations. It is possible to compute these to satisfactory accuracy for relatively unstable problems over short enough time intervals by taking the discretization interval Δt small enough and using

methods consistent and stable in the sense of Dahlquist [21]. As indicated in the introduction to this chapter, attempting to resolve all components of a fundamental matrix accurately when the initial value formulation is unstable can have similar problems to using a non-stiff integrator to solve a stiff initial value problem. In the multiple shooting case the requirement is that the breakpoints be chosen sufficiently close together for the computed matrix to be an adequate approximation to the exact multiple shooting matrix [75].

Example 5.2.2 *Differential equations with constant coefficients. Let*

$$\mathbf{f}(t, \mathbf{x}) = M\mathbf{x} - \mathbf{q}.$$

If M is non-defective then weak stability requires that the eigenvalues $\lambda_i(M)$ satisfy $Re\{\lambda_i\} \leq 0,\ i = 1, 2, \cdots, m$. These inequalities must be strict for strong stability. Now consider the one step discretization

$$\mathbf{x}_{i+1} = T_\Delta(M)\,\mathbf{x}_i + \mathbf{v}_i.$$

Here $T_\Delta(M)$ is the amplification matrix, and Δt is the discretization interval. The condition for a stiff discretization is that

$$Re\{\lambda_i(M)\} \leq 0 \Rightarrow |\lambda_i(T_\Delta)| \leq 1, \quad i = 1, 2, \cdots, m.$$

For the trapezoidal rule

$$\begin{aligned} |\lambda_i(T_\Delta)| &= \left| \frac{1 + \Delta t \lambda_i(M)/2}{1 - \Delta t \lambda_i(M)/2} \right|, \\ &\leq 1 \text{ if } Re\{\lambda_i(M)\} \leq 0 \end{aligned}$$

The stability discussion for the constant coefficient case does not generalise too easily in the sense of providing a readily computed stability criterion for the case where M is a function of t. Some partial results can be based on the concept of kinematic eigenvalues [6]. A more general tool is provided by the "logarithmic norm" [101]

$$\mu(A) = \lim_{h \to 0+} \frac{\|I + hA\| - 1}{h}, \tag{5.2.37}$$

where $\| \bullet \|$ is a subordinate matrix norm. Because $\mu(A)$ can be negative it is clearly not a norm. An important example that illustrates this point corresponds to the case of constant M with negative eigenvalues where the spectral norm is used in the computation of $\mu(M)$. A key result is the differential inequality

$$\frac{d\|\mathbf{x}\|}{dt^+} \leq \mu(M)\,\|\mathbf{x}\|, \tag{5.2.38}$$

where the derivative is the upper right Dini derivative

$$\frac{d\psi}{dt^+} = \lim_{h\to 0+} sup \frac{\psi(t+h) - \psi(t)}{h}. \tag{5.2.39}$$

Here $\mu(M)$ can be a function of t. The logarithmic norm can be extended to unbounded and nonlinear operators.

The spirit of the above discussion is provided by a concern for problems which are dominated by considerations of exponential growth and decay. This works well enough for linear systems. However, it is an incomplete picture when attempting to analyse the stability of a nonlinear system by studying the stability of a local linear approximation. Here the local linearisation can predict instability in the sense of arbitrarily large departures Δ_∞ of initially close Δ_0 trajectories when the nonlinear system supports attractive regions which capture solution trajectories typically referred to as chaotic . The instability of the nonlinear initial value problem in this case is one of phase rather than amplitude. To classify this behaviour let $X(t)$ be a fundamental matrix of the linearised differential equation, and define

$$\Lambda = \lim_{t\to\infty} \left(X^T X\right)^{1/2t}. \tag{5.2.40}$$

Then

$$\lambda_1 = \log\left(\|\Lambda\|\right), \tag{5.2.41}$$

where $\| \bullet \|$ is the spectral norm, is called the first Lyapunov exponent . A positive value of λ_1 is commonly used as an indicator of chaotic situations. The logarithms of the other eigenvalues of Λ give higher Lyapunov exponents and these provide further structural information on the properties of the attractor. For a discussion in a computational context see [20].

Boundary value stability introduces rather different considerations. *It could be expected to be more relevant to the estimation problem because now the information determining the parameter estimates generically is distributed in a manner more resembling the setting of multi-point boundary values*. Here the problem is

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(t, \mathbf{x}), \quad B(\mathbf{x}) = B_1\mathbf{x}(0) + B_2\mathbf{x}(1) = \mathbf{b}.$$

Behaviour of perturbations about a solution trajectory $\mathbf{x} = \mathbf{x}^*(t)$ is governed to first order by the linearised equation

$$L(\mathbf{z}) = \frac{d\mathbf{z}}{dt} - \nabla_x \mathbf{f}(t, \mathbf{x}^*(t))\,\mathbf{z} = 0. \tag{5.2.42}$$

Here (computational) stability is closely related to the existence of a (modest) bound for the Green's matrix :

$$\mathcal{G}(t,s) = Z(t)\,B(Z)^{-1}B_1 Z(0)\,Z(s)^{-1}, \quad t > s, \tag{5.2.43}$$

$$= -Z(t)\,B(Z)^{-1}B_2 Z(1)\,Z(s)^{-1}, \quad t < s, \tag{5.2.44}$$

where $Z(t)$ is a fundamental matrix for the differential equation linearised about the trajectory described by $\mathbf{x}^*$, and

$$B(Z) = B_1 Z(0) + B_2 Z(1).$$

Definition 5.1 *Let*

$$\alpha = \sup_{0\le s,t\le 1} \|\mathcal{G}(t,s)\|_\infty . \tag{5.2.45}$$

Then α is called the stability constant *. In this context, modest α means that small perturbations in the problem data lead to small perturbations in the resultant solution.*

Stability of the linear boundary value problem (5.2.42) expressed by (5.2.45) is closely linked to the property of dichotomy .

Definition 5.2 *: Equation (5.2.42) possesses a* (strong) dichotomy *if there exists a constant projection P, depending on the choice of Z, such that, given the splitting of the solution space*

$$S_1 \leftarrow \{ZP\mathbf{w},\ \mathbf{w}\in R^m\}, \quad S_2 \leftarrow \{Z(I-P)\mathbf{w},\ \mathbf{w}\in R^m\}, \tag{5.2.46}$$

there exists $\kappa > 0$ such that

$$\boldsymbol{\phi}\in S_1 \Rightarrow \frac{\|\boldsymbol{\phi}(t)\|_\infty}{\|\boldsymbol{\phi}(s)\|_\infty} \le \kappa, \quad t \ge s,$$

$$\boldsymbol{\phi}\in S_2 \Rightarrow \frac{\|\boldsymbol{\phi}(t)\|_\infty}{\|\boldsymbol{\phi}(s)\|_\infty} \le \kappa, \quad t \le s.$$

Note that such a κ always exists for $t, s \in [0,1]$. Computational interest is in relatively small values of κ given such a range restriction.

Remark 5.2.3 *The significance of dichotomy in the computational context is a consequence of a result of de Hoog and Mattheij [22]. They showed that if the boundary conditions are separated so that*

$$\text{rank}(B_1) = r, \quad \text{rank}(B_2) = m - r,$$

and Z is chosen such that $B(Z) = I$ then $P = B_1 Z(0)$ is a projection and it is possible to choose $\kappa = \alpha$. A weaker, but similar relation was shown if the boundary conditions are not separated. It follows that there is a basic connection between stability and dichotomy for linear boundary value problems.

The significance of dichotomy is that it provides a global labelling on $0 \leq t \leq 1$ of the rapidly increasing solutions $\phi \in S_2$ and the rapidly decreasing solutions $\phi \in S_1$. This classification is fuzzy at the edges for solutions which do not lie in either extreme category, and it flags as possibly dangerous solutions which would try to flip between the categories. This classification has another important interpretation. It means that boundary control at $t = 1$ is important to ensure rapidly increasing solution components from S_2 do not dominate as lack of such control could allow a large perturbation about $\mathbf{x}^*(t)$. In similar fashion rapidly decreasing solution components in S_1 must be pinned down at $t = 0$.

Dichotomy together with compatible boundary conditions provides the boundary value problem property which is analogous to the stability requirement important in the solution of initial value problems.

Example 5.2.3 *To illustrate the importance of compatible boundary conditions consider the Mattheij system of differential equations [6] given by*

$$M = \begin{bmatrix} 1 - 19\cos 2t & 0 & 1 + 19\sin 2t \\ 0 & 19 & 0 \\ -1 + 19\sin 2t & 0 & 1 + 19\cos 2t \end{bmatrix}, \tag{5.2.47}$$

$$\mathbf{q} = \begin{bmatrix} e^t\left(-1 + 19\left(\cos 2t - \sin 2t\right)\right) \\ -18e^t \\ e^t\left(1 - 19\left(\cos 2t + \sin 2t\right)\right) \end{bmatrix}. \tag{5.2.48}$$

Here the right hand side is chosen so that the slowly varying functions $\mathbf{x}(t) = e^t\mathbf{e}$ satisfy the differential equations. The fundamental matrix is

$$X(t,0) = \begin{bmatrix} e^{-18t}\cos t & 0 & e^{20t}\sin t \\ 0 & e^{19t} & 0 \\ -e^{-18t}\sin t & 0 & e^{20t}\cos t \end{bmatrix}.$$

It is characterised by rapidly varying fast and slow solutions determined by the terms e^{19t}, e^{20t}, and e^{-18t}. Because the fast and slow solutions are already separated it is straightforward to verify that the system supports a dichotomy – even one with exponentially decreasing bounds – with $P = \mathbf{e}_2\mathbf{e}_2^T + \mathbf{e}_3\mathbf{e}_3^T$. For boundary data which gives two terminal conditions and one initial condition :

$$B_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \; B_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; \mathbf{b} = \begin{bmatrix} e \\ e \\ 1 \end{bmatrix},$$

the solution obtained using a trapezoidal rule discretization scheme and standard orthogonal factorization of the linear equations gives the results displayed in Table 5.1. These computations are apparently satisfactory. However, there

	$\Delta t = .1$			$\Delta t = .01$		
$\mathbf{x}(0)$	1.0000	.9999	.9999	1.0000	1.0000	1.0000
$\mathbf{x}(1)$	2.7183	2.7183	2.7183	2.7183	2.7183	2.7183

Table 5.1: Boundary point values - stable computation

is an interesting twist which becomes more apparent in the next case which considers unstable conditions. For boundary data with two initial and one terminal condition

$$B_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \; B_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; \mathbf{b} = \begin{bmatrix} 1 \\ e \\ 1 \end{bmatrix},$$

the results are given in Table 5.2. They indicate clearly an unstable behaviour in the first and third equations. The second equation is uncoupled from the other two, and for it the terminal condition used is appropriate. One source of the problems revealed would be the instability caused by the use of initial conditions on an unstable problem. However, if this were the single cause then it is somewhat surprising that the erroneous terms are larger in the problem with larger Δt. Something else is going on here, and it has to do with the breakdown of the trapezoidal rule discretization. Consider the simple equation

$$\frac{dx}{dt} = \lambda x.$$

The trapezoidal rule discretization gives

$$\left(1 - \frac{\Delta t \lambda}{2}\right) x_{i+1} - \left(1 + \frac{\Delta t \lambda}{2}\right) x_i = 0.$$

If $\lambda \leq 0$ then the marching procedure giving x_{i+1} in terms of x_i is always stable in the sense that the amplification factor is always less equal 1 in modulus. However, if $\lambda > 0$ then the recurrence breaks down when $\lambda = 2/\Delta t$, and produces an oscillating result if $\lambda > 2/\Delta t$ with amplification factor tending to -1 when $\Delta t \lambda$ is large (so called super stability) . This is not mirrored exactly here, but the first and third equations are coupled and have solutions which include terms proportional to e^{20t} which correspond exactly to the value of λ which causes the super stability blow up in the initial value problem in the above simple model when $\Delta t = .1$. Of at least as much interest is the

apparently stable computation associated with the second (uncoupled) equation. Posed stably by virtue of a terminal condition, it manages to extract the relatively slowly varying solution information despite the trapezoidal rule recurrence giving an amplification approximately 10 *times larger than the exact figure* $e^{20\Delta t}$ *for solutions of the homogeneous equation. The computations on the coarser grid provide interesting evidence of a form of stiff stability available when the differential system supports nontrivial dichotomy. This property of the trapezoidal rule is identified for systems of constant coefficients in [27]. A sufficient condition is that the eigenvalues* μ_i *of the amplification matrix satisfy*

$$|\lambda_i| < 0 \Rightarrow |\mu_i| < 1,$$
$$|\lambda_i| > 0 \Rightarrow |\mu_i| > 1,$$

for solutions of the differential equation that are proportional to $\exp \lambda_i t$, $i = 1, 2, \cdots, m$. *This requires that the dichotomy is preserved in a weak sense; and the authors introduce the term* di-stability *for this property. The trapezoidal rule is di-stable for constant coefficient equations. This property is verified readily for the above recurrence which gives*

$$|\mu| = \frac{\left|\left(1 + \frac{\Delta t \lambda}{2}\right)\right|}{\left|\left(1 - \frac{\Delta t \lambda}{2}\right)\right|}$$

The key observation is that if increasing solutions are mapped into solutions that increase in magnitude then they are still controlled by boundary conditions that are compatible with the dichotomy. This has the effect of continuing to suppress any contribution to a slowly varying solution arising from the effects of discretization error in the rapidly changing terms. Note that it is magnitude not sign of the difference equation amplification factors which is important for this purpose.

	$\Delta t = .1$			$\Delta t = .01$		
$\mathbf{x}(0)$	1.0000	.9999	1.0000	1.0000	1.0000	1.0000
$\mathbf{x}(1)$	-7.9197+11	2.7183	-4.7963+11	2.0369+2	2.7183	1.3169+2

Table 5.2: Boundary point values - unstable computation

The "natural" boundary matrices for the interesting case $\Delta t = .1$ calculated using the procedure developed in subsection 5.4.2 are given in Table 5.3. These bear out the importance of weighting the boundary data to reflect the stability requirements of a mix of fast and slow solutions.They give solution values identical with those reported in Table 5.1.

B_1			B_2		
.99955	0.0000	.02126	-.01819	0.0000	-.01102
0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
.02126	0.0000	.00045	.85517	0.0000	.51791

Table 5.3: natural boundary matrices in case $\Delta t = .1, \rho = 1$

Useful information on the stability of the differential equations associated with the computation of the cyclic reduction quantities $\mathbf{w}$, V, W can be obtained by considering the case M constant and diagonalizable by similarity transformation ($M = T\Lambda T^{-1}$).

Lemma 5.1 *Let $S = 0$. Then $N : R^{2m} \to R^{2m}$ is similar to a matrix with m Jordan blocks $J_i = \begin{bmatrix} \lambda_i & 1 \\ & \lambda_i \end{bmatrix}$, $i = 1, 2, \cdots, m$.*

Proof.

$$N = \begin{bmatrix} M & I \\ & M \end{bmatrix} = \begin{bmatrix} T & \\ & T \end{bmatrix} \begin{bmatrix} \Lambda & I \\ & \Lambda \end{bmatrix} \begin{bmatrix} T^{-1} & \\ & T^{-1} \end{bmatrix}.$$

It can now be cast in the desired form by making symmetric row and column permutations of the eigenvalue matrix without destroying the property of similarity. ■■

Lemma 5.2 *Let $S = \left(M + M^T\right)/2$ and assume that M is non-defective. Then corresponding to each eigenvalue λ of M for which $-\lambda$ is not also an eigenvalue is a pair $\pm\lambda$ of eigenvalues of N. If $\pm\lambda$ are both eigenvalues of M then both correspond to (2×2) Jordan blocks in the similarity normal form of N.*

Proof.

$$\begin{aligned} \det(N - \lambda I) &= \det\left(\begin{bmatrix} M - \lambda I & I \\ & -\left(M^T + \lambda I\right) \end{bmatrix}\right) \\ &= (-1)^m \det(M - \lambda I) \det\left(M^T + \lambda I\right). \end{aligned}$$

showing that the eigenvalues occur in $\pm$ pairs. The eigenvector corresponding to the eigenvalue λ_i is $\left[\mathbf{t}_i^T, 0\right]^T$ where $\mathbf{t}_i$ is the corresponding right eigenvector of M. Let $\mathbf{s}_i$ be the left eigenvector of M corresponding to λ_i.Then the eigenvector $\mathbf{v}$ of N corresponding to the eigenvalue $\lambda = -\lambda_i$ is

$$\mathbf{v} = \left[\left((M + \lambda_i I)^{-1} \mathbf{s}_i\right)^T, -\mathbf{s}_i^T\right]^T$$

unless $-\lambda_i$ is also an eigenvalue of M. If M is non-defective and has a $\pm$ pair of eigenvalues then it is necessary to allow N to have a principal vector of grade 2 associated with each eigenvalue. To calculate this when $\lambda = \lambda_i$ set

$$(N - \lambda_i I)^2 = \begin{bmatrix} (M - \lambda_i I)^2 & M - M^T - 2\lambda_i I \\ 0 & \left(M^T + \lambda_i I\right)^2 \end{bmatrix}.$$

Because M is non-defective it follows that the grade 2 vector will have the form $\begin{bmatrix} \mathbf{v}_i \\ \mathbf{u}_i \end{bmatrix}$ where $\mathbf{u}_i$ satisfies

$$\left(M^T + \lambda_i I\right) \mathbf{u}_i = 0.$$

Now $\mathbf{v}_i$ must satisfy

$$(M - \lambda_i I)^2 \mathbf{v}_i + (M - \lambda_i I) \mathbf{u}_i = 0.$$

This is equivalent to the system

$$(M - \lambda_i I) \mathbf{v}_i + \mathbf{u}_i = \gamma \mathbf{t}_i$$

where γ must be chosen such that the singular system is consistent. This requires

$$\mathbf{s}_i^T (\mathbf{u}_i - \gamma \mathbf{t}_i) = 0.$$

It is now routine to compute $\mathbf{v}$. To complete the lemma note that the existence of principal vectors of grade 2 implies the existence of 2×2 Jordan blocks. ■■

Example 5.2.4 *The simplest instance of the occurence of principal vectors of grade 2 when M has a pair $\pm\lambda$ of eigenvalues corresponds to*

$$M = \begin{bmatrix} 0 & 1 \\ \mu^2 & 0 \end{bmatrix}, \; N - \mu I = \begin{bmatrix} -\mu & 1 & 1 & 0 \\ \mu^2 & -\mu & 0 & 1 \\ 0 & 0 & -\mu & -\mu^2 \\ 0 & 0 & -1 & -\mu \end{bmatrix}. \tag{5.2.49}$$

Here N has an eigenvector $[1, \mu, 0, 0]^T$ associated with the eigenvalue μ, but no eigenvector of the form $[x, x, \mu, -1]^T$. The equation determining $\mathbf{v}$ is

$$\begin{bmatrix} -\mu & 1 \\ \mu^2 & -\mu \end{bmatrix} \mathbf{v} + \begin{bmatrix} \mu \\ -1 \end{bmatrix} = \gamma \begin{bmatrix} 1 \\ \mu \end{bmatrix}.$$

It follows that

$$\gamma = \frac{\mu^2 - 1}{2\mu}, \quad \mathbf{v} = \frac{\mu^2 + 1}{2\mu\,(\mu + 1)} \begin{bmatrix} -1 \\ \mu \end{bmatrix}.$$

Remark 5.2.4 *The problem with the compactification case is evident from these results. Assume that M has no zero or pure imaginary eigenvalues. Then corresponding to each Jordan block in N are new solutions of the form polynomial times exponential. These are of the same type (fast or slow) as the exponential terms. This follows because the characteristic powers of t which appear in the new solutions multiplying an exponential term can be ignored for the purposes of this classification. If the number of fast solutions differs from the number of slow solutions then the possibility of satisfying the boundary conditions (5.2.13) on both V and W in a stable fashion appears unlikely. For example, let (5.2.1) be stable so that all solutions are slow. In this case computation of V, W is most likely difficult as only half the slow solutions can be pinned down at $t = 0$. In contrast, the condition that there be equal numbers of fast and slow solutions is automatically satisfied in the orthogonal reduction case.*

Numerical evidence of this numerical stability of the orthogonal factorization system of differential equations for dichotomous systems can be presented using the Mattheij example. This problem (5.2.47), (5.2.48) shows significant potential for highlighting instability. Another problem which presents similar difficulties [6] has data

$$M = \begin{bmatrix} -10\cos 20t & 10 + 10\sin 20t \\ -10 + 10\sin 20t & 10\cos 20t \end{bmatrix},$$
$$\mathbf{q} = \begin{bmatrix} e^t(-9 + 10(\cos 20t - \sin 20t)) \\ e^t(11 - 10(\cos 20t + \sin 20t)) \end{bmatrix}.$$

This problem has the fundamental matrix

$$X(t,0) = \begin{bmatrix} \cos 10t & \sin 10t \\ -\sin 10t & \cos 10t \end{bmatrix} \begin{bmatrix} e^{10t} & 0 \\ 0 & e^{-10t} \end{bmatrix}.$$

To check stability, the particular integral $\mathbf{w}(t)$ has been computed for both problems (see (5.2.34)). In the first case

$$L_1(K)\mathbf{q} = \begin{bmatrix} e^t(-364 + 38(\cos 2t - \sin 2t)) \\ -360e^t \\ e^t(-362 - 38(\cos 2t + \sin 2t)) \end{bmatrix},$$

while in the second

$$L_1(K)\mathbf{q} = \begin{bmatrix} -219e^t \\ -179e^t \end{bmatrix}.$$

The first order system (5.2.35) for the differential equation (5.2.32) has been integrated using the standard midpoint rule (box scheme) and the solution

values at the ends of the interval of integration ($(0, \pi)$ in the first example, $(0, 1)$ in the second) which are not fixed by the boundary conditions (5.2.13) are tabulated for numbers of mesh points $n = 2^k,\ k = 6, 8, 10, 12, 14$.

k	6	8	10	12	14
$\mathbf{w}(0)_4$	20.304	20.478	20.489	20.490	20.490
$\mathbf{w}(0)_5$	19.994	20.000	20.000	20.000	20.000
$\mathbf{w}(0)_6$	-15.595	-15.876	-15.894	-15.895	-15.895
$\mathbf{w}(\pi)_4$	-366.22	-370.12	-370.36	-370.38	-370.38
$\mathbf{w}(\pi)_5$	-416.41	-416.52	-416.53	-416.53	-416.53
$\mathbf{w}(\pi)_6$	373.39	377.02	377.25	377.27	377.27
$\mathbf{w}(0)_3$	1.0449	1.0035	1.0009	1.0007	1.0007
$\mathbf{w}(0)_4$	20.957	21.001	21.003	21.003	21.003
$\mathbf{w}(1)_3$	-51.694	-51.652	-51.649	-51.649	-51.649
$\mathbf{w}(1)_4$	2.5846	2.7097	2.7175	2.7180	2.7180

Examples of particular integral computations.

The general trend of the results indicate stable computations. In fact the terminal values were among the largest recorded, and these are perfectly compatible with the magnitudes of the coefficients in the differential systems.

Exercise 5.2.5 *Show that the Green's matrix (5.2.43), (5.2.44) has a unit jump when $t = s$. What is the significance of this result for the differential equation satisfied by the Green's matrix.*

Exercise 5.2.6 *Use a sketch to explain why rapidly increasing solutions of a linear differential equation should be tied down by terminal boundary conditions while rapidly decreasing solutions need to be tied down by initial conditions if the boundary problem is to have a stably computable solution.*

5.3 Nonlinear differential equations

An important computation associated with the embedding approach to the parameter estimation problem is the explicit solution of the nonlinear differential equation (5.1.1) for a given parameter vector and given boundary values. This problem is written

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}\left(t, \mathbf{x}\right), \quad B\left(\mathbf{x}\right) = \mathbf{b}. \tag{5.3.1}$$

The form of the boundary conditions has to be selected in the embedding method. Here it is assumed that the boundary operator $B(\mathbf{x})$ has the form

$$B(\mathbf{x}) = B_1\mathbf{x}(0) + B_2\mathbf{x}(1). \tag{5.3.2}$$

It is further assumed that the system (5.3.1) has a well determined solution $\mathbf{x}^*(t)$ which satisfies the Kantorovich conditions for the application of Newton's method.

Stability in nonlinear problems becomes a property of the linear problem governing the behaviour of perturbations about a current trajectory. In this sense it is a local property. Easy nonlinear problems are associated with relatively slow perturbation growth. Such problems can be expected to have the property that Newton's method applied to solve the discretized problem will have a reasonable domain of convergence. The key property here is the connection between the Jacobian matrix in the solution of the discretised nonlinear BVP and the discretized linearization of the BVP about the current trajectory. This is explored in the Appendix to this chapter. Linear IVP/BVP stability requirements discussed in the previous section are inflexible in the sense that emphasis on dichotomy means that solutions must not depart too far from the classification as increasing/decreasing in order to guarantee a moderate stability constant. Such a departure signals that the characteristic stability property of being capable of following a slowly varying solution with an appropriately coarse mesh need not hold if the linearised equation has rapidly varying solutions. Important conflicting examples occur in the linearised equations that can be associated with dynamical system trajectories. These include solution trajectories which have the properties:

- they can have a stable character – for example, limiting trajectories which attract neighbouring orbits; and
- they have linearised systems which switch between increasing and decreasing modes in a manner characteristic of oscillatory behaviour . If this switch is rapid then it could be difficult to satisfy the dichotomy partitioning inequalities with a modest bound. This is likely to make more difficult the solution of the nonlinear problem by Newton's method.

The approach followed here begins by first discretizing the differential equation to obtain an approximating system of nonlinear equations. For example, using trapezoidal rule integration on the mesh considered in (5.1.15) gives

$$\mathbf{x}_{i+1} - \mathbf{x}_i = \frac{\Delta t_i}{2}\left(\mathbf{f}(t_{i+1}, \mathbf{x}_{i+1}, \boldsymbol{\beta}) + \mathbf{f}(t_i, \mathbf{x}_i, \boldsymbol{\beta})\right),\ i = 1, 2, \cdots, n-1. \tag{5.3.3}$$

Augmenting these equations with the boundary equation, which is here assumed to be linear, leads to the nonlinear system

$$\mathbf{F}(\mathbf{x}_c) = 0, \quad B(\mathbf{x}_c) = \mathbf{b} \tag{5.3.4}$$

where $\mathbf{x}_c$ is the composite vector

$$\mathbf{x}_c = \left[\mathbf{x}_1^T, \cdots, \mathbf{x}_n^T\right]^T. \tag{5.3.5}$$

This system use can be solved by the following steps:

1 Newton step - solve for $\mathbf{h}_c^*$

$$\mathbf{F}'(\mathbf{x}_c)\,\mathbf{h}_c^* = -\mathbf{F}(\mathbf{x}_c), \quad B(\mathbf{h}_c^*) = \mathbf{b} - B(\mathbf{x}_c),$$

where $\mathbf{F}'$ is the variational operator (Jacobian of the nonlinear system) derived from $\mathbf{F}$. It is assumed that B is linear otherwise it too must be linearized.

2 Perform a line search on an appropriate objective $\Phi(\lambda)$. Note that in contrast to the problem of maximising a likelihood the problem here is typically a descent computation so that the aim is to seek lower values in the linesearch step. Relevant cases include:

Residual sum squares To balance differential equation and boundary residuals note that individual differential equation residuals are typically $O\left(\frac{1}{n}\right)$ while the boundary residuals are typically $O(1)$.

$$\Phi(\lambda) = \|F(\mathbf{x}_c + \lambda\mathbf{h}_c^*)\|^2 + \|\mathbf{b} - B(\mathbf{x}_c + \lambda\mathbf{h}_c^*)\|^2. \tag{5.3.6}$$

Affine invariant objective Here

$$\Phi(\lambda) = \|\mathbf{h}_c\|^2, \tag{5.3.7}$$

where $\mathbf{h}_c$ satisfies the system of equations :

$$\mathbf{F}'(\mathbf{x}_c)\,\mathbf{h}_c = -\mathbf{F}(\mathbf{x}_c + \lambda\mathbf{h}_c^*), \quad B(\mathbf{h}_c) = \mathbf{b} - B(\mathbf{x}_c + \lambda\mathbf{h}_c^*).$$

The (remote) possibility of the iteration cycling when this strategy is used has been noted (4.2.9). However, $\mathbf{h}_c$ is a descent vector for the affine invariant objective (5.3.7) in this case.

Let the linesearch terminate with $\lambda = \lambda^*$.

3 Update $\mathbf{x}_c$ and test for convergence.

$$\mathbf{x}_c \leftarrow \mathbf{x}_c + \lambda^* \mathbf{h}_c^*.$$

Remark 5.3.1 *Here* $\mathbf{h}_c$ *transforms correctly with* $\mathbf{x}_c$ *as a contravariant vector. This contrasts with the use of the monitors (4.2.8) and (4.2.9) for quasi-likelihood optimization where rescaling of the unknown log likelihood makes no sense and the invariance is to transformations of the parameters to be estimated. It is the form of affine invariant monitor considered here that was shown to have the possibility of cycling in [5]. An improved implementation which at least partially overcomes this problem is given in [14].*

Example 5.3.1 *The problem considered derives from the similarity solutions to the flow between two infinite rotating discs . This problem was considered in [75] and the discussion makes for an interesting comparison. The aim there was to demonstrate the advantages of the multiple shooting method by solving a sequence of problems that had proved difficult by the methods commonly used at the time. The stability advantage was emphasised by the use of the 32 bit single precision computations available on an IBM 360/50.*

The governing fifth order system of differential equations is:

$$\begin{aligned}
\frac{dx_1}{dt} &= -2x_2, \\
\frac{dx_2}{dt} &= x_3, \\
\frac{dx_3}{dt} &= x_1 x_3 + x_2^2 - x_4^2 + k, \\
\frac{dx_4}{dt} &= x_5, \\
\frac{dx_5}{dt} &= 2x_2 x_4 + x_1 x_5;
\end{aligned}$$

and the corresponding boundary conditions are:

$$\begin{aligned}
x_1(0) = x_2(0) = 0,\ x_4(0) = 1, \\
x_1(b) = x_2(b) = 0,\ x_4(b) = s.
\end{aligned}$$

This problem corresponds to a form of nonlinear eigenvalue problem as k *has to be determined so that all six boundary conditions are satisfied. The smoothing approach sets* $k = x_6(t)$ *and adjoins the additional differential equation* $\frac{dx_6(t)}{dt} = 0$. *This reduces the problem to a boundary value problem for a system of* 6 *ordinary differential equations. The parameter* s *in the boundary*

conditions corresponds to the ratio of the speed of rotation of the two discs. Here b is the distance between the two discs and corresponds to the square root of the Reynolds number for the problem. The initial approximation used is $\mathbf{x} = 0$ *in all cases.*

The system of differential equations is turned into a system of nonlinear equations by using the trapezoidal rule integration formula. When combined with the boundary conditions the result is a system of $6n$ *equations in* $6n$ *unknowns which can be solved by Newton's method.*

Results of numerical computations are given in Table 1 below. Three cases are considered corresponding to $s = 0.5,\ 0.0,\ -0.3,\ b = 9$. *Starting values are* $\mathbf{x}_i = 0,\ i = 1, 2, \cdots, n$. *Other settings are* $n = 101$, *iteration tolerance* 10^{-10}, *and line search parameter* $\theta = .25$. *The iteration tolerance is applied to the objective function which is defined as* $\sqrt{\Phi(\lambda)}$ *where* $\Phi(\lambda)$ *is the affine invariant objective in the first case, and the sum of squares of residuals in the second. Note that the objective varies from iteration to iteration in the first case, so that two entries in the table are made for each iteration in the affine case, but maintains the same form for each iteration in the second so that the starting value for each iteration is the terminating value from the previous one.*

In this problem, difficulty typically increases with increasing separation b and decreasing rotation speed ratio s. However, another form of difficulty can occur in the form of multiple solutions (the $s = 0$ *case gets particularly difficult for this reason at intermediate values between* $b = 9$ *and* $b = 18$*) . The feature of the results is the manner in which the sum of squares of residuals outperforms the affine invariant objective. This superiority is maintained in the more difficult problems tried with larger b.*

Example 5.3.2 *A classic example to illustrate nonlinear stability is provided by the Van der Pol equation :*

$$\frac{d^2x}{dt^2} - \lambda\left(1 - x^2\right)\frac{dx}{dt} + x = 0. \tag{5.3.8}$$

This is a "reliably" difficult example with difficulty increasing with λ. *The limit cycle behaviour is illustrated in the scilab plot of initial value trajectories figure 5.1. This shows convergence to the limit cycles for* $\lambda = 1, 10$.

The Van der Pol equation can also be posed as a boundary value problem . The transformation $s = 4t/T$ *maps a* $1/2$ *period onto* $[0, 2]$. *If a new variable*

it	cov			ss	
	λ	objs	objt	λ	obj
s=0.5					
0					3.3541-01
1	.25	2.3096 00	1.7705 00	1	1.6807-01
2	1	2.2310 00	1.4067 00	1	5.9296-03
3	1	1.5757 00	8.7796-01	1	1.3200-03
4	1	1.38771-01	2.0114-02	1	5.1152-06
5	1	2.4072-02	2.2878-04	1	7.0834-11
6	1	2.3238-04	1.4847-08		
7	1	1.4849-08	2.4315-15		
s=0.0					
0					3.999-01
1	.25	1.7769 00	1.4134 00	1	1.2075-01
2	1	2.2370 00	1.5573 00	1	1.8768-02
3	.25	1.5529 00	5.6232-01	1	1.0154-03
4	.25	1.5421 00	1.0481 00	1	1.5290-04
5	1	5.3675-01	7.1059-02	1	1.6576-07
6	1	6.3186-02	3.5046-03	1	3.3204-12
7	1	3.4912-03	4.4195-06		
8	1	4.4287-06	1.6675-11		
s=-0.3					
0					3.1213-01
1	.25	1.6152 00	1.2853 00	1	1.0603-01
2	.25	1.9894 00	1.4501 00	1	1.5674-02
3	1	1.0389 00	8.80146-01	1	5.1339-03
4	1	4.1931-01	4.2213-02	1	5.2968-04
5	1	5.1575-02	8.2002-04	1	2.4096-05
6	1	8.4793-04	3.7020-07	1	2.4622-09
7	1	3.7013-07	1.4444-13	1	1.6204-16

Table 5.4: Rotating disc flow: numerical results for $b = 9$

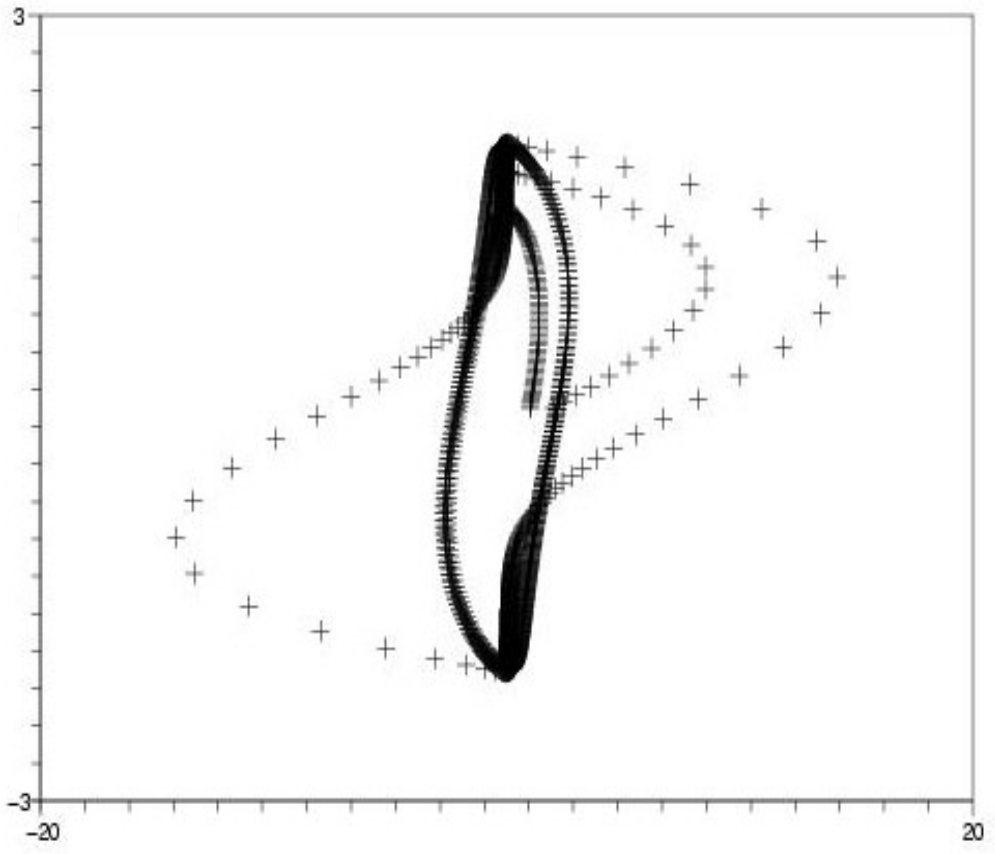

Figure 5.1: Van der Pol equation, initial value trajectories

$x_3 = T/4$ is also introduced then the differential equation becomes

$$\begin{aligned}\frac{dx_1}{ds} &= x_2,\\ \frac{dx_2}{ds} &= \lambda\left(1 - x_1^2\right) x_2 x_3 - x_1 x_3^2,\\ \frac{dx_3}{ds} &= 0.\end{aligned}$$

Appropriate boundary data is

$$B_0 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, B_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \mathbf{b} = 0.$$

The first two conditions specify that the state derivative vanish at successive extrema. The periodicity condition expressed in the third condition states that

the extremal values will alternate in sign. This system has the trivial solution $\mathbf{x} = 0$*, so starting values need to avoid this. Continuation with* $\Delta\lambda = 1$ *has been used in the reported computations starting with the known solution for* $\lambda = 0$*. Meshes* $h = 1/100, 1/1000$ *have been used, as has a mesh based on extrema of Chebyshev polynomials shifted to* $[0, 2]$*. Equispaced meshes such as these can be expected to be inefficient as the need for adaptivity is clear from the scilab figure 5.1. In this case similar accuracy is obtained for the equispaced mesh corresponding to* $h = 1/1000$ *and the Chebyshev grid with* 101 *extrema. The results for* $\lambda = 10$ *are summarised in figure 5.2. The values for the additional cycles are obtained by reflection.*

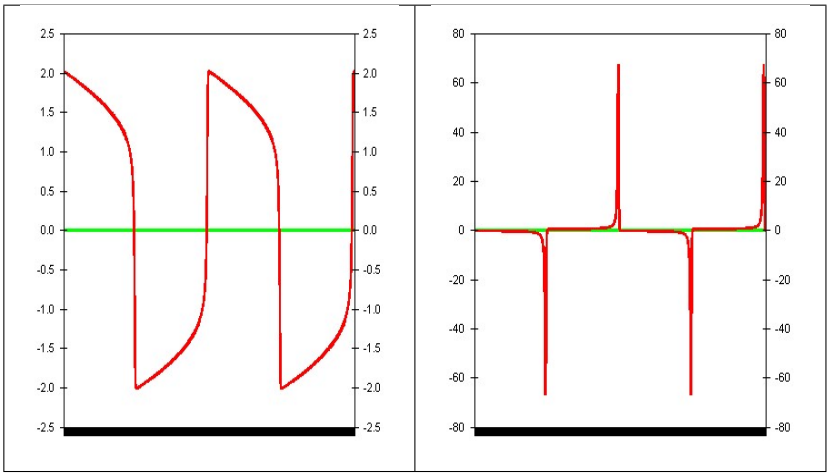

Figure 5.2: Van der Pol equation limit cycle, $\lambda = 10$

Exercise 5.3.1 *Show that* $\mathbf{h_c}$ *is a direction of descent in (5.3.7).*

5.4 The Estimation Problem

5.4.1 Basic questions

It is convenient to use the smoothing formulation of the estimation problem (Remark 5.1.1) in this subsection. Important questions that need to be asked about the two approaches to the estimation problem suggested in the introduction to this chapter are:

1. Are they equivalent? If so:

2. Are they consistent?

Superficially the embedding and simultaneous methods look rather different. This is not misleading. The relatively arbitrary component in the embedding method has been noted, while the simultaneous method has a surprising depth of structure. Perhaps the most obvious feature in common is that they address the same problem! However, progress is possible on the question of equivalence .

Theorem 5.1 *An isolated local minimum of the sums of squares of residuals for either the embedding or simultaneous method is also an isolated local minimum of the sum of squares of residuals of the other. An appropriate choice of boundary matrices B_1, B_2 is assumed.*

Proof. Let $S_S(\mathbf{x})$ be the sum of squares of residuals in the simultaneous method corresponding to feasible $\mathbf{x}$, and let $S_E(\mathbf{x}, \mathbf{b})$ be the sum of squares of residuals in the embedding method corresponding to given boundary vector $\mathbf{b}$. Let $\mathbf{x}_S$ be an isolated local minimum of the simultaneous method in a ball $R(\mathbf{x}_S, \rho)$ of radius ρ for some $\rho > 0$. Then direct substitution gives

$$B_1\mathbf{x}_S(0) + B_2\mathbf{x}_S(1) = \mathbf{b}_S\,.$$

Because $\mathbf{x}_S$ satisfies (5.1.15) the corresponding sum of squares is defined for the embedding method and $S_E(\mathbf{x}_S, \mathbf{b}_S) = S_S(\mathbf{x}_S)$. Assume $\mathbf{x}_S$, $\mathbf{b}_S$ is not a corresponding local minimum of $S_E(\mathbf{x}, \mathbf{b})$. Then there exists $\mathbf{x} = \mathbf{x}_P \in R(\mathbf{x}_S, \rho)$, and $\mathbf{b} = \mathbf{b}_P$ such that

$$S_E(\mathbf{x}_P, \mathbf{b}_P) < S_E(\mathbf{x}_S, \mathbf{b}_S)\ .$$

However, $\mathbf{x}_P$ is feasible for the simultaneous method. Thus

$$S_S(\mathbf{x}_P) = S_E(\mathbf{x}_P, \mathbf{b}_P) < S_S(\mathbf{x}_S)\ .$$

This is a contradiction. It follows that $\mathbf{x}_S$, $\mathbf{b}_S$ provides a local minimum for both methods. The argument can be reversed to show that if $\mathbf{x}_E$, $\mathbf{b}_E$ is a local minimum of the embedding method then it is a local minimum of the simultaneous method also. ■

This is a non–constructive argument. A more interesting result would be one that addressed more of the structure of the methods. Thus it is of interest to show that satisfaction of necessary conditions for either the embedding or simultaneous methods can be deduced from satisfaction of the other.

Theorem 5.2 *Let $\mathbf{x}_S, \boldsymbol{\lambda}_S$ satisfy the necessary conditions on the simultaneous method, then $\mathbf{x}_S, \mathbf{b}_S = B_1\mathbf{x}_S(0) + B_2\mathbf{x}_S(1)$ satisfy the necessary conditions on the embedding method . Let $\mathbf{x}_E, \mathbf{b}_E = B_1\mathbf{x}_E(0) + B_2\mathbf{x}_E(1)$ satisfy the necessary conditions on the embedding method then there exists $\boldsymbol{\lambda}_E$, $\boldsymbol{\lambda}_E^T\nabla_x\mathbf{c} = \nabla_x S_E$ such that $\mathbf{x}_E, \boldsymbol{\lambda}_E$ satisfies the necessary conditions on the simultaneous method .*

Proof. The necessary conditions for a stationary point of the simultaneous method Lagrangian $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = S_S(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{c}(\mathbf{x})$ give

$$\boldsymbol{\lambda}_S^T C_S = -\nabla_x S_S, \quad \mathbf{c}(\mathbf{x}_S) = 0,$$

where $C_S = \nabla_x \mathbf{c}(\mathbf{x}_S)$ with $\mathbf{c}$ given by (5.1.15). This gives

$$\begin{bmatrix} \boldsymbol{\lambda}_S^T & 0 \end{bmatrix} \begin{bmatrix} C_S \\ B_1 \cdots B_2 \end{bmatrix} = -\nabla_x S_S$$

so that

$$\begin{aligned} 0 = \begin{bmatrix} \boldsymbol{\lambda}_S^T & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ I_m \end{bmatrix} &= -\nabla_x S_S \begin{bmatrix} C_S \\ B_1 \cdots B_2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ I_m \end{bmatrix}, \\ &= -\nabla_x S_S \frac{\partial \mathbf{x}}{\partial \mathbf{b}}, \end{aligned}$$

interpreting the right hand side as the solution of the trapezoidal rule discretization of (5.1.13). In the embedding form of the smoothing problem $\mathbf{b}$ provides the adjustable parameters so this shows that the necessary conditions for the embedding method are satisfied.

If the necessary conditions on the embedding method are satisfied then

$$\nabla_x S_E \begin{bmatrix} C_E \\ B_1 \cdots B_2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ I_m \end{bmatrix} = 0.$$

It follows that there is $\mathbf{s} \in R^{(n-1)p}$ such that

$$\nabla_x S_E \begin{bmatrix} C_E \\ B_1 \cdots B_2 \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{s}^T & 0 \end{bmatrix}.$$

Thus

$$\nabla_x S_E = \begin{bmatrix} \mathbf{s}^T & 0 \end{bmatrix} \begin{bmatrix} C_E \\ B_1 \cdots B_2 \end{bmatrix}$$

whence

$$\nabla_x S_E = \mathbf{s}^T C_E$$

showing that the necessary conditions on the simultaneous method are satisfied at $\mathbf{x}_E$ with $\boldsymbol{\lambda}_E = \mathbf{s}$. Note it is implicit in this argument that $\mathbf{c}(\mathbf{x}_E) = 0$ and so is feasible for the simultaneous method. ■

Consistency for the embedding method reduces exactly to the question considered in Theorem 3.4 if an exact (multiple shooting) form of integration is used to solve the embedded boundary value problem because the use of an exact integration scheme is equivalent to using the true model. *It follows that the embedding method is an exact maximum likelihood method for the sum of squares objective corresponding to normally distributed observation errors assumed in this section.* This argument can be extended to an approximate form of integration such as that based on the trapezoidal rule by taking into account truncation error effects . It turns out that the same form of argument applies. The consistency argument for the approximate integration case relies on the estimates $\widehat{\boldsymbol{\beta}}_n, \widehat{\mathbf{b}}_n$ computed using exact integration producing small residuals when used as hypothetical initial guesses for the finite difference scheme estimation for n large. This permits the Kantorovich theorem to be applied to show that the finite difference estimates are increasingly close to the estimates produced by exact integration as $n \to \infty$. This happens because the contribution to the residuals made as a consequence of the truncation error contribution to the approximate integration of the differential equation is relatively small compared to the stochastic errors provided a suitable integration mesh is chosen.

The first requirement in the exact integration case is to ensure that the log likelihood is well defined as a function of the parameters in a suitable neighborhood of $\boldsymbol{\beta}^*, \mathbf{b}^*$.

Assumption 5.3 *There exist positive constants k_1, k_2, k_3 and boundary matrices B_1, B_2 such that for all $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq k_1$, $\|\mathbf{b} - \mathbf{b}^*\| \leq k_2$ the boundary value problem for (5.1.1) has an unique solution $\mathbf{x}(t, \boldsymbol{\beta}, \mathbf{b})$ satisfying $\max_{t_j \in \mathbf{T}_n} \|\mathbf{x}(t_j, \boldsymbol{\beta}, \mathbf{b}) - \mathbf{x}(t_j, \boldsymbol{\beta}^*, \mathbf{b}^*)\| \leq k_3 \left\| \begin{array}{c} \boldsymbol{\beta} - \boldsymbol{\beta}^* \\ \mathbf{b} - \mathbf{b}^* \end{array} \right\|$, where $\mathbf{T}_n$ is the set of data points. This solution can be found by an application of Newton's method; and the conditions for an application of the Kantorovich Theorem (Theorem 3.3) are satisfied.*

Theorem 5.4 *Assume that Assumption 5.3 holds for the embedding method applied to a sequence of regular experiments as $n \to \infty$, and that*

$$\frac{1}{n} \sum_{i=1}^{n} \nabla_{(\boldsymbol{\beta}, \mathbf{b})} \mathbf{x}(t_i, \boldsymbol{\beta}^*, \mathbf{b}^*)^T \mathcal{O}^T \mathcal{O} \nabla_{(\boldsymbol{\beta}, \mathbf{b})} \mathbf{x}(t_i, \boldsymbol{\beta}^*, \mathbf{b}^*) \succ 0 \tag{5.4.1}$$

for all n large enough, then the embedding method is consistent . That is there exists n_0 such that, for $n \geq n_0$,

$$\left[\begin{array}{c} \widehat{\boldsymbol{\beta}}_n \\ \widehat{\mathbf{b}}_n \end{array} \right] \underset{n\to\infty}{\overset{a.s.}{\to}} \left[\begin{array}{c} \boldsymbol{\beta}^* \\ \mathbf{b}^* \end{array} \right], n \to \infty. \tag{5.4.2}$$

Proof. Assumption 5.3 ensures that the signal approximating the data is well defined in the neighbourhood of the true parameter values. Thus it is necessary only to verify that the Hessian matrix of the negative log likelihood is positive definite at the true solution in order to apply the consistency theorem (Theorem 3.4) derived in Chapter 3. Here the negative of the normal likelihood is given, up to an additive constant, by

$$-2\sigma^2 \mathcal{L}\left(\mathbf{y} : \left[\begin{array}{c} \boldsymbol{\beta} \\ \mathbf{b} \end{array} \right], \mathbf{T}_n\right) = \Sigma_{i=1}^n \left\| \mathbf{y}_i - \mathcal{O}\mathbf{x}\left(t_i, \boldsymbol{\beta}, \mathbf{b}\right)\right\|_2^2 .$$

A sufficient condition for positive definiteness of the Hessian is

$$\frac{1}{n}\mathcal{E}^* \left\{ \nabla_{(\boldsymbol{\beta},\mathbf{b})} \mathcal{L}\left(\mathbf{y} : \left[\begin{array}{c} \boldsymbol{\beta}^* \\ \mathbf{b}^* \end{array} \right], \mathbf{T}_n\right)^T \nabla_{(\boldsymbol{\beta},\mathbf{b})} \mathcal{L}\left(\mathbf{y} : \left[\begin{array}{c} \boldsymbol{\beta}^* \\ \mathbf{b}^* \end{array} \right], \mathbf{T}_n\right) \right\} \succ 0$$

as this ensures positive definiteness in an open neighbourhood of $\left[\begin{array}{c} \boldsymbol{\beta}^* \\ \mathbf{b}^* \end{array} \right]$ for all n large enough as the limiting Hessian tends almost surely to its expectation. This condition is just the positive definiteness assumption (5.4.1) in the statement of the Theorem. ■

As in Chapter 3, equation (3.2.10) there is an associated rate equation expressed in terms of the variance . Here this is

$$\mathcal{V}\left\{ \left[\begin{array}{c} \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^* \\ \widehat{\mathbf{b}}_n - \mathbf{b}^* \end{array} \right] \right\} = O\left(\frac{1}{n}\right). \tag{5.4.3}$$

Now it follows from the assumption of continuous dependence (5.3) that a similar rate governs the difference in state variable values.

The above consistency and convergence rate results assuming exact integration can be extended to two important applications of approximate integration :

1. when each differential equation discretization grid $\mathbf{K}_n$ corresponds to the observation grid $\mathbf{T}_n$; and

2. when the discretization is made on a fine enough fixed grid $\{t_j \in \mathbf{K}, j = 1, 2, \cdots, |K|\}$ independent of $\mathbf{T}_n$ as $n \to \infty$.

In the first case the condition for a regular experiment together with the additional requirement that the differential equation be integrated satisfactorily ensures that the maximum mesh spacing $\Delta t \to 0,\ n \to \infty$. In the second case Δt is fixed and finite. This means that truncation error effects persist in the solution of the discretized problem as $n \to \infty$.

The first step is to ensure that both the true and approximate log likelihoods are well defined in a suitable neighbourhood of $\boldsymbol{\beta}^*, \mathbf{b}^*$. This requires a reformulation of Assumption 5.3 to take into account also the need for the solution of discretised system based on the trapezoidal rule to be well determined.

Assumption 5.5 *There exist positive constants $\kappa_1, \kappa_2, \kappa_3, \kappa_4$ and boundary matrices B_1, B_2 such that for all $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq \kappa_1,\ \|\mathbf{b} - \mathbf{b}^*\| \leq \kappa_2$ both the boundary value problem for (5.1.1) has an unique solution $\mathbf{x}(t, \boldsymbol{\beta}, \mathbf{b})$ and there exists Δt_0 such that for all $\Delta t < \Delta t_0$ the discretized system possesses an unique solution $\mathbf{x}_\Delta(t_k, \boldsymbol{\beta}, \mathbf{b})$ defined on the discretization points $t_j \in \mathbf{K}$ such that $\max_k(\|\mathbf{x}(t_k, \boldsymbol{\beta}, \mathbf{b}) - \mathbf{x}_\Delta(t_k, \boldsymbol{\beta}, \mathbf{b})\|) \leq \kappa_4 (\Delta t)^2$. These solutions can be found by an application of Newton's method; and the conditions for an application of the Kantorovich Theorem (Theorem 3.3) are satisfied. As before the boundary value problem solution satisfies the continuous dependence condition:*

$$\max_{t_j \in \mathbf{K}} \|\mathbf{x}(t_k, \boldsymbol{\beta}, \mathbf{b}) - \mathbf{x}(t_k, \boldsymbol{\beta}^*, \mathbf{b}^*)\| \leq \kappa_3 \left\| \begin{array}{c} \boldsymbol{\beta} - \boldsymbol{\beta}^* \\ \mathbf{b} - \mathbf{b}^* \end{array} \right\|.$$

For an analysis of the convergence of the discretised system to the continuous one see [6].

Also needed is a form of the Kantorovich Theorem 3.3 packaged together with equation (3.2.6). This permits the minimum $\widehat{\mathbf{x}}$ of a given function $f(\mathbf{x})$ to be related to the corresponding minimum $\widehat{\mathbf{x}}_\varepsilon$ of the perturbed function $f_\varepsilon(\mathbf{x}) = f(\mathbf{x}) + \varepsilon g(\mathbf{x})$ for small values of ε.

Theorem 5.6 *Assume there exists $\rho > 0$ such that $\widehat{\mathbf{x}}$ is the unique minimum of $f(\mathbf{x})$ in the ball $S(\widehat{\mathbf{x}}, \rho)$ in which the conditions of Theorem 3.3 hold. Further, that for ε small enough, these conditions translate to the application of Newton's method to $f_\varepsilon(\mathbf{x})$ for $\mathbf{x} \in S(\widehat{\mathbf{x}}, \rho)$. In summary:*

(i) $\|\nabla^2 f_\varepsilon(\mathbf{u}) - \nabla^2 f_\varepsilon(\mathbf{v})\| \leq K_1 \|\mathbf{u} - \mathbf{v}\|,\ \mathbf{u}, \mathbf{v} \in S(\widehat{\mathbf{x}}, \rho)$;

(ii) $\left\|\nabla^2 f_\varepsilon(\widehat{\mathbf{x}})^{-1}\right\| \leq K_2$;

(iii) $\left\|\nabla^2 f_\varepsilon(\widehat{\mathbf{x}})^{-1} \nabla f_\varepsilon(\widehat{\mathbf{x}})\right\| \leq K_3,\ \xi = K_1 K_2 K_3 < \frac{1}{2}$;

(iv) $2K_3 < \rho$.

Then $f_\varepsilon(\mathbf{x})$ has a unique minimum $\widehat{\mathbf{x}}_\varepsilon \in S(\widehat{\mathbf{x}}, \rho)$ as a consequence of Theorem 3.3, while equation (3.2.6) gives

$$\|\widehat{\mathbf{x}}_\varepsilon - \widehat{\mathbf{x}}\| < 2K_3.$$

Note that K_3 can be chosen to be proportional to ε as $\nabla f_\varepsilon(\widehat{\mathbf{x}}) = \varepsilon \nabla g_\varepsilon(\widehat{\mathbf{x}})$. This ensures that the conditions of the theorem can be met for ε small enough.

Theorem 5.7 *Assume Assumption 5.5 holds, that $\Delta t(n) = o\left(n^{-1/2}\right)$ as $n \to \infty$, and that n is sufficiently large that*

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathbf{x}_\Delta(t_i, \boldsymbol{\beta}^*, \mathbf{b}^*)^T \mathcal{O}^T \mathcal{O} \nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathbf{x}_\Delta(t_i, \boldsymbol{\beta}^*, \mathbf{b}^*) \succ 0. \tag{5.4.4}$$

Then the embedding method based on the trapezoidal rule discretization is consistent :

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}}_{\Delta t(n)} \\ \widehat{\mathbf{b}}_{\Delta t(n)} \end{bmatrix} \underset{n\to\infty}{\overset{a.s.}{\to}} \begin{bmatrix} \boldsymbol{\beta}^* \\ \mathbf{b}^* \end{bmatrix}.$$

Proof. The result is a consequence of successive applications of Theorem 5.6 to show that $(\widehat{\mathbf{x}}_{\Delta(n)})_c$ approaches $(\widehat{\mathbf{x}}_n)_c$ as $n \to \infty$ where, as before, the subscript c indicates a composite vector. The assumptions ensure that the conditions of this theorem are satisfied provided the choice of K_3 for each n gives $K_3^n \to 0,\ n \to \infty$ with corresponding initial guess $\widehat{\mathbf{x}}_n$. They also ensure that the linearised equations that determine $\nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathbf{x}(t, \boldsymbol{\beta}, \mathbf{b})$ and $\nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathbf{x}_{\Delta(n)}(t, \boldsymbol{\beta}, \mathbf{b})$ have stably computable solutions which differ by $O(\Delta t(n)^2)$. Let $\mathcal{L}_\Delta$ denote the log likelihood evaluated using the solution of the discretised system and $\mathbf{r}_t^\Delta$ the corresponding residuals, n being understood . It is assumed they are evaluated at $\begin{bmatrix} \widehat{\boldsymbol{\beta}}_n \\ \widehat{\mathbf{b}}_n \end{bmatrix}$ for each $\Delta t(n)$. Then

$$\begin{aligned}
\frac{1}{n}\nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathcal{L}_\Delta &= \frac{1}{n}\nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathcal{L}_\Delta - \frac{1}{n}\nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathcal{L}, \\
&= \frac{1}{n\sigma^2}\sum_{t\in T_n} \left\{ \mathbf{r}_t^{\Delta T} \mathcal{O} \nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathbf{x}_\Delta(t) - \mathbf{r}_t^T \mathcal{O} \nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathbf{x}(t) \right\}, \\
&= \frac{1}{n\sigma^2}\sum_{t\in T_n} \Big\{ \left(\mathbf{r}_t^\Delta - \mathbf{r}_t\right)^T \mathcal{O} \nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathbf{x}_\Delta(t) \\
&\quad + \mathbf{r}_t^T \mathcal{O} \left(\nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathbf{x}_\Delta(t) - \nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathbf{x}(t) \right) \Big\}.
\end{aligned} \tag{5.4.5}$$

The deterministic contribution to $\frac{1}{n}\nabla_{(\boldsymbol{\beta},\mathbf{b})}\mathcal{L}_\Delta$ comes from the terms

1. $\left(\mathbf{r}_t^{\Delta} - \mathbf{r}_t\right)^T \mathcal{O} \nabla_{(\boldsymbol{\beta},\mathbf{b})} \mathbf{x}_{\Delta}(t)$, and

2. $\left(\mathbf{x}^*(t) - \mathbf{x}(t)\right)^T \mathcal{O}^T \mathcal{O} \left(\nabla_{(\boldsymbol{\beta},\mathbf{b})} \mathbf{x}_{\Delta}(t) - \nabla_{(\boldsymbol{\beta},\mathbf{b})} \mathbf{x}(t)\right)$.

Here use has been made of $\mathbf{y}_t = \boldsymbol{\varepsilon}_t + \mathcal{O}\mathbf{x}_t^*$. Summation of the first set contributes a term which is of the order of the truncation error $o(n^{-1})$ when the regular experiment condition is applied, while summation of the the second depends not only on the truncation error estimate but also on the convergence of $(\widehat{\mathbf{x}}_n)_c$ to its limiting value and so is almost surely smaller. The stochastic contribution comes from the terms $\boldsymbol{\varepsilon}_t^T \mathcal{O} \left(\nabla_{(\boldsymbol{\beta},\mathbf{b})} \mathbf{x}_{\Delta}(t) - \nabla_{(\boldsymbol{\beta},\mathbf{b})} \mathbf{x}(t)\right)$. The law of large numbers shows that these terms make a small contribution almost surely even before the truncation error component is taken into account. It follows that $\frac{1}{n} \nabla_{(\boldsymbol{\beta},\mathbf{b})} \mathcal{L}_{\Delta} \overset{a.s.}{\underset{n\to\infty}{\to}} 0$, $n \to \infty \Rightarrow K_3^n \overset{a.s.}{\underset{n\to\infty}{\to}} 0$. Thus

$$\left\| \begin{bmatrix} \widehat{\boldsymbol{\beta}}_n - \widehat{\boldsymbol{\beta}}_{\Delta} \\ \widehat{\mathbf{b}}_n - \widehat{\mathbf{b}}_{\Delta} \end{bmatrix} \right\| \leq 2K_3^n \overset{a.s.}{\underset{n\to\infty}{\to}} 0.$$

The consistency result now follows from (5.4.2). ■

Corollary 5.1 *If the value of Δt is fixed small enough instead of proceeding to the limit as $\Delta t \to 0$ then there is a ball centred on $\begin{bmatrix} \boldsymbol{\beta}^* \\ \mathbf{b}^* \end{bmatrix}$ such that*

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}}_{\Delta} \\ \widehat{\mathbf{b}}_{\Delta} \end{bmatrix} \subset S\left(\begin{bmatrix} \boldsymbol{\beta}^* \\ \mathbf{b}^* \end{bmatrix}, O\left(\Delta t^2\right) \right), \; n \to \infty.$$

This follows because the truncation error $O(\Delta t^2)$ is a factor in all the terms in (5.4.5). Thus $K_3^n = O(\Delta t^2)$ for all n large enough.

Remark 5.4.1 *These results have consequences also for the performance of the Gauss-Newton iteration. When $\Delta t \to 0$ the asymptotic convergence rate is essentially the same as that of the exact procedure and approaches second order. When Δt is fixed the rate contains a truncation error term of $O(\Delta t^2)$ and so asymptotes to a fast, first order method.*

Exercise 5.4.1 *Complete the proof of Theorem 5.1. That is given a local minimum of $S_E(\mathbf{x}, \mathbf{b})$ show that it provides a local minimum of $S_S(\mathbf{x})$.*

5.4.2 Embedding method details

The embedding method assumes that boundary conditions exist such that the linearization of the differential system about the true solution of the estimation problem has a well determined solution. A sufficient condition for this is that the linearized system of differential equations possesses a non-trivial dichotomy . However, the Van der Pol example shows that nonlinear problems can possess a well determined nonlinear stability behaviour in circumstances in which a fixed partion into increasing and decreasing solutions of the linearised problem is not available . Di-stability may not provide a panacea here. This means general solution methods should incorporate such facilities as adaptive meshing to ensure reasonable integration accuracy and a continuation facility to provide good enough starting values for the iterative solver [6]. The embedding method has the advantage of being sufficiently modular that well performed boundary value solvers can be readily incorporated into the estimation procedure. The first step in implementing the embedding algorithm is to determine suitable boundary conditions as a key part of developing the boundary value framework . Two examples taken from the honours thesis [17] illustrate problems caused by inadequate adjoined conditions. In both examples initial conditions are imposed on the system of differential equations. Potential problems are then illustrated using plots of the sum of squares of residuals response surface as problem parameters are varied about imposed true values. The basic idea for the first plot is taken from the paper [93]. The differential equation system considered is the Fitzhugh-Nagumo equation

$$\frac{dx_1}{dt} = 3\left(x_1 - \frac{x_1^3}{3} + x_2\right), \tag{5.4.6}$$

$$\frac{dx_2}{dt} = -\frac{1}{3}\left(x_1 - a + bx_2\right), \tag{5.4.7}$$

with true parameter values $a = b = .2$, and subject to initial conditions $\mathbf{x}(0)^T = [-1, 1]$. The response surface is plotted for parameter values in the range $-1 \leq a, b \leq 1$ for terminal integration values $t_n = 20, 40, 70, 100$. The plot given in [93] corresponds to the case $t_n = 100$. This system is often used as a nonlinear example displaying periodic behaviour. There is evidence of this behaviour here, but the plots also show regions of solution roughness which increase with t_n in a manner suggesting chaotic behaviour, and there is an interesting change in the character of the response surface for larger values of b for each of the t_n values which could suggest bifurcation. There appears to be a reasonable domain of attraction for initial parameter values selected in the immediate neighbourhood of $(.2, .2)$, but initial choices

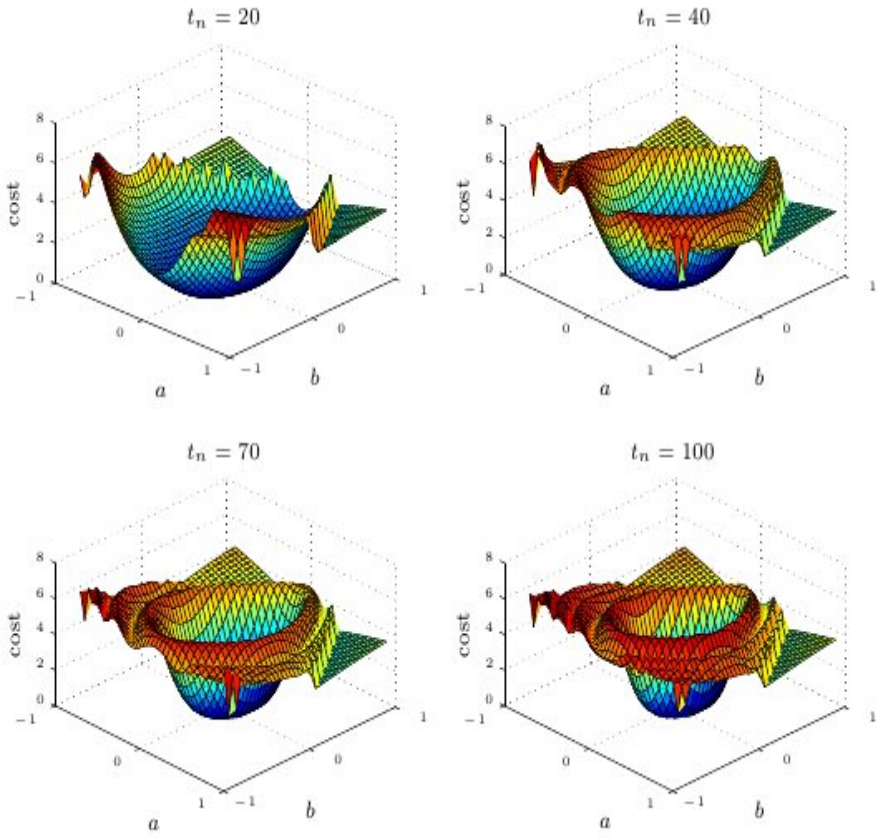

Figure 5.3: Estimation response surface plots for Fitzhugh-Nagumo equations

outside this region are unlikely to provide successful starting values for the estimation problem.

The second example provides dramatic evidence that an initial value formulation of the embedding method is fraught with difficulty when the differential system is chaotic. We consider the Lorenz equations

$$\frac{dx_1}{dt} = 10\left(x_2 - x_1\right), \tag{5.4.8}$$

$$\frac{dx_2}{dt} = x_1\left(28 - x_3\right) - x_2, \tag{5.4.9}$$

$$\frac{dx_3}{dt} = x_1 x_2 - \frac{8}{3}x_3. \tag{5.4.10}$$

This system is chaotic with Lyapunov exponents $\lambda_1 = .905, \lambda_2 = 0., \lambda_3 = -14.57$. The instability of the system is illustrated in Figure 5.4 where plots of trajectories for two sets of initial values $\mathbf{x}^*\left(0\right)^T = [1, 1, 30]$, and $\widehat{\mathbf{x}}\left(0\right)^T = [-0.1, 2, 31]$ are displayed. Reasonably close initially, the trajectories begin to diverge seriously about $t = 1$, the divergence being more in phase rather than amplitude. The response surface plots are very revealing. Here $\mathbf{x}^*\left(0\right)$ is taken as the true vector of unknowns corresponding to the boundary value parameters in the embedding method. The response surfaces are plotted as functions of $x_1(0), x_2(0)$ and correspond to terminal integration values $t_n = 1, 3, 5, 10$. The plots correspond to the choice $n = 1000$. The instability of the response surfaces with respect to the initial value parameters as t_n increases is clearly evident . This conclusion is valid despite the instability of the forward integration. The system actually possesses a backward stability property [20] which means that the visual impression is correct even if the numerical detail cannot be accurate. The crucial point is that the initial value formulation of the embedding method in the case of chaotic system dynamics is a recipe for serious disappointment. We have seen already that inappropriate choice of conditions in the case of strong dichotomy leads to serious problems. It follows that the adjoining of auxiliary conditions in the embedding method must be done appropriately and the a priori choice of initial values, if made at all, should be done with caution .

The recommended strategy is to choose the adjoined conditions equal to the (natural) conditions determined in Remark 5.2.1 when these are computed for the linearised system corresponding to $\mathbf{x}$ variation evaluated at the initial solution estimate $\mathbf{x}_c^0, \boldsymbol{\beta}^0$. These have the great advantage that they can be set automatically. Typically we set $\rho = 1$ unless there is evidence of

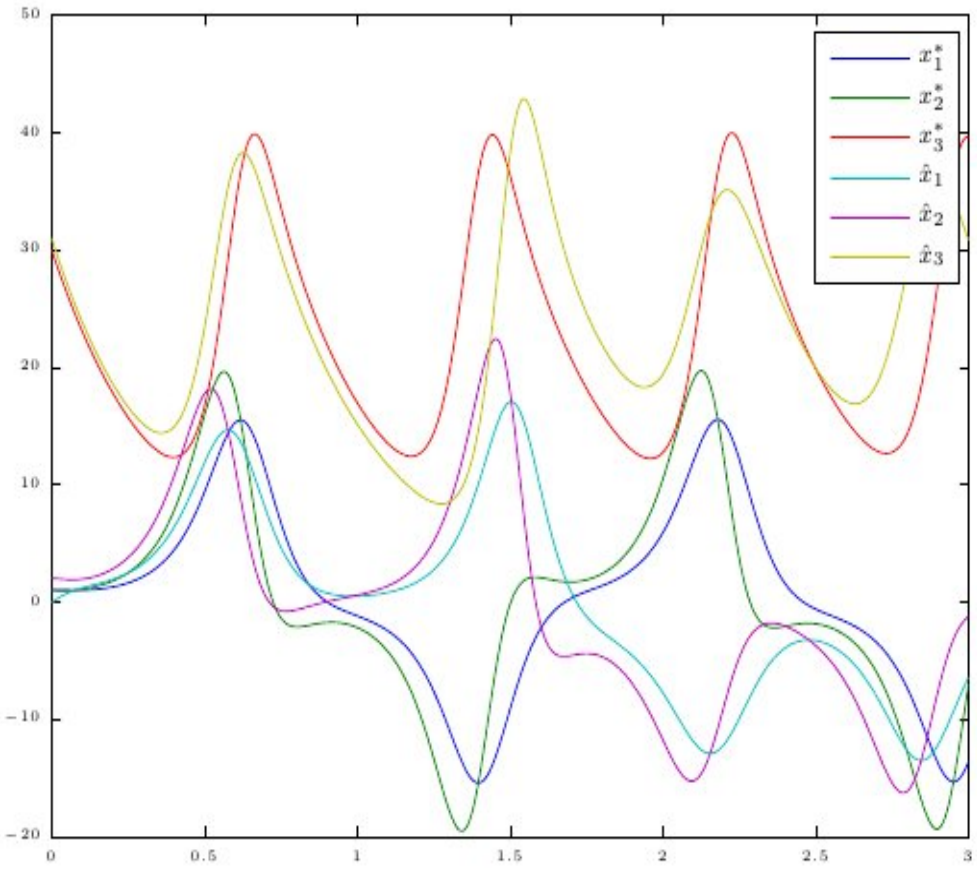

Figure 5.4: Diverging trajectory plots for Lorenz system

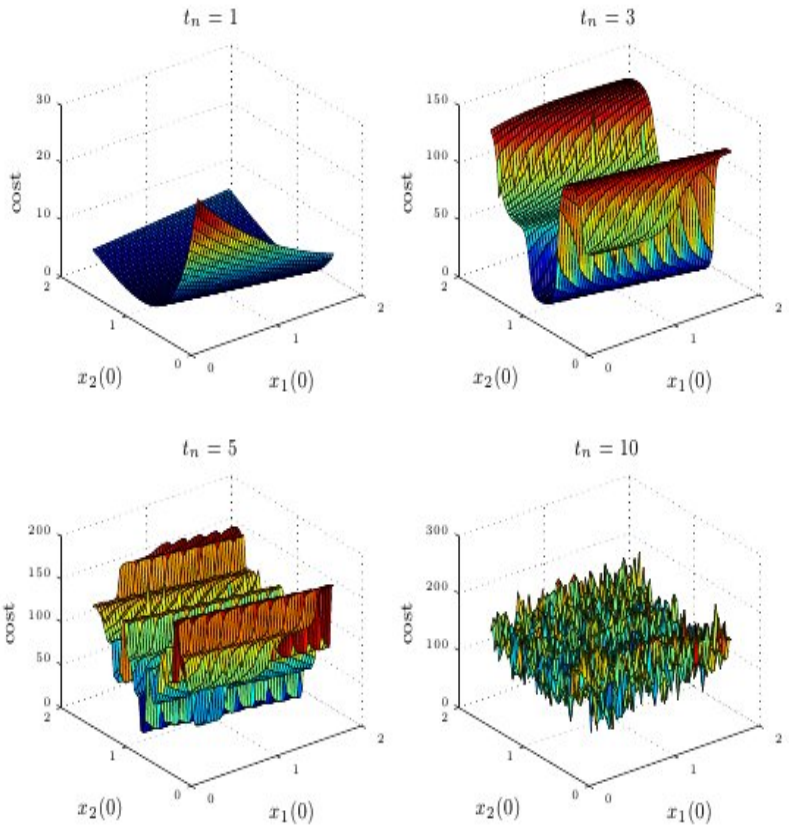

Figure 5.5: Estimation response surface plots for Lorenz equations

scaling problems. Let the matrix of the linearised system be written

$$\mathbf{F}'\left(\mathbf{x}_c^0, \boldsymbol{\beta}^0\right) = \begin{bmatrix} X_{11} & X_{12} & & & & \\ & X_{22} & X_{23} & & & \\ & & & \cdots & & \\ & & & & X_{(n-1)(n-1)} & X_{(n-1)n} \end{bmatrix}.$$

The procedure used here to determine the natural boundary conditions starts by permuting the first block column of F' to the last column position by postmultiplying by a permutation matrix P. This gives

$$\mathbf{F}'\left(\mathbf{x}_c^0, \boldsymbol{\beta}^0\right) P = \begin{bmatrix} X_{12} & & & & X_{11} \\ X_{22} & X_{23} & & & 0 \\ & & \cdots & & \vdots \\ & & X_{(n-1)(n-1)} & X_{(n-1)n} & 0 \end{bmatrix}.$$

An orthogonal factorization of the permuted matrix to upper triangular form yields the desired form:

$$Q^T\mathbf{F}'\left(\mathbf{x}_c^0, \boldsymbol{\beta}^0\right) P = \begin{bmatrix} R_{11} & R_{12} & 0 & \cdots & 0 & R_{1n} \\ & R_{22} & R_{23} & \cdots & 0 & R_{2n} \\ & & & \cdots & \vdots & \vdots \\ & & & & R_{(n-1)(n-1)} & R_{(n-1)n} \end{bmatrix}. \quad (5.4.11)$$

To connect the result of this factorization with the cyclic reduction results note that in equations (5.2.17) and (5.2.18) it is shown that the cyclic reduction factorization can be related to an orthogonal times upper triangular factorization of a permuted matrix

$$\mathbf{F}' = \overline{QR}P_R,$$

while the above development gives

$$\mathbf{F}' = QRP^T.$$

Thus

$$\mathbf{F}'^T\mathbf{F}' = P_R^T\overline{R}^T\overline{R}P_R = P^TR^TRP.$$

It follows that for each i, j there corresponds s, t such that

$$\left(\overline{R}^T\overline{R}\right)_{ij} = \left(R^TR\right)_{st}$$

Thus there is a close relation which identifies corresponding elements produced by the two transformations.

Constraint matrices

$$H\left(\mathbf{x}_c^0, \boldsymbol{\beta}^0\right) = R_{(n-1)n}, \quad G\left(\mathbf{x}_c^0, \boldsymbol{\beta}^0\right) = R_{(n-1)(n-1)}, \tag{5.4.12}$$

in the sense of the cyclic reduction calculations can be identified by analogy with (5.2.12), and this identification permits the required boundary matrices B_1, B_2 to be computed by the orthogonal factorization (5.2.16). For this choice of B_1 and B_2 to be satisfactory initially requires the initial estimates of the additional parameters $\mathbf{b}$ to be adequate in the sense of leading to sensible corrections in the initial Newton step. This is something like well posed assumption for the estimation problem. However, it would be necessary to recompute B_1, B_2 and re-embed if there is evidence of serious deterioration in the conditioning of the calculation of $\mathbf{x}_c$. One requirement is that $\|Z\|$ evaluated at the current iterate be small where

$$Z = \left[\begin{array}{cc} H\left(\mathbf{x}_c, \boldsymbol{\beta}\right) & G\left(\mathbf{x}_c, \boldsymbol{\beta}\right) \end{array} \right] \left[\begin{array}{c} B_1^T \\ B_2^T \end{array} \right].$$

If Z is small then there has been little loss of the initial orthogonality at the current stage of the computation. If this is the case then computational problems possibly suggest inherent problems with the differential system. If Z is not small, and resetting the boundary conditions does not improve matters, then the most likely causes are either the selection of poor initial estimates, suggesting the use of a continuation strategy , or a differential system that is not stably posed as a two-point boundary value problem.

The steps involved in the embedding form of the estimation algorithm are as follows:

1. Provide starting values for $\boldsymbol{\beta}$ and $\mathbf{b}$. The provision of a suitable initial estimate for $\boldsymbol{\beta}$ is an expected requirement and could follow from external information. However, the initial choice of $\mathbf{b}$ is linked to the choice of embedding and satisfactory values could be harder to find.

   The choice $\mathbf{b} = 0$ is made in the following example.

2. Solve the embedded nonlinear system to provide the comparison values $\mathbf{x}_c$ in order to compute the residual vector at the current point $(\boldsymbol{\beta}, \mathbf{b})$. It is assumed here that the problem log likelihood is a sum of squares of residuals corresponding to a normal distribution of data errors. Calculation of $\mathbf{x}_c$ involves solving the nonlinear differential system so it involves an inner iteration which terminates with $\mathbf{F}'\left(\mathbf{x}_c, \boldsymbol{\beta}\right)$ available as a byproduct of the Newton iteration.

3. Use the scoring method to generate corrections to the current parameter vector $\boldsymbol{\beta}$ and boundary vector $\mathbf{b}$. This requires integrating the variational equations (compare (5.1.10), (5.1.11), (5.1.12), (5.1.13)) which are linear but have matrix arguments:

$$\mathbf{F}'(\mathbf{x}_c, \boldsymbol{\beta}) \frac{\partial \mathbf{x}_c}{\partial \boldsymbol{\beta}} = \nabla_\beta \mathbf{F}(\mathbf{x}_c, \boldsymbol{\beta}), \qquad B\left(\frac{\partial \mathbf{x}_c}{\partial \boldsymbol{\beta}}\right) = 0, \tag{5.4.13}$$

$$\mathbf{F}'(\mathbf{x}_c, \boldsymbol{\beta}) \frac{\partial \mathbf{x}_c}{\partial \mathbf{b}} = 0, \qquad B\left(\frac{\partial \mathbf{x}_c}{\partial \mathbf{b}}\right) = I. \tag{5.4.14}$$

4. Perform a linesearch using the likelihood as monitor. Note that this requires values of $\mathbf{x}_c$ in the search direction and these require the integration of the nonlinear system. This puts an emphasis on being able to accept the initial (unit) step in the linesearch parameter. The current parameter vector $\boldsymbol{\beta}$ and boundary vector $\mathbf{b}$ are then updated, and the convergence test is applied.

Example 5.4.1 *Again the Mattheij example : Consider the modification of the Mattheij problem (5.2.47), (5.2.48) with parameters* $\beta_1^* = \gamma$, *and* $\beta_2^* = 2$. *This system possesses the solution* $\mathbf{x}(t, \boldsymbol{\beta}^*) = e^t \mathbf{e}$ *for arbitrary* γ. *The modified system is:*

$$M(t) = \begin{bmatrix} 1 - \beta_1 \cos \beta_2 t & 0 & 1 + \beta_1 \sin \beta_2 t \\ 0 & \beta_1 & 0 \\ -1 + \beta_1 \sin \beta_2 t & 0 & 1 + \beta_1 \cos \beta_2 t \end{bmatrix},$$

$$\mathbf{f}(t) = \begin{bmatrix} e^t \left(-1 + \gamma \left(\cos 2t - \sin 2t\right)\right) \\ -(\gamma - 1) e^t \\ e^t \left(1 - \gamma \left(\cos 2t + \sin 2t\right)\right) \end{bmatrix}.$$

In the numerical experiments reported in Table 5.5 the natural boundary conditions with $\rho = 1$ *have been set at the first iteration . The aim is to recover estimates of* $\boldsymbol{\beta}^*, \mathbf{b}^*$ *from simulated data* $e^{t_i} O\mathbf{e} + \boldsymbol{\varepsilon}_i$, $\boldsymbol{\varepsilon}_i \sim N(0, .01I)$ *using the Gauss-Newton algorithm. The computation is stopped when* $\nabla F \mathbf{h} < 10^{-8}$.

Here the angle between the initially selected boundary conditions and those that would be available at subsequent iterations remains small. We have

$$\| \begin{bmatrix} B_1 & B_2 \end{bmatrix}_1 \begin{bmatrix} B_1 & B_2 \end{bmatrix}_k^T - I \|_F < 10^{-3},$$

$k > 1$, *where* k *is the iteration number, and the norm is the Frobenius norm.*

The second example compares the performance of initial value and boundary value methods on the Lorenz equations (5.4.8), (5.4.9), (5.4.10). The "true" solution data is obtained by solving the initial value problem with $\mathbf{x}^*(0)^T = [1, 1, 30]$.

$O = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}$	$O = \begin{bmatrix} .5 & 0 & .5 \\ 0 & 1 & 0 \end{bmatrix}$
$n = 51,\ \gamma = 10,\ \sigma = .1$	$n = 51,\ \gamma = 10,\ \sigma = .1$
14 iterations	5 iterations
$n = 51,\ \gamma = 20,\ \sigma = .1$	$n = 51,\ \gamma = 20,\ \sigma = .1$
11 iterations	9 iterations
$n = 251,\ \gamma = 10,\ \sigma = .1$	$n = 251,\ \gamma = 10,\ \sigma = .1$
9 iterations	4 iterations
$n = 251,\ \gamma = 20,\ \sigma = .1$	$n = 251,\ \gamma = 20,\ \sigma = .1$
8 iterations	5 iterations

Table 5.5: Summary of Gauss-Newton results for the modified Mattheij problem

Example 5.4.2 *It proved necessary to adjust the form of the estimation data on the initial value form of this problem to obtain any results at all. The final form used was*

$$\mathbf{y}_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{x}^*(t_i) + \mathbf{e}_i, i = 1, 2, \cdots, n,$$

where $\mathbf{e}_i \sim N(\mathbf{0}, I)$*, and* $n = 200$*. Starting estimates for* $\mathbf{b}$ *and* $\boldsymbol{\beta}$ *are generated by adding random noise to the exact values*

$$\boldsymbol{\beta} = \boldsymbol{\beta}^* + \delta\boldsymbol{\beta},$$
$$\mathbf{b} = \mathbf{b}^* + \delta\mathbf{b},$$

with $\delta\mathbf{b} \sim N(0, 3I)$*,* $\delta\boldsymbol{\beta} \sim N(0, 0.15I)$*. Although a convergent iteration was obtained,the result did not correspond to the expected solution. Results are summarised in Figure 5.6: The natural boundary matrices corresponding to the true solution* $\mathbf{x}^*(t)$ *are*

$$B_1 = \begin{bmatrix} -0.0155 & 0.0084 & -0.2942 \\ 0.0483 & 0.0061 & 0.8043 \\ -0.9790 & 0.1958 & 0.0574 \end{bmatrix},$$

and

$$B_2 = \begin{bmatrix} -0.4504 & -0.7696 & 0.3434 \\ 0 & -0.4694 & -0.3611 \\ 0 & 0 & 0 \end{bmatrix},$$

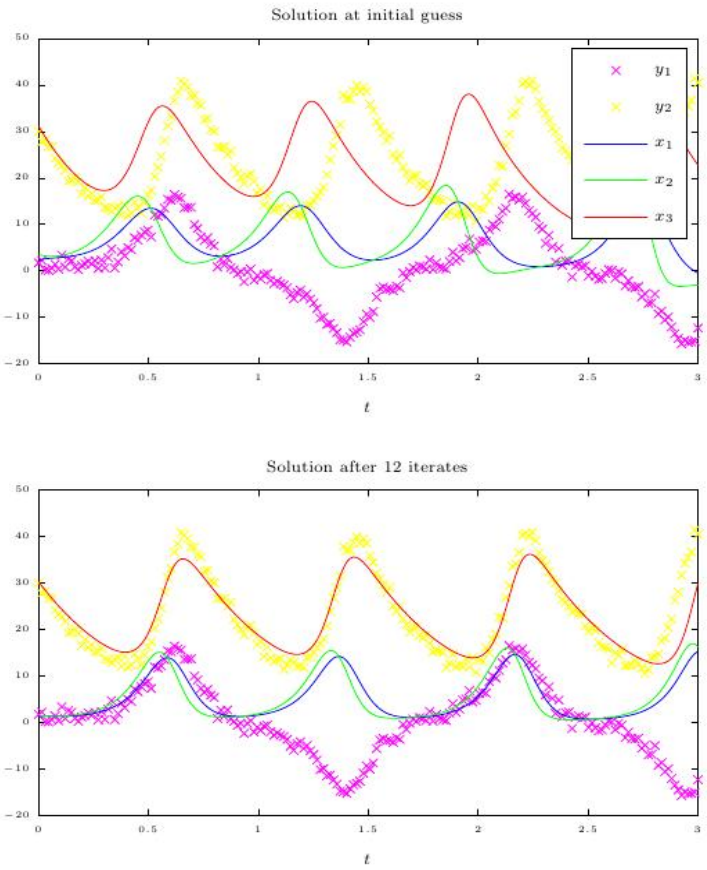

Figure 5.6: Initial and converged solutions for the initial value formulation

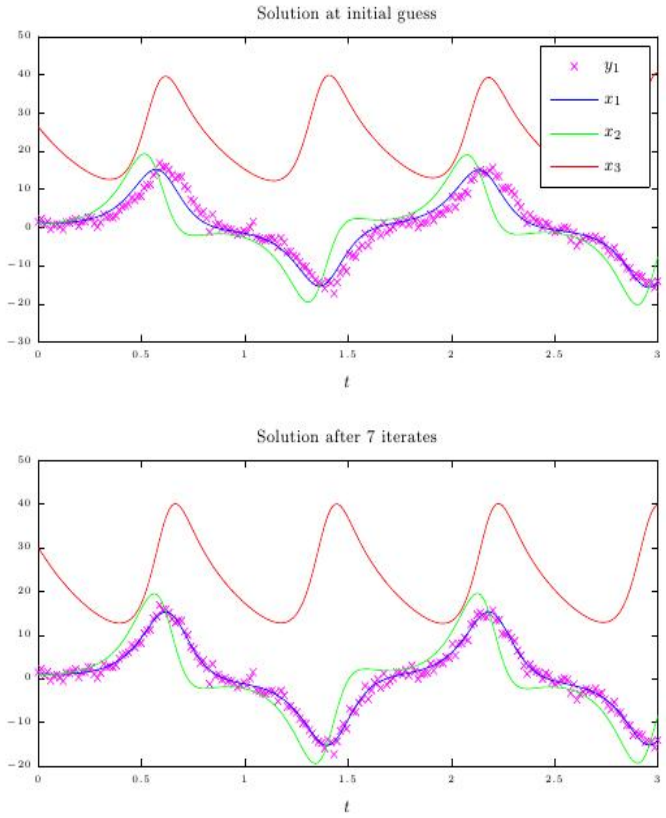

Figure 5.7: Initial and converged solutions for the boundary value formulation

and otherwise similar initialisation is used. These boundary matrices are used here to define the particular form of the embedding method. The data used corresponded to the choice $y(t_i)_1 = x^*(t_i)_1 + \varepsilon_i, \varepsilon_i \sim N(0,1), i = 1, 2, \cdots, 200$ *corresponding to a failed case for the IVP. This time a satisfactory computation is achieved. This choice corresponds to the variable corresponding to the drive equation. Results are summarised in Figure 5.7:*

An example of a system with two positive Lyapunov exponents is given

by the Lorenz (1996) equations

$$\frac{dx_1}{dt} = x_5 (x_2 - x_3) - x_1 + f, \tag{5.4.15}$$
$$\frac{dx_2}{dt} = x_1 (x_3 - x_4) - x_2 + f, \tag{5.4.16}$$
$$\frac{dx_3}{dt} = x_2 (x_4 - x_5) - x_3 + f, \tag{5.4.17}$$
$$\frac{dx_4}{dt} = x_3 (x_5 - x_1) - x_4 + f, \tag{5.4.18}$$
$$\frac{dx_5}{dt} = x_4 (x_1 - x_2) - x_5 + f, \tag{5.4.19}$$

with f=8.17. The estimation problem based on this system is discussed in some detail in [1] who note that the presence of two positive Lyapunov exponents requires a modification of their basic synchronized initial value approach which involves penalising both the first and third differential equations ,(5.4.15) and (5.4.17), and a consequent need to collect values of the corresponding solution values as data items at each observation point . This constraint on the provision of estimation problem data does not apply to the boundary value embedding approach. This point is illustrated here using a single solution value in the data sequence in the embedding algorithm.

Example 5.4.3 *This example illustrates the robustness of the boundary value embedding method in this more complicated situation. The data sequence is chosen as*

$$y_i = x^* (t_i)_5 + \varepsilon_i, i = 1, 2, \cdots, 200,$$

where $\varepsilon_i \sim N(0, \sigma^2)$, $\sigma = 0.3$, *and the* $*$ *is used to denote exact quantities. The computed boundary matrices are*

$$B_1 = \begin{bmatrix} -0.2466 & 0.3415 & -0.0282 & 0.4924 & 0.6841 \\ -0.1790 & -0.2863 & -0.1201 & 0.3459 & -0.1249 \\ 0.3798 & -0.3457 & -0.6443 & 0.2565 & 0.2039 \\ -0.7452 & 0.0797 & -0.5354 & -0.1267 & -0.2342 \\ -0.2778 & -0.5959 & 0.1805 & -0.4544 & 0.5535 \end{bmatrix},$$

and

$$B_2 = \begin{bmatrix} -0.2674 & -0.0961 & 0.1433 & -0.0095 & 0.0998 \\ 0 & 0.4998 & 0.2637 & 0.6186 & -0.1854 \\ 0 & 0 & -0.4222 & -0.1365 & -0.1302 \\ 0 & 0 & 0 & -0.2775 & -0.0612 \\ 0 & 0 & 0 & 0 & -0.1493 \end{bmatrix}.$$

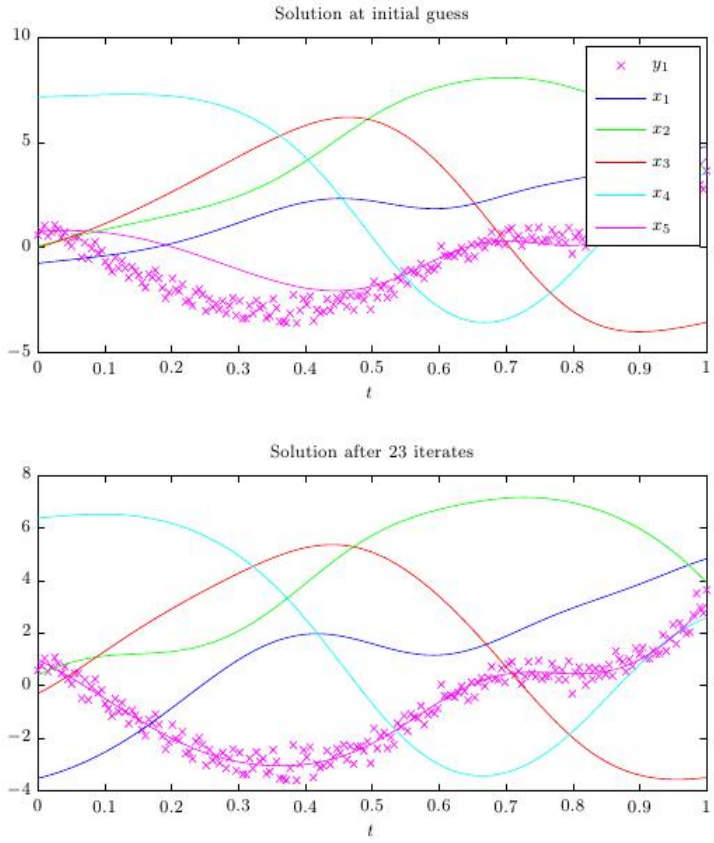

Figure 5.8: Initial and converged solutions for the Lorenz 1996 model

starting estimates for $\mathbf{b}$ *and* $\boldsymbol{\beta}$ *are generated by adding random noise to the exact values*

$$\boldsymbol{\beta} = \boldsymbol{\beta}^* + \delta\boldsymbol{\beta},$$
$$\mathbf{b} = \mathbf{b}^* + \delta\mathbf{b},$$

with $\delta\mathbf{b} \sim N(0,1), \delta\boldsymbol{\beta} \sim N(0, 0.15)$. *The results for the Gauss-Newton algorithm are displayed in figure 5.8. They show that rate of convergence measured by number of iterations in this application, while still reasonably satisfactory, proved distinctly slower than in the previous example with little evidence of the asymptotic second order convergence expected for large* n. *This may be a hint that a different choice of observed data could be more satisfactory. The choice made in [1] is a consequence of the choice of the first and*

third equations for penalisation in order to synchronise with the corresponding data items. The data choice made here is $y_1(t_j) = x_5(t_j, \boldsymbol{\beta}^*), j = 1, 2, \cdots, n$ *and has no corresponding structural justification.*

Exercise 5.4.2 *Consider the multiple shooting method (exact integration). Show that*

$$R_{(n-1)(n-1)}X(1,0) + R_{(n-1)n} = 0, \tag{5.4.20}$$

where R *is the matrix corresponding to (5.4.11) and* $X(t,0)$ *is the fundamental matrix satisfying* $X(0,0) = I$*. Hence show that in this case the natural boundary condition matrices are given by*

$$\begin{bmatrix} B_1 & B_2 \end{bmatrix} = \rho Q_1^T, \tag{5.4.21}$$

where Q_1 *is defined by the orthogonal factorization*

$$\begin{bmatrix} I \\ X(1,0)^T \end{bmatrix} = Q_1 U,$$

and ρ *is the scale factor in (5.2.16).*

Why is this not a practical method for evaluating these conditions in general?

5.4.3 The simultaneous method

The name implies that estimates of the solution vector $\mathbf{x}$ and parameter vector $\boldsymbol{\beta}$ are refined simultaneously [105]. This is in contrast to the embedding method with its nontrivial inner iteration which requires the solution of the variational boundary value problems (5.1.10), (5.1.11), (5.1.12), (5.1.13) needed for the current Gauss-Newton iteration. The constrained optimization setting of the simultaneous method is more complicated as Lagrange multiplier estimates must be computed, but this approach has the advantage that no form of boundary information is required. It is convenient here to use the smoothing form of the problem in which the parameters are treated as additional state variables (Remark 5.1.1), and the augmented differential equation written

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}\left(t, \mathbf{x}\right). \tag{5.4.22}$$

The observation equations have the form

$$\mathbf{y}_i = O\mathbf{x}^*\left(t_i\right) + \boldsymbol{\varepsilon}_i, \quad i = 1, 2, \cdots, n, \tag{5.4.23}$$

and $\mathbf{x}_c$, the blocked vector with sub-block components $\mathbf{x}(t_i) = \mathbf{x}_i, i = 1, 2, \cdots, n$, is chosen to minimize $\frac{1}{2n}\sum_{i=1}^{n} \|\mathbf{r}_i(\mathbf{x}_i)\|^2$ where

$$\mathbf{r}_i = \mathbf{y}_i - O\mathbf{x}_i.$$

The feature of the simultaneous method is that the objective is minimized subject to equality constraints obtained by discretizing the differential equation. Using the trapezoidal rule assuming an equispaced solution grid with spacing Δt which corresponds also to the points at which the observations $\mathbf{y}_i$ are made is the simplest of the available possibilities. Here the constraints have the form

$$\mathbf{c}_i(\mathbf{x}_i, \mathbf{x}_{i+1}) = 0, \quad i = 1, 2, \cdots, n-1.$$

where

$$\mathbf{c}_i = \mathbf{x}_{i+1} - \mathbf{x}_i - \frac{\Delta t}{2}\left\{\mathbf{f}(t_i, \mathbf{x}_i) + \mathbf{f}(t_{i+1}, \mathbf{x}_{i+1})\right\}, \tag{5.4.24}$$

$$= \mathbf{c}_{ii}(\mathbf{x}_i) + \mathbf{c}_{i(i+1)}(\mathbf{x}_{i+1}), \quad i = 1, 2, \cdots, n-1. \tag{5.4.25}$$

Let

$$\Phi_n(\mathbf{x}_c) = \frac{1}{2n}\sum_{i=1}^{n} \|\mathbf{r}_i(\mathbf{x}_i)\|_2^2.$$

Associated with the simultaneous method is the Lagrangian

$$\begin{aligned}\mathcal{L}_n(\mathbf{x}_c, \boldsymbol{\lambda}_c) &= \Phi_n(\mathbf{x}_c) + \sum_{i=1}^{n-1} \boldsymbol{\lambda}_i^T \mathbf{c}_i, \\ &= \Phi_n(\mathbf{x}_c) + \boldsymbol{\lambda}_1^T \mathbf{c}_{11}(\mathbf{x}_1) + \sum_{i=2}^{n-1}\left(\boldsymbol{\lambda}_{i-1}^T \mathbf{c}_{(i-1)i}(\mathbf{x}_i) + \boldsymbol{\lambda}_i^T \mathbf{c}_{ii}(\mathbf{x}_i)\right) \\ &+ \boldsymbol{\lambda}_{n-1}^T \mathbf{c}_{(n-1)n}(\mathbf{x}_n). \end{aligned} \tag{5.4.26}$$

Note that, as a consequence of the structure of the trapezoidal rule (5.4.25), the Lagrangian is separable in the sense that it can be represented as a sum of terms each of which depends only on the individual sub-block components $\mathbf{x}_i$ of the state vector $\mathbf{x}_c$. The necessary conditions for a solution of the constrained problem are

$$\nabla_{\mathbf{x}_i}\mathcal{L}_n = 0, \ i = 1, 2, \cdots, n, \quad \mathbf{c}(\mathbf{x}_c) = 0. \tag{5.4.27}$$

Here the gradient of the Lagrangian gives the equations

$$-\frac{1}{n}\mathbf{r}_1^T O + \boldsymbol{\lambda}_1^T \nabla_{\mathbf{x}_1}\mathbf{c}_{11} = 0, \tag{5.4.28}$$

$$-\frac{1}{n}\mathbf{r}_i^T O + \boldsymbol{\lambda}_{i-1}^T \nabla_{\mathbf{x}_i}\mathbf{c}_{(i-1)i} + \boldsymbol{\lambda}_i^T \nabla_{\mathbf{x}_i}\mathbf{c}_{ii} = 0, \quad i = 2, 3, \cdots, n-1, \tag{5.4.29}$$

$$-\frac{1}{n}\mathbf{r}_n^T O + \boldsymbol{\lambda}_{n-1}^T \nabla_{\mathbf{x}_n}\mathbf{c}_{(n-1)n} = 0, . \tag{5.4.30}$$

The Newton equations determining corrections $\mathbf{dx}_c, \mathbf{d\lambda}_c$ to current estimates of state and multiplier vector solutions of these equations are:

$$\nabla^2_{\mathbf{x}}\mathcal{L}_n \mathbf{dx}_c + \nabla^2_{\mathbf{x\lambda}}\mathcal{L}_n \mathbf{d\lambda}_c = -\nabla_{\mathbf{x}}\mathcal{L}_n^T, \tag{5.4.31}$$

$$\nabla_{\mathbf{x}}\mathbf{c}\left(\mathbf{x}_c\right)\mathbf{dx}_c = C\mathbf{dx}_c = -\mathbf{c}\left(\mathbf{x}_c\right), \tag{5.4.32}$$

where (setting $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}_n = 0$ and making use of the block separability of the Lagrangian)

$$\nabla^2_{\mathbf{x}}\mathcal{L}_n = \operatorname{diag}\left\{\frac{1}{n}O^TO - \left(\boldsymbol{\lambda}_{i-1} + \boldsymbol{\lambda}_i\right)^T \frac{\Delta t}{2}\nabla^2_{\boldsymbol{x}_i}\mathbf{f}\left(t_i, \mathbf{x}_i\right), \quad i = 1, 2, \cdots, n\right\}, \tag{5.4.33}$$

$$\nabla^2_{\boldsymbol{\lambda}\mathbf{x}}\mathcal{L}_n = C^T, \tag{5.4.34}$$

$$C_{ii} = -I - \frac{\Delta t}{2}\nabla_{\mathbf{x}_i}\mathbf{f}\left(t_i, \mathbf{x}_i\right), \tag{5.4.35}$$

$$C_{i(i+1)} = I - \frac{\Delta t}{2}\nabla_{\mathbf{x}_{i+1}}\mathbf{f}\left(t_{i+1}, \mathbf{x}_{i+1}\right). \tag{5.4.36}$$

Note that the choice of the trapezoidal rule makes $\nabla^2_{\mathbf{x}}\mathcal{L}_n$ block diagonal, and that the constraint matrix $C : R^{nm} \to R^{(n-1)m}$ is block bidiagonal.

Equations (5.4.31) and (5.4.32) are often rewritten using the linear dependence of the Lagrangian on $\boldsymbol{\lambda}_c$ as

$$\nabla^2_{\mathbf{x}}\mathcal{L}_n \mathbf{dx}_c + \nabla^2_{\mathbf{x\lambda}}\mathcal{L}_n \boldsymbol{\lambda}^u_c = -\nabla_{\mathbf{x}}\Phi^T, \tag{5.4.37}$$

$$\nabla_{\mathbf{x}}\mathbf{c}\left(\mathbf{x}_c\right)\mathbf{dx}_c = C\mathbf{dx}_c = -\mathbf{c}\left(\mathbf{x}_c\right), \tag{5.4.38}$$

where $\boldsymbol{\lambda}^u_c = \boldsymbol{\lambda}_c + \mathbf{d\lambda}_c$. This practice is followed in developing algorithms for sequential quadratic programming in [73] for example.

A priori the assumption 5.5 that the estimation problem has a well determined solution implies a well determined solution of the Newton equations for good enough initial approximations. In this optimization context this is equivalent to C possessing full row rank, and to the second order sufficiency conditions holding [73]. The possible catch here arises because the augmented matrix associated with equations (5.4.31), (5.4.32) is symmetric but indefinite. The Bunch-Parlett algorithm [46] provides a suitable solution procedure for such systems as it possesses similar stability properties to complete pivoting. It has the disadvantage that it achieves its stability by an interchange strategy that has the possibility of destroying the considerable sparsity structure of the augmented matrix in this case. For this reason strategies which exploit the sparsity structure to allow systematic elimination of variables in a fixed order are popular. However, this amounts to the use of fixed pivoting sequences, and this introduces the possibility of numerical

instability. Implications of this strategy for several variations of systematic elimination strategies are considered below.

Choice of starting data in the simultaneous method requires initial estimates of $\mathbf{x}$ and $\boldsymbol{\lambda}$ in addition to initial estimates of the parameter values. There is some structure in $\boldsymbol{\lambda}$ which follows from (5.4.29). Note that

$$\nabla_{\mathbf{x}_i}\mathbf{c}_i = -I - \frac{\Delta t}{2}\nabla_{\mathbf{x}_i}\mathbf{f}_i,\ \nabla_{\mathbf{x}_{i+1}}\mathbf{c}_i = I - \frac{\Delta t}{2}\nabla_{\mathbf{x}_{i+1}}\mathbf{f}_{i+1}.$$

Grouping terms in (5.4.29) gives the recurrence

$$-\boldsymbol{\lambda}_{i-1} + \boldsymbol{\lambda}_i - \frac{\Delta t}{2}\nabla_{\mathbf{x}_i}\mathbf{f}\left(t_i, \mathbf{x}_i\right)^T\left(\boldsymbol{\lambda}_{i-1} + \boldsymbol{\lambda}_i\right) = -\frac{1}{n}\mathcal{O}^T\mathbf{r}_i, \tag{5.4.39}$$

while equations (5.4.28) and (5.4.30) provide boundary conditions both on this recurrence and the discretization of the target differential equation. For simplicity consider the case where $\Delta t = O\left(\frac{1}{n}\right)$, r_i is a scalar and the observation structure is based on a vector representer $\mathbf{o}^T$. Then

$$\begin{aligned}\mathcal{O}^T r_i &= \left\{\varepsilon_i + \mathbf{o}^T\left(\mathbf{x}_i^* - \mathbf{x}_i\right)\right\}\mathbf{o}, \\ &= \sqrt{n}\left\{\frac{\varepsilon_i}{\sqrt{n}} + \frac{1}{\sqrt{n}}\mathbf{o}^T\left(\mathbf{x}_i^* - \mathbf{x}_i\right)\right\}\mathbf{o}.\end{aligned} \tag{5.4.40}$$

Let $\mathbf{w}_i = \sqrt{n}\boldsymbol{\lambda}_i,\ i = 1, 2, \cdots, n-1$, then equation (5.4.39) becomes

$$-\mathbf{w}_{i-1} + \mathbf{w}_i - \frac{\Delta t}{2}\nabla_{\mathbf{x}_i}\mathbf{f}\left(t_i, \mathbf{x}_i\right)^T\left(\mathbf{w}_{i-1} + \mathbf{w}_i\right) = -\frac{r_i}{\sqrt{n}}\mathbf{o}. \tag{5.4.41}$$

It follows from (5.4.40) evaluated at $\widehat{\mathbf{x}}_n,\ \widehat{\boldsymbol{\lambda}}_n$ that the stochastic component of the forcing term is normally distributed by assumption and has variance $\left(\sigma^2/n\right)\mathbf{o}\mathbf{o}^T$. The remaining right hand side term is small as a result of the dependence on $\mathbf{o}^T\left(\mathbf{x}_i^* - \widehat{\mathbf{x}}_i\right)$. Here the equivalence between the embedding and simultaneous methods is valuable .This shows that the error term in the maximum likelihood estimate of the augmented parameter vector in the embedding formulation has a leading error term which is normally distributed and has scale $O(\frac{1}{\sqrt{n}})$ (Theorem 5.7). Further, the discussion of consistency shows that the truncation error in the trapezoidal rule leads to a smaller order correction term in the maximum likelihood estimate. It follows that the contribution of the $\mathbf{o}^T\left(\mathbf{x}_i^* - \widehat{\mathbf{x}}_i\right)$ term to the solution of (5.4.41) is of smaller order than the $\boldsymbol{\varepsilon}$ term. Thus the assumption of normality together with the scale of the variance permits the identification of (5.4.39) with a consistent discretization of the adjoint to the linearised constraint differential equation system subject to a forcing term which contains a Wiener process

component(compare equations (6.2.3) - (6.2.6)) [58]. The interesting and important consequence is that (5.4.41) indicates that the multipliers $\boldsymbol{\lambda}_i \overset{a.s.}{\underset{n\to\infty}{\to}} 0,\ i = 1, 2, \cdots, n-1$, on a scale which is $O\left(n^{-1/2}\right)$. Note also that the scale of the forcing term depends on σ so that the multipliers will be small when the experimental error is small.

Remark 5.4.2 *This estimate of the asymptotic behaviour of the multipliers is consistent with the conditions (5.4.28) and (5.4.30). Only one of these is needed to determine the recurrence with the satisfaction of the second being equivalent to satisfying the optimality conditions. The occurrence of a form of the adjoint differential equation in the necessary conditions is reminiscent of the Pontryagin maximum principle [66].*

Example 5.4.4 *The effect of the random walk term can be isolated in the smoothing problem with data:*

$$\begin{aligned}
\frac{d\mathbf{x}}{dt} &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{x}, \\
y_i &= \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x}_i + \varepsilon_i = 1 + \varepsilon_i,\ \varepsilon_i \sim N(0,1), \\
t_i &= \frac{(i-1)}{(n-1)},\ i = 1, 2, \cdots, n,
\end{aligned}$$

where the data corresponds to the exact solution $\mathbf{x}(t) = \mathbf{e}_1$. The trapezoidal rule is exact for this differential equation so the truncation error component is absent in (5.4.40). The scaled solution $\mathbf{w}_i,\ i = 1, 2, \cdots, n-1$ obtained for a particular realisation of the ε_i for $n = 501,\ \sigma = 5.$ is plotted in figure 5.9. The relation between the scale of the standard deviation σ and that of $\mathbf{w}$ seems typical. This provides a good illustration that the $n^{-1/2}$ scaling leads to an $O(1)$ result. Note the dominance of the first (red) component of the scaled multiplier vector in (5.4.41) is compatible with the structure of the observation vector $\mathbf{o} = \mathbf{e}_1$. The second component corresponds to a summed version of the first and shows significant cancellation.

The above observations suggest that $\boldsymbol{\lambda}_c = 0$ could be a suitable initial choice for the Lagrange multipliers in the simultaneous method . Another estimation possibility starts with an initial guess to the state vector $\mathbf{x}_c$ and estimates $\boldsymbol{\lambda}_c$ by minimizing $\|\nabla \mathcal{L}_n\|_2^2$. This latter approach could prove attractive when an initial solution of the boundary problem is made, as in the embedding approach, in order to use an adaptive procedure to introduce an appropriately graded mesh.

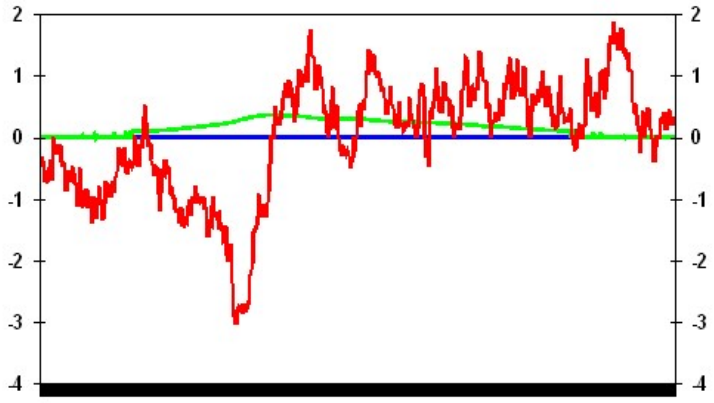

Figure 5.9: Scaled multiplier vector components

5.4.4 Computational considerations in the simultaneous methods

The embedding method has the advantages both of straight forward implementation and an effective minimization procedure provided by the Gauss-Newton method. The simultaneous method has a deeper mathematical programming setting which has its own advantages to set against additional complexity, being in principle capable of including the real world constraints required by industrial and commercial applications into the algorithmic setting. This has spawned several full scale mathematical programming packages including IPOPT [107] which has, at least in part, grown out of implementations of the simultaneous method as described here.

There are two basic approaches to using the structure of the Newton equations (5.4.31), (5.4.32) in order to produce compact solution procedures for the simultaneous method. These are the null-space method and the elimination methods. Both partition the unknowns into blocks that can be computed compactly in a sequence of steps, and it is this partitioning that is equivalent to using a fixed pivoting sequence in the solution of the augmented matrix equations. If the partitioning is done by an orthogonal transformation then the result is known as the null-space method . If unknowns are eliminated directly using the linearised constraint equations then the resulting approaches are elimination methods . Both approaches are standard methods in sequential quadratic programming . Also considered here is a modification of the Newton iteration due to Bock [12] which has some superficial resemblance to the Gauss-Newton method in that it avoids the computation of second derivative information. Here it proves convenient to consider the Bock method in

the context of the null-space method, and it is shown that it has a similar large sample convergence rate to the Gauss-Newton method when the errors are independent and normally distributed. However, note the requirement that the errors be normally distributed. This stands in contrast to the Gauss-Newton method where they are required only to be independent and have bounded variance.

A null-space method

This is related to the solution procedure for generalised least squares problems discussed in Chapter 1, Section 1.2, but the structure and interpretation of the component terms is somewhat different. In the present context it provides one of the basic solution strategies used in implementing algorithms for sequential quadratic programming problems [73]. Let $C^T = Q \begin{bmatrix} U \\ 0 \end{bmatrix}$, where Q is orthogonal and $U : R^{(n-1)m} \to R^{(n-1)m}$ is upper triangular, $Q = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}$, $Q_1 : R^{(n-1)m} \to R^{nm}$, $Q_2 : R^m \to R^{nm}$. Then the Newton equations (5.4.31), (5.4.32) can be written

$$\begin{bmatrix} Q^T\nabla^2_{\mathbf{x}}\mathcal{L}_n Q & \begin{bmatrix} U \\ 0 \end{bmatrix} \\ \begin{bmatrix} U^T & 0 \end{bmatrix} & 0 \end{bmatrix} \begin{bmatrix} Q^T\mathbf{dx}_c \\ \mathbf{d\lambda}_c \end{bmatrix} = \begin{bmatrix} -Q^T\nabla_{\mathbf{x}}\mathcal{L}_n^T \\ -\mathbf{c} \end{bmatrix}. \tag{5.4.42}$$

The solution of this system can be found by solving in sequence:

$$U^T\left(Q_1^T\mathbf{dx}_c\right) = -\mathbf{c}, \tag{5.4.43}$$

$$Q_2^T\nabla^2_{\mathbf{x}}\mathcal{L}_n Q_2\left(Q_2^T\mathbf{dx}_c\right) = -Q_2^T\left(\nabla^2_{\mathbf{x}}\mathcal{L}_n Q_1\left(Q_1^T\mathbf{dx}_c\right) + \nabla_{\mathbf{x}}\mathcal{L}_n^T\right), \tag{5.4.44}$$

$$U\mathbf{d\lambda}_c = -Q_1^T\left(\nabla^2_{\mathbf{x}}\mathcal{L}_n\mathbf{dx}_c + \nabla_{\mathbf{x}}\mathcal{L}_n^T\right). \tag{5.4.45}$$

It is necessary to form $Q_2^T\nabla^2_{\mathbf{x}}\mathcal{L}_n Q_2$ explicitly. If second order sufficiency holds then its invertibility at $\mathbf{x}_c = \mathbf{x}_c^*$ is guaranteed. However, Q must be kept as a product of its component transformations to preserve sparsity. Similarly, formation of $Q_2^T\nabla^2_{\mathbf{x}}\mathcal{L}_n Q_1$ explicitly is to be avoided.

Here the fixed pivot sequence involves first the calculation of $Q_1^T\mathbf{dx}_c$. This computation depends on the differential equation discretization only. Stability of this step is easily checked for the trapezoidal rule approximation of the Mattheij example. If $\mathbf{x}_c = 0$ then $Q_1^T\mathbf{dx}_c$ approximates $Q_1^T\mathbf{x}_c^*$ where $\mathbf{x}_i^* = \begin{bmatrix} e^{t_i} \\ e^{t_i} \\ e^{t_i} \end{bmatrix}$. This follows because $Q_1^T\mathbf{dx}_c$ is independent of boundary conditions on the differential equation. It thus must approximate Q_1^T times corresponding sample values of the particular integral. Results for the

interesting case $n = 11$ are given in table 5.6. Approximate and exact results are identical to five figures for $n = 101$. This confirms the di-stability of the discretization has been carried over to this implementation despite the factorization being carried out on the transpose of the usual matrix. The second step involves the inversion of $Q_2^T \nabla_{xx}^2 \mathcal{L}_n Q_2$ in order to compute $Q_2^T \mathbf{dx}_c$. Here second order sufficiency guarantees stability as a consequence of the assumption that the problem is well posed. This step completes the computation of $\mathbf{dx}_c$. The final step computes $\mathbf{d\lambda}_c$. This computation solves a least squares problem using an orthogonal factorization of the design matrix and possesses a stability that is familiar in other contexts [39]. Computational experience with the Mattheij problem has proved satisfactory. Thus the suggested conclusion is that the null space method proves satisfactory for boundary value stable methods, and may have potentially greater applicability. It is a point of some interest that the splitting of the computation of the state variables is based on an orthogonal transformation. This is in contrast to the splittings used in variants of the elimination method.

test results $n = 11$	particular integral $Q_1^T x$
.87665 -.97130 -1.0001	.87660 -.97134 -1.0001
.74089 -1.0987 -1.3432	.74083 -1.0988 -1.3432
.47327 -1.2149 -1.6230	.47321 -1.2150 -1.6231
.11498 -1.3427 -1.8611	.11491 -1.3428 -1.8612
-.32987 -1.4839 -2.0366	-.32994 -1.4840 -2.0367
-.85368 -1.6400 -2.1250	-.85376 -1.6401 -2.1250
-1.4428 -1.8125 -2.1018	-1.4429 -1.8125 -2.1019
-2.0773 -2.0031 -1.9444	-2.0774 -2.0032 -1.9444
-2.7309 -2.2137 -1.6330	-2.7310 -2.2138 -1.6331
-3.3719 -2.4466 -1.1526	-3.3720 -2.4467 -1.1527

Table 5.6: Stability test on null space method

Example 5.4.5 *Mattheij Problem test. Figure 5.10 shows state variable and scaled multiplier plots for a Newton's method implementation of the null space approach. This differs from the Gauss-Newton iteration suggested in the embedding method in incorporating a full Hessian evaluation. These results complement the embedding results presented in Example 5.4.1. The data for the estimation problem is based on the observation functional representer* $\mathcal{O} =$

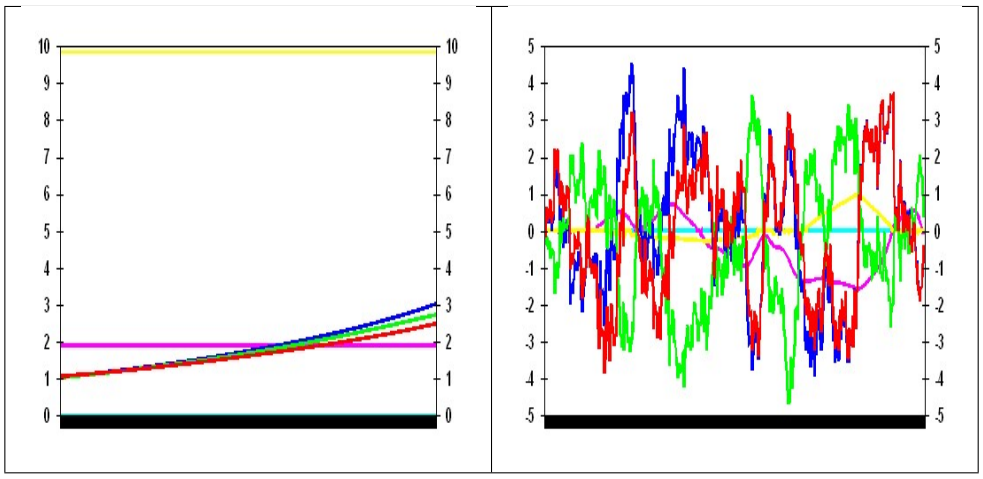

Figure 5.10: State variables and multiplier values for Mattheij Problem

$\begin{bmatrix} .5 & 0 & .5 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 \end{bmatrix}$ *with the true signal values being perturbed by random normal values having standard deviation* $\sigma = .5$. *The number of observations generated is* $n = 501$. *The initial values of the state variables are perturbed from their true values by up to 10%. The initial parameter values correspond to the true values* 10, 2 *perturbed also by up to 10%. Very rapid convergence (4 iterations) is obtained.*

The Bock iteration

The Newton iteration works with the augmented matrix appropriate to the problem. This is necessarily indefinite even if $\nabla_x^2 \mathcal{L}_n$ is positive definite. However, linear constraints make no contribution to the Hessian of the Lagrangian and the second derivative terms arising from the constraints are $O(1/n)$ through the factor Δt. Thus their contribution is smaller than that of the terms arising from the objective function when the $O\left(1/n^{1/2}\right)$ scale appropriate for the Lagrange multipliers is taken into account. Also, it is required that the initial augmented matrix be nonsingular if $\boldsymbol{\lambda}_c = 0$ is an acceptable initial estimate. This suggests that ignoring the strict second derivative contribution from the constraints could lead to an iteration with asymptotic convergence properties similar to those of the Gauss-Newton method in least squares problems where the ignored terms have a similar $O(\frac{1}{\sqrt{n}})$ dependence. This behaviour in what will be called the Bock algorithm has been observed by, for example, [16], [12] where the original observation was made, [13], and [63]. However, the resulting iteration differs from Gauss-Newton in a number

of significant ways. For example, the discussion of the asymptotic rate of the Gauss-Newton iteration requires independence and bounded variance while here the multiplier estimate makes explicit use of the additional assumption that the measurement errors are normally distributed .

Theorem 5.8 is developed to support these observations. It can be seen as an analogue of Theorem 4.5. It makes use of the detailed structure of the iteration matrix. Let H_n be the matrix of the approximate Hessian associated with the Bock iteration corresponding to the augmented matrix in the Newton iteration. Then

$$H_n = \begin{bmatrix} \frac{1}{n}\operatorname{diag}\left\{\mathcal{O}^T\mathcal{O},\ i=1,2,\cdots,n\right\} & C^T \\ C & 0 \end{bmatrix}. \tag{5.4.46}$$

It is convenient to define the composite vector $\mathbf{s}(\boldsymbol{\lambda}_c)$ by

$$\mathbf{s}(\boldsymbol{\lambda})_i = \boldsymbol{\lambda}_{i-1} + \boldsymbol{\lambda}_i,\ i=1,2,\cdots,n,$$

where $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}_n = 0$. Note that, as a consequence of the smoothing form of the estimation problem,

$$\nabla_x^2 \mathcal{L}_n(\mathbf{x}_c, \boldsymbol{\lambda}_c) = \frac{1}{n}\operatorname{diag}\left\{\mathcal{O}^T\mathcal{O},\ i=1,2,\cdots,n\right\} - \frac{\Delta t}{2}\mathcal{B}(\mathbf{x}, \boldsymbol{\lambda}),$$

where

$$\mathcal{B}(\mathbf{x}, \boldsymbol{\lambda}) = \operatorname{diag}\left\{\nabla_x^2 \mathbf{s}(\boldsymbol{\lambda})_i^T \mathbf{f}(t_i, \mathbf{x}_i),\ i=1,2,\cdots,n\right\}.$$

The main algebraic results required are summarised in the following Lemmas.

Lemma 5.3 *Let G be an invertible matrix given by:*

$$G = \begin{bmatrix} A & B & C \\ D & E & 0 \\ F & 0 & 0 \end{bmatrix},$$

then

$$G^{-1} = \begin{bmatrix} 0 & 0 & F^{-1} \\ 0 & E^{-1} & -E^{-1}DF^{-1} \\ C^{-1} & -C^{-1}BE^{-1} & C^{-1}BE^{-1}DF^{-1} - AF^{-1} \end{bmatrix}$$

Lemma 5.4 *Let G be the invertible matrix defined in Lemma 5.3 , and let W have the sparsity pattern:*

$$W = \begin{bmatrix} R & S & 0 \\ T & U & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

then

$$G^{-1}W = \begin{bmatrix} 0 & 0 & 0 \\ E^{-1}T & E^{-1}U & 0 \\ C^{-1}R - C^{-1}BE^{-1}T & C^{-1}S - C^{-1}BE^{-1}U & 0 \end{bmatrix}$$

Lemma 5.5 *Let G and W have the structure defined in Lemmas 5.3 and 5.4. Then:*

$$\varpi\left\{G^{-1}W\right\} = \varpi\left\{E^{-1}U\right\}.$$

where $\varpi(\cdot)$ denotes the spectral radius of the indicated matrix.

The fixed point form for the Bock iteration is

$$\begin{bmatrix} \mathbf{x}_c \\ \boldsymbol{\lambda}_c \end{bmatrix}_{i+1} = F_n\left(\begin{bmatrix} \mathbf{x}_c \\ \boldsymbol{\lambda}_c \end{bmatrix}_i\right),$$
$$F_n\left(\begin{bmatrix} \mathbf{x}_c \\ \boldsymbol{\lambda}_c \end{bmatrix}\right) = \begin{bmatrix} \mathbf{x}_c \\ \boldsymbol{\lambda}_c \end{bmatrix} - H_n^{-1}\begin{bmatrix} \nabla_x \mathcal{L}_n^T \\ \mathbf{c} \end{bmatrix}.$$

The condition for $\begin{bmatrix} \widehat{\mathbf{x}}_c \\ \widehat{\boldsymbol{\lambda}}_c \end{bmatrix}$ to be an attractive fixed point of the Bock iteration is that

$$\varpi\left\{F_n'\left(\begin{bmatrix} \widehat{\mathbf{x}}_c \\ \widehat{\boldsymbol{\lambda}}_c \end{bmatrix}\right)\right\} = \varpi\left\{H_n^{-1}\begin{bmatrix} -\frac{\Delta t}{2}\mathcal{B}\left(\widehat{\mathbf{x}}_c, \widehat{\boldsymbol{\lambda}_c}\right) & 0 \\ 0 \quad 0 & 0 \end{bmatrix}\right\} < 1.$$

Theorem 5.8 *Assume that the solutions of the recurrence (5.4.41) are bounded almost surely. Also, let $S = \begin{bmatrix} S_1 & S_2 \end{bmatrix}$ be the orthogonal transformation that takes C^T to upper triangular form in the null space method, and assume*

$$\left\|\left(S_2^T \operatorname{diag}\left\{\mathcal{O}^T\mathcal{O},\ i = 1, 2, \cdots, n\right\} S_2\right)^{-1}\right\| = O(1). \tag{5.4.47}$$

Then

$$\varpi\left\{F_n'\left(\begin{bmatrix} \widehat{\mathbf{x}}_c \\ \widehat{\boldsymbol{\lambda}}_c \end{bmatrix}\right)\right\} \overset{a.s.}{\underset{n\to\infty}{\to}} 0.$$

Proof. The orthogonal similarity transform of F_n' by $\begin{bmatrix} S & 0 \\ 0 & I \end{bmatrix}$ leaves the spectral radius invariant. Set $S^* = \begin{bmatrix} S & 0 \\ 0 & I \end{bmatrix}$. Then it follows that

$$\varpi\left\{F_n'\left(\begin{bmatrix} \widehat{\mathbf{x}}_c \\ \widehat{\boldsymbol{\lambda}}_c \end{bmatrix}\right)\right\} = \varpi\left\{\left(\left(S^*\right)^T H_n S^*\right)^{-1}\left(S^*\right)^T\begin{bmatrix} \frac{\Delta t}{2}\mathcal{B}\left(\widehat{\mathbf{x}}_n, \widehat{\boldsymbol{\lambda}}_n\right) & 0 \\ 0 \quad 0 & 0 \end{bmatrix} S^*\right\}$$

An application of Lemmas 5.3–5.5 permits the identifications

$$\varpi\left\{F_n'\left(\left[\begin{array}{c}\widehat{\mathbf{x}}_c\\ \widehat{\boldsymbol{\lambda}}_c\end{array}\right]\right)\right\}=\varpi\left\{\left(S_2^T\,\mathrm{diag}\left\{\frac{1}{n}\mathcal{O}^T\mathcal{O},\ i=1,\cdots,n\right\}S_2\right)^{-1}\right.$$
$$\left.S_2^T\frac{\Delta t}{2}\mathcal{B}\left(\widehat{\mathbf{x}}_c,\widehat{\boldsymbol{\lambda}}_c\right)S_2\right\}$$

As $n\Delta t=O\left(1\right)$ it follows from (5.4.47) that it is necessary only to show that

$$\left\|S_2^T\mathcal{B}\left(\widehat{\mathbf{x}}_c,\widehat{\boldsymbol{\lambda}}_c\right)S_2\right\|\overset{a.s.}{\underset{n\to\infty}{\to}}0.$$

Here the spectral norm (which is equal to the spectral radius by symmetry) is dominated by the spectral radius of the symmetric, $m\times m$ block, block diagonal matrix $\mathcal{B}\left(\widehat{\mathbf{x}}_c,\widehat{\boldsymbol{\lambda}}_c\right)$. The desired result now follows because the diagonal blocks of this matrix all tend to 0 with $\widehat{\boldsymbol{\lambda}}_c,\ n\to\infty$. ■

Remark 5.4.3 *The assumption (5.4.47) has independent interest. It is clearly linked to second order sufficiency in the case that the constraints are linear so it becomes also an algebraic condition determining identifiability . In one case it is very easy. Let $\mathcal{O}=I$ and assume no explicit parameters so the problem reduces to pure smoothing. In this case the matrix to be inverted is just the identity.*

Remark 5.4.4 *Because the dimension of the augmented matrix increases with n it is not sufficient to show the elements get small like $O\left(n^{-1/2}\right)$. The key step is the focussing of attention on an $m\times m$ matrix as in the case where the parameters make up a set of fixed dimension. The problem with increasing n is illustrated by the example*

$$\varpi\left\{\frac{1}{\sqrt{n}}\mathbf{e}\mathbf{e}^T\right\}=\sqrt{n}.$$

Exercise 5.4.3 *Let* $\left\|\left[\begin{array}{c}\mathbf{x}_c\\ \boldsymbol{\lambda}_c\end{array}\right]-\left[\begin{array}{c}\widehat{\mathbf{x}}_n\\ \widehat{\boldsymbol{\lambda}}_n\end{array}\right]\right\|$ *be small. Show that the* relative error *in the Newton correction compared to the Bock correction satisfies*

$$\frac{\left\|\left[\begin{array}{c}\Delta_x\\ \Delta_\lambda\end{array}\right]\right\|}{\left\|\left[\begin{array}{c}\mathbf{dx}_c\\ \mathbf{d}\boldsymbol{\lambda}_c\end{array}\right]\right\|}=O\left(n^{-\frac{1}{2}}\right).$$

where $\begin{bmatrix} \Delta_x \\ \Delta_\lambda \end{bmatrix}$ *represent the difference between the Bock and Newton corrections computed at* $\begin{bmatrix} \mathbf{x}_c \\ \boldsymbol{\lambda}_c \end{bmatrix}$. *This result relies on the sparsity structure of the augmented matrix associated with the Newton iteration matrix, and it assumes* $O(n)$ *estimates for both* $\|U^{-1}\|$ *and* $\left\|\left(Q_2^T \nabla^2_{\mathbf{x}} \mathcal{L}_n Q_2\right)^{-1}\right\|$. *The first implied inequality is generic for consistent discretization formulae such as the trapezoidal rule, while the second is compatible with second order sufficiency and the* $1/n$ *scaling of the objective function.*

Example 5.4.6 *The Mattheij example (Example 5.4.5) provides an interesting comparison between the Newton and Bock iterations. The Bock iteration does not compare favourably with the Newton iteration when run under identical conditions in the case* $\sigma = .5$, $n = 501$ *as it diverges for many choices of seed for the random number generator. Neither test made use of a line search. However, when* σ *is reduced to* .1 *then the performance of the two methods is essentially identical and convergence very rapid. Note that the size of the Lagrange multipliers is directly proportional to the choice of* σ.

Elimination of variables

A standard approach, followed by [63] for example, is to use the linearity of (5.4.32) to eliminate components of the state vector in what is essentially a pivoting step. The resulting reduced system is then solved by a standard SQP procedure. In contrast to the null-space method which uncouples the equations by finding in sequence $Q^T \mathbf{dx}_c$, $\mathbf{d}\boldsymbol{\lambda}_c$, the elimination procedures computes first $\mathbf{dx}_c$, and then $S^T \mathbf{d}\boldsymbol{\lambda}_c$ where orthogonal S is defined by the mode of factorization in

$$C = S \begin{bmatrix} U_1 & U_{12} \end{bmatrix}. \tag{5.4.48}$$

In the first case considered $U_1 : R^{(n-1)m} \to R^{(n-1)m}$ is upper triangular, and $U_{12} : R^m \to R^{(n-1)m}$. Note that here S is made up of components that operate on individual block rows of C. Transforming (5.4.31), (5.4.32) gives

$$\begin{bmatrix} \nabla^2_{\mathbf{x}} \mathcal{L}_n & \begin{bmatrix} U_1^T \\ U_{12}^T \end{bmatrix} \\ \begin{bmatrix} U_1 & U_{12} \end{bmatrix} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{dx}_c \\ S^T \mathbf{d}\boldsymbol{\lambda}_c \end{bmatrix} = - \begin{bmatrix} \nabla_{\mathbf{x}} \mathcal{L}_n^T \\ S^T \mathbf{c} \end{bmatrix}. \tag{5.4.49}$$

Let $\mathbf{dx}_c = \begin{bmatrix} \mathbf{dx}^{(1)} \\ \mathbf{dx}^{(2)} \end{bmatrix}$ be partitioned in conformity with the factors of C then

$$\mathbf{dx}^{(1)} = -U^{-1} U_{12} \mathbf{dx}^{(2)} - U^{-1} S^T \mathbf{c}. \tag{5.4.50}$$

This expression is of interest because it gives the solution of the differential equation system in terms of a terminal value condition provided by $\mathbf{dx}^{(2)}$. This component of the calculation can be compatible with dichotomy structure in the case of rapidly varying solutions only if these correspond to rapidly increasing solutions (backward stability) . A similar point is made in section 7.2.3. of [6],

It is necessary to compute $\mathbf{dx}^{(2)},\ S^T\mathbf{d\lambda}_c$ in order to complete the solution. This information can be found from the system

$$\left[\begin{array}{cc} I & \\ -U_{12}^T U_1^{-T} & I \end{array}\right] \nabla_{\mathbf{x}}^2\mathcal{L}_n \left\{ -\left[\begin{array}{c} U_1^{-1}U_{12} \\ -I \end{array}\right] \mathbf{dx}^{(2)} - \left[\begin{array}{c} U_1^{-1}S^T\mathbf{c} \\ 0 \end{array}\right] \right\} + \left[\begin{array}{c} U_1^T \\ 0 \end{array}\right] S^T\mathbf{d\lambda}_c = -\left[\begin{array}{cc} I & \\ -U_{12}^T U_1^{-T} & I \end{array}\right] \nabla_{\mathbf{x}}\mathcal{L}_n^T. \tag{5.4.51}$$

This leads to the equation determining $\mathbf{dx}^{(2)}$:

$$\left[\begin{array}{cc} U_{12}^T U_1^{-T} & -I \end{array}\right] \nabla_{\mathbf{x}}^2\mathcal{L}_n \left[\begin{array}{c} U_1^{-1}U_{12} \\ -I \end{array}\right] \mathbf{dx}^{(2)} = -\left[\begin{array}{cc} U_{12}^T U_1^{-T} & -I \end{array}\right] \left\{ \nabla_{\mathbf{x}}^2\mathcal{L}_n \left[\begin{array}{c} U_1^{-1}S^T\mathbf{c} \\ 0 \end{array}\right] + \nabla_{\mathbf{x}}\mathcal{L}_n^T \right\}. \tag{5.4.52}$$

The computation is concluded by solving:

$$U_1^T S^T \mathbf{d\lambda}_c = -\left[\begin{array}{cc} I & 0 \end{array}\right] \left\{ \nabla_{\mathbf{x}}\mathcal{L}_n^T - \nabla_{\mathbf{x}}^2\mathcal{L}_n \left(\left[\begin{array}{c} U_1^{-1}U_{12} \\ -I \end{array}\right] \mathbf{dx}^{(2)} + \left[\begin{array}{c} U_1^{-1}S^T\mathbf{c} \\ 0 \end{array}\right] \right) \right\}. \tag{5.4.53}$$

The appearance of the explicit terminal boundary condition calculation in this form of elimination suggests that an alternative approach is required in the factorization (5.4.48) if the differential equation is not backward stable. A boundary value oriented approach corresponding to that used in the embedding method is recommended in [63] and can be formulated as follows. Start by permuting the first column of C to the last position with permutation matrix P, and then perform an orthogonal factorization as in the embedding method. This gives

$$CP = Q \left[\begin{array}{cc} U_1 & U_{12} \\ & U_2 \end{array}\right], \tag{5.4.54}$$

where now $U_1 : R^{(n-2)m} \to R^{(n-2)m},\ U_{12} : R^{2m} \to R^{(n-2)m},\ U_2 : R^{2m} \to R^m$. Setting

$$P^T\mathbf{dx}_c = \left[\begin{array}{c} \mathbf{dx}^{(1)} \\ \mathbf{dx}^{(2)} \end{array}\right], \quad \mathbf{dx}^{(2)} = \left[\begin{array}{c} \mathbf{dx}_n \\ \mathbf{dx}_1 \end{array}\right],$$

then the necessary conditions become

$$\left[\begin{array}{cc} P^T\nabla_{\mathbf{x}}^2\mathcal{L}_n P & \left[\begin{array}{cc} U_1^T & \\ U_{12}^T & U_2^T \end{array}\right] \\ \left[\begin{array}{cc} U_1 & U_{12} \\ & U_2 \end{array}\right] & \end{array}\right] \left[\begin{array}{c} P^T\mathbf{dx}_c \\ Q^T\mathbf{d\lambda}_c \end{array}\right] = -\left[\begin{array}{c} P^T\nabla_{\mathbf{x}}\mathcal{L}_n^T \\ Q^T\mathbf{c} \end{array}\right]. \quad (5.4.55)$$

The constraint equation gives

$$U_1\mathbf{dx}^{(1)} = -U_{12}\mathbf{dx}^{(2)} - Q_1^T\mathbf{c}, \quad (5.4.56)$$

$$U_2\mathbf{dx}^{(2)} = -Q_2^T\mathbf{c}. \quad (5.4.57)$$

To simplify the Lagrangian component proceed as before by pre-multiplying by $\left[\begin{array}{cc} I & \\ -U_{12}^TU_1^{-T} & I \end{array}\right]$. This gives:

$$\left[\begin{array}{cc} I & \\ -U_{12}^TU_1^{-T} & I \end{array}\right] P^T\nabla_{\mathbf{x}}^2\mathcal{L}_n P \left[\begin{array}{c} -U_1^{-1}\left(U_{12}\mathbf{dx}^{(2)} + Q_1^T\mathbf{c}\right) \\ \mathbf{dx}^{(2)} \end{array}\right]$$
$$+ \left[\begin{array}{cc} U_1^T & \\ & U_2^T \end{array}\right] Q^T\mathbf{d\lambda}_c = -\left[\begin{array}{cc} I & \\ -U_{12}^TU_1^{-T} & I \end{array}\right] P^T\nabla_{\mathbf{x}}\mathcal{L}_n^T.$$

Thus, in addition to (5.4.57), the pair $\mathbf{dx}^{(2)}$, $Q_2^T\mathbf{d\lambda}_c$ satisfy the equation

$$\left[\begin{array}{cc} U_{12}^TU_1^{-T} & -I \end{array}\right] P^T\nabla_{\mathbf{x}}^2\mathcal{L}_n P \left[\begin{array}{c} U_1^{-1}U_{12} \\ -I \end{array}\right] \mathbf{dx}^{(2)} + U_2^TQ_2^T\mathbf{d\lambda}_c$$
$$= -\left[\begin{array}{cc} U_{12}^TU_1^{-T} & -I \end{array}\right] \left(P^T\nabla_{\mathbf{x}}^2\mathcal{L}_n P \left[\begin{array}{c} Q_1^T\mathbf{c} \\ 0 \end{array}\right] - P^T\nabla_{\mathbf{x}}\mathcal{L}_n^T\right). \quad (5.4.58)$$

The computation is now completed by solving for $Q_1^T\mathbf{d\lambda}_c$:

$$U_1^TQ_1^T\mathbf{d\lambda}_c = -\left[\begin{array}{cc} I & 0 \end{array}\right] P^T\left\{\nabla_{\mathbf{x}}^2\mathcal{L}_n\mathbf{dx} + \nabla_{\mathbf{x}}\mathcal{L}_n^T\right\}, \quad (5.4.59)$$

and transforming back to recover $\mathbf{d\lambda}_c$.

5.5 Appendix: Conditioning of finite difference equations

The trapezoidal rule discretization of (5.2.1) gives

$$\left(I - \frac{\Delta t}{2}M_{i+1}\right)\mathbf{x}_{i+1} - \left(I + \frac{\Delta t}{2}M_i\right)\mathbf{x}_i = \frac{\Delta t}{2}\left(\mathbf{q}_{i+1} + \mathbf{q}_i\right). \quad (5.5.1)$$

It follows immediately that any solution process producing components $\mathbf{x}_i$ that are $O(1)$ must effectively be combining quantities at least n times. This already suggests that the conditioning of a well specified boundary value problem is likely to be at least $O(n)$. To verify this it is convenient to start with the exact relation

$$\mathbf{x}_{i+1} - X(t_{i+1}, t_i)\,\mathbf{x}_i = \mathbf{v}_i,$$

where $X(t_{i+1}, t_i)$ is the fundamental matrix with $X(t_i, t_i) = I$, and where the particular integral term $\mathbf{v}_i = O(1/n)$. The exact solution values of the boundary value problem on the grid $0 = t_1 < t_2 < \cdots < t_n = (n-1) * \Delta t = 1$ satisfy the system of equations

$$A\mathbf{x}_c = \begin{bmatrix} -X_1 & I & & & \\ & -X_2 & I & & \\ & & \cdots & & \\ & & & -X_{n-1} & I \\ B_1 & & & & B_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \vdots \\ \mathbf{v}_n \end{bmatrix} = \mathbf{v}_c. \quad (5.5.2)$$

Here $X_i = X(t_{i+1}, t_i)$. The $\|*\|_\infty$ norm of A is clearly $O(1)$. To estimate the norm of $A^{-1} = G$ consider

$$AG_{*i} = \begin{bmatrix} 0 \\ \vdots \\ I \\ \vdots \\ 0 \end{bmatrix}, \quad (5.5.3)$$

where the $m \times m$ unit matrix occurs in block i of the right hand side. This equation can be solved by a forward recursion which gives, for $i = 2, 3, \cdots, n$,

$$\begin{aligned} G_{ji} &= \prod_{s=j-1}^{1} X_s G_{1j},\ j < i, \\ &= \prod_{s=i}^{1} X_s G_{1i} + I,\ j = i+1, \\ &= \prod_{s=j-1}^{1} X_s G_{1i} + \prod_{s=j-1}^{i} X_s,\ j > i+1. \end{aligned}$$

The boundary value conditions expressed in the last block row of (5.5.2) now give

$$[B_1 + B_2 X(1, 0)]\, G_{1i} = -B_2 \prod_{s=n-1}^{i} X_s.$$

In particular, this gives for $j < i$,

$$G_{ji} = -X(t_j, 0)\left(B_1 + B_2 X(1,0)\right)^{-1} B_2 X(1, t_i) = G(t_j, t_i),$$

where $G(t_j, t_i)$ is the Green's matrix defined for $t_j < t_i$ in (5.2.44). Also, equation (5.5.3) can be solved for G_{ni} using a backward recursion. This proves appropriate for G_{ji} with $j > i$. The relevant equations are

$$\begin{aligned} G_{ji} &= \prod_{s=j}^{n-1} X_s^{-1} G_{ni},\ j > i, \\ &= \prod_{s=i-1}^{n-1} X_s^{-1} G_{ni} - X_{i-1}^{-1},\ j = i-1, \\ &= \prod_{s=i-1}^{n-1} X_s^{-1} G_{ni} - \prod_{s=j}^{n-1} X_s^{-1},\ j < i-1. \end{aligned}$$

G_{ni} can now be found from the boundary conditions. This gives

$$\left(B_1 X(1,0)^{-1} + B_2\right) G_{ni} = B_1 \prod_{s=1}^{i-1} X_s^{-1}.$$

For $j > i$ this gives

$$\begin{aligned} G_{ji} &= X(t_j, 0)\left(B_1 + B_2 X(1,0)\right)^{-1} B_1 X(t_i, 0)^{-1}, \\ &= G(t_j, t_i),\ j > i. \end{aligned}$$

where $G(t_j, t_i)$ is the Green's matrix defined in (5.2.43). An immediate consequence is

$$\left\|A^{-1}\right\|_\infty = \|G\|_\infty \leq n\alpha, \tag{5.5.4}$$

where α is the stability constant (5.2.45).

This result applies also to the trapezoidal rule approximation for small enough Δt. It suffices to write (5.5.1) as

$$\mathbf{x}_{i+1} - \left(I - \frac{\Delta t}{2} M_{i+1}\right)^{-1} \left(I + \frac{\Delta t}{2} M_i\right) \mathbf{x}_i = \frac{\Delta t}{2} \left(I - \frac{\Delta t}{2} M_{i+1}\right)^{-1} (\mathbf{q}_{i+1} + \mathbf{q}_i).$$

The norm inequality is a consequence of this equation and has the very similar bound because

$$X(t_{i+1}, t_i) = \left(I - \frac{\Delta t}{2} M_{i+1}\right)^{-1} \left(I + \frac{\Delta t}{2} M_i\right) + O\left(\Delta t^3\right).$$

Remark 5.5.1 *This result extends to consistent discretizations, especially those that correspond to diagonal Padé approximants generated by collocation formulae [6]. However, it needs to be noted that this is an asymptotic result valid for small enough Δt. It does not immediately provide direct information about conditioning of (for example) the trapezoidal rule in regions where advantage is being taken of di-stability in order to work with relatively large Δt so that the discretization could possess large components through the occurrence of super stability.*

Chapter 6

Some nonparametric estimation techniques in parameter estimation problems

6.1 Introduction

Up until now the assumption of a correct model has been made explicitly. This assumption has been made in the context of a parametric approach in the sense that the model is characterised by the truth represented by a particular set of values given to a finite dimensional vector of parameters. These values are assumed to specify the unique member of a class of explicitly defined systems which generates the possible observed data sets under the given experimental conditions. This vector of parameters is to be estimated from the particular realisation of the data provided. In contrast, a strict nonparametric model would be one that smooths the given data by removing the observational noise from the underlying signal without, in any sense, attempting to find a deeper mechanistic interpretation for the fitted results. For example, such a nonparametric procedure is provided by the local piecewise polynomial

$$z(t) = \sum_{i=1}^{k} P_i(t) K\left(\frac{t - t_i}{\Delta_i}\right),$$

where the P_i are polynomials of fixed, low degree with coefficients to be determined by regressing $z(t)$ on the given data y_j, $j = 1, 2, \cdots, n$, and where $K\left(\frac{t-t_i}{\Delta_i}\right)$ is a kernel or spread function (typically nonnegative) which has a peak at $t = t_i$, and a spread or support depending on Δ_i which characterizes how K decays with distance from $t = t_i$. This class of functions

has many admirable properties including an optimal rate of convergence of $O\left(k^{-1}\right)$ as the data and the number of kernel functions are increased in an appropriately controlled fashion [28]. Using it directly to represent the differential equation solution in the parameter estimation problems considered in the previous chapter is somewhat more problematical. For example, it is necessary to use the observed data to estimate the coefficients in the $P_i\left(t\right)$ which can be substituted into the differential system to produce a vector of residuals as a key step to computing an objective function on which to base the estimation of the auxiliary parameters. Generation of these approximate solution functions would not be all that dissimilar to the calculations used to test the suitability of the form of observation data (Remark 5.1.2). That is given an estimate of $\mathbf{h}^T\mathbf{x}(t) = \eta(t)$ and all necessary derivatives estimate $\mathbf{x}(t)$. An attractive feature in theory is that there is no a priori restriction to linear estimation problems. Calculation of the piecewise polynomial functions is a linear step and the estimation phase is typically an unconstrained optimisation problem. Proceeding stepwise to generate the objective function minima for a hierarchy of differential systems provides a possible basis for model selection.

Remark 6.1.1 C*are is needed in selecting the form of the objective function because simple minded procedures such as the minimization of unweighted sums of squares of residuals can be unsatisfactory in certain circumstances. The problem occurs in the first phase when the fitted estimate of the solution function results in residual errors in which the random component is correlated even if the errors in the original observations are independent. For example, the classic method of Prony [91] (3.5.13) , which can be taken as an archetype for this problem, is known to be inconsistent [54]. The correct maximum likelihood problem is given by (3.5.14) and takes explicit account of the induced correlation. The possibility of the problem occurring in kernel function estimation depends at least in part on the overlapping of the supports of neighbouring kernel functions.*

An alternative approach to the model selection problem available in some circumstances makes use of an association between a linear system of differential equations and a class of stochastic extensions of this system [109]. This approach is best known as a method for analyzing and computing smoothing splines [60] . Smoothing splines provide a data representation $\eta\left(t\right)$ given iid, normally distributed errors in the observations y_i, $i = 1, 2, \cdots, n$. The spline is found as the solution of the optimization problem

$$\min_{\eta} \sum_{i=1}^{n} \left(y_i - \eta\left(t_i\right)\right)^2 + \tau \int_0^1 \left(\frac{d^k\eta}{dt^k}\right)^2 dt, \tag{6.1.1}$$

and the resulting function represents a compromise between approximation to the observed data and smoothness as represented by the contribution of integral term, a compromise which is reflected in the value of τ. Typically it is required also to estimate τ from the given data. This procedure is often thought of as a nonparametric fitting technique. However, as τ becomes large, the optimal $\eta(t)$ is forced to approach the null space of the differential operator $\frac{d^k\eta}{dt^k}$ so that at least a limiting structure can be discerned. The key to taking this further is due to Wahba [108] who considered the stochastic differential equation

$$\frac{d^k x}{dt^k} = \sigma\sqrt{\lambda}\frac{dw}{dt} \tag{6.1.2}$$

where $w(t)$ is a unit scale Wiener process independent of the observation process, λ provides a scale for the process and corresponds to $1/\tau$ in (6.1.1), and σ corresponds to the standard deviation of the observational error. In this context, when suitably interpreted, the spline is given by

$$\eta(t) = \mathcal{E}\left\{x(t)\,|y_1, y_2, \cdots, y_n, \lambda\right\}.$$

Thus optimal mean square prediction properties can be anticipated, but this does not mean that function recovery is particularly efficient [109]. Note that the scale parameter λ typically would enter into the estimation process. Assuming the estimation process is consistent then small estimated λ implies that a linear combination of fundamental solutions of the differential operator in (6.1.2) - that is a polynomial of degree $k-1$ - provides a good model for the observed data. An immediate generalisation uses a more general differential operator instead of $\frac{d^k}{dt^k}$ to develop an analogously structured spline smoother, the g-spline . These tools can be used for model selection. For example, assume that an appropriate model is to be found among the members of a structured family of differential operators of increasing complexity such as those associated with selections of basis functions from the general separable regression model (5.2.2). For each of these selections we can construct the differential operator (5.2.3) and then fit the associated g-spline smoother to the given data. The computed values of λ provide a data driven labelling of these possible models, with the parsimonious model being the first for which the estimate of λ is insignificant in an appropriate sense.

6.2 Generalised spline formulation using the Kalman Filter

Parallel to the Wahba approach several authors have considered equivalent stochastic settings which have the advantage that the Kalman filter can be

used for computing the smoothing splines ([113], [111], [59]). In both [113] and [59] the development is in terms of g-splines . Consider the m'th order differential equation

$$\mathcal{M}x=x^{(m)}+m_1x^{(m-1)}+\cdots+m_mx. \tag{6.2.1}$$

To this equation corresponds the first order system

$$\begin{aligned}&\frac{d\mathbf{x}}{dt}=M\mathbf{x},\\ &M=\begin{bmatrix}0 & 1 & & \\ \cdots & \cdots & \cdots & \cdots\\ & & 0 & 1\\ -m_m & -m_{m-1} & \cdots & -m_1\end{bmatrix},\end{aligned} \tag{6.2.2}$$

and associated families of fundamental matrices $X(t,\xi)$ where

$$\frac{dX}{dt}=MX,\ X(\xi,\xi)=I.$$

Consider also the stochastic differential equation

$$d\mathbf{x}=M\mathbf{x}dt+\sigma\sqrt{\lambda}\mathbf{b}dw \tag{6.2.3}$$

where w is a unit Wiener process. This is the first order formulation which corresponds to the Wahba spline model (6.1.2) when the vector of coefficients in (6.2.2) is $\mathbf{m}=0$, and $\mathbf{b}=\mathbf{e}_m$. It corresponds to the limiting case as $\beta\to 0$ of

$$M=\begin{bmatrix}0 & 1\\ \beta^2 & 0\end{bmatrix},\ X(t,s)=\begin{bmatrix}\cosh\beta(t-s) & \frac{\sinh\beta(t-s)}{\beta}\\ \beta\sinh\beta(t-s) & \cosh\beta(t-s)\end{bmatrix}.$$

This specification characterises tension smoothing splines . Here the deterministic component of the dynamics equation possesses a nontrivial dichotomy so that it is unstable for $\beta>0$.

Equation (6.2.3) provides a generalisation not only to g-splines written in first order form, but also to more general differential systems. If the variation of parameters formula is used to integrate this equation then we obtain the dynamics equation

$$\mathbf{x}_{i+1}=X(t_{i+1},t_i)\mathbf{x}_i+\sigma\sqrt{\lambda}\mathbf{u}_i, \tag{6.2.4}$$

where

$$\mathbf{u}_i=\int_{t_i}^{t_{i+1}}X(t_{i+1},s)\mathbf{b}\frac{dw}{ds}ds. \tag{6.2.5}$$

From this it follows that

$$\mathbf{u}_i \backsim N\left(0, \sigma^2 R\left(t_{i+1}, t_i\right)\right),$$

where

$$R\left(t_{i+1}, t_i\right) = \lambda \int_{t_i}^{t_{i+1}} X\left(t_{i+1}, s\right) \mathbf{b}\mathbf{b}^T X\left(t_{i+1}, s\right)^T ds. \tag{6.2.6}$$

Taken in conjunction with the observation process

$$y_i = \mathbf{h}^T \mathbf{x}_i + \varepsilon_i,\ \varepsilon_i \backsim N\left(0, \sigma^2\right), \tag{6.2.7}$$

this provides the basis for the development of generalised smoothing splines using the framework provided by the Kalman filter.

6.2.1 Smoothness considerations

The smoothness properties of $\mathbf{x}\left(t|n\right)$ provide the justification for calling $\mathbf{h}^T\mathbf{x}\left(t|n\right)$ a generalised smoothing spline as they connect the developments here, which follows that given in [81], with those provided by the variational approach. The interpolation smoother (1.4.15) gives

$$\mathbf{x}\left(t|n\right) = X\left(t, t_i\right)\mathbf{x}_{i|i} + A\left(t, t_i\right)\left(\mathbf{x}_{i+1|n} - \mathbf{x}_{i+1|i}\right),\ t_i \leq t \leq t_{i+1}.$$

Thus, assuming smoothness of the coefficient matrix in the differential equation, the smoothness of $\mathbf{x}\left(t|n\right)$ depends on the smoothness of $A\left(t, t_i\right)$. To examine this it is convenient to write $A\left(t, t_i\right)$ in the form

$$A\left(t, t_i\right) = \left\{X\left(t, t_i\right) S_{i|i} X_i + \Gamma\left(t, t_i\right)\right\} S_{i+1|i}^{-1},$$

where

$$\begin{aligned}
\Gamma\left(t, t_i\right) &= \mathcal{V}\left\{\mathbf{u}\left(t, t_i\right)\right\} X\left(t_{i+1}, t\right)^T, \\
&= \sigma^2\lambda \int_{t_i}^{t} X\left(t, s\right) \mathbf{b}\mathbf{b}^T X\left(t, s\right)^T X\left(t_{i+1}, t\right)^T ds, \\
&= \sigma^2\lambda \int_{t_i}^{t} X\left(t, s\right) \mathbf{b}\mathbf{b}^T X\left(t_{i+1}, s\right)^T ds.
\end{aligned}$$

It follows that

$$\frac{dA\left(t, t_i\right)}{dt} = MA + \sigma^2\lambda \mathbf{b}\mathbf{b}^T X\left(t_{i+1}, t\right)^T S_{i+1|i}^{-1},$$

giving

$$\frac{d\mathbf{x}\left(t|n\right)}{dt} = M\mathbf{x}\left(t|n\right) + \sigma^2\lambda \mathbf{b}\mathbf{b}^T X\left(t_{i+1}, t\right)^T S_{i+1|i}^{-1}\left(\mathbf{x}_{i+1|n} - \mathbf{x}_{i+1|i}\right). \tag{6.2.8}$$

An immediate consequence is that between the data points $\mathbf{x}\left(t|n\right)$ has the same smoothness as solutions of the homogeneous differential equation.

Remark 6.2.1 *The choices* $\mathbf{h} = \mathbf{e}_1$, $\mathbf{b} = \mathbf{e}_m$ *correspond to the Wahba form for the polynomial smoothing spline which is a special case in which* M *has the structure given by (6.2.2). Multiplying (6.2.8) by* $\mathbf{e}_m^T$ *and expressing the left hand side as a differential equation for the first component of* $\mathbf{x}(t|n)$ *using*

$$\mathbf{x}^T = \left[\begin{array}{cccc} x_1 & \frac{dx_1}{dt} & \cdots & \frac{d^{m-1}x_1}{dt^{m-1}} \end{array} \right]$$

gives

$$\mathcal{M}x_1(t|n) = \sigma^2 \lambda \mathbf{e}_m^T X(t_{i+1}, t)^T S_{i+1|i}^{-1} \left(\mathbf{x}_{i+1|n} - \mathbf{x}_{i+1|i} \right), \tag{6.2.9}$$

$$= \sigma^2 \lambda \mathbf{e}_m^T X(t_{i+1}, t)^T \mathbf{d}. \tag{6.2.10}$$

The right hand side here is a general solution (parametrised by $\mathbf{d}$*) of the formal adjoint* $\mathcal{M}^+$ *of* $\mathcal{M}$*. To see this in a simple case note that*

$$\frac{dX^T(s,t)}{dt} = -M^T X^T(s,t).$$

If

$$M = \left[\begin{array}{cc} 0 & 1 \\ -m_2 & -m_1 \end{array} \right]$$

then the equation to consider is

$$\frac{d\mathbf{v}}{dt} = \left[\begin{array}{cc} 0 & m_2 \\ -1 & m_1 \end{array} \right] \mathbf{v}.$$

Expanding gives

$$\frac{dv_1}{dt} = m_2 v_2,$$
$$\frac{dv_2}{dt} = -v_1 + m_1 v_2.$$

Differentiating to eliminate v_1 *gives*

$$\frac{d^2 v_2}{dt^2} - \frac{d}{dt}(m_1 v_2) + m_2 v_2 = 0,$$

showing that v_2 *satisfies the formal adjoint equation. The general case follows by an extension of this argument and is readily verified by induction. Operating on (6.2.10) with* $\mathcal{M}^+$ *gives*

$$\mathcal{M}^+ \mathcal{M} x_1(t|n) = 0.$$

This result is due to [60] in this context. For the case of the cubic polynomial smoothing spline $m_1 = m_2 = 0$. The corresponding differential equation is

$$\frac{d^4 x_1}{dt^4} = 0,$$

recovering a well known result.

Thus the interesting points are the data points t_i. Continuity of $\frac{d\mathbf{x}(t|n)}{dt}$ at t_i requires

$$\mathbf{b}^T S_{i|i-1}^{-1} \left(\mathbf{x}_{i|n} - \mathbf{x}_{i|i-1}\right) = \mathbf{b}^T X^T \left(t_{i+1}, t_i\right) S_{i+1|i}^{-1} \left(\mathbf{x}_{i+1|n} - \mathbf{x}_{i+1|i}\right).$$

Using the smoothing formula (1.4.11) and (1.4.12) gives

$$\mathbf{x}_{i|n} - \mathbf{x}_{i|i} = S_{i|i} X^T \left(t_{i+1}, t_i\right) S_{i+1|i}^{-1} \left(\mathbf{x}_{i+1|n} - \mathbf{x}_{i+1|i}\right),$$

so that the smoothness condition becomes

$$\mathbf{b}^T S_{i|i-1}^{-1} \left(\mathbf{x}_{i|n} - \mathbf{x}_{i|i-1}\right) = \mathbf{b}^T S_{i|i}^{-1} \left(\mathbf{x}_{i|n} - \mathbf{x}_{i|i}\right). \tag{6.2.11}$$

An application of (1.4.10) gives

$$S_{i|i}^{-1} = S_{i|i-1}^{-1} + \sigma^{-2} \mathbf{h}\mathbf{h}^T,$$

so that the difference between the two sides of (6.2.11) is

$$\begin{aligned}
D &= \mathbf{b}^T \left\{ \sigma^{-2} \mathbf{h}\mathbf{h}^T \left(\mathbf{x}_{i|n} - \mathbf{x}_{i|i}\right) - S_{i|i-1}^{-1} \left(\mathbf{x}_{i|i} - \mathbf{x}_{i|i-1}\right) \right\}, && (6.2.12)\\
&= \mathbf{b}^T \left\{ \sigma^{-2} \mathbf{h}\mathbf{h}^T \left(\mathbf{x}_{i|n} - \mathbf{x}_{i|i-1}\right) - S_{i|i}^{-1} S_{i|i-1} \frac{\widetilde{y}_i}{\sigma^2 + \mathbf{h}^T S_{i|i-1} \mathbf{h}} \mathbf{h} \right\}, && (6.2.13)\\
&= \mathbf{b}^T \left\{ \sigma^{-2} \mathbf{h}\mathbf{h}^T \left(\mathbf{x}_{i|n} - \mathbf{x}_{i|i-1}\right) - \frac{\left(I + \sigma^{-2} \mathbf{h}\mathbf{h}^T S_{i|i-1}\right) \left(y_i - \mathbf{h}^T \mathbf{x}_{i|i-1}\right)}{\sigma^2 + \mathbf{h}^T S_{i|i-1} \mathbf{h}} \mathbf{h} \right\},\\
&= -\sigma^{-2} \mathbf{b}^T \mathbf{h} \left\{ y_i - \mathbf{h}^T \mathbf{x}_{i|n} \right\}, && (6.2.14)
\end{aligned}$$

where $\widetilde{y}_i$ is the innovation and the Kalman filter equations (1.4.6), (1.4.7), and (1.4.8) have been used in the simplification. This vanishes provided $\mathbf{b}^T \mathbf{h} = 0$.

To extend this result to higher derivatives note that:

1. It is necessary only to consider derivatives of the term involving $\mathbf{b}\mathbf{b}^T X^T \left(t_{i+1}, t\right)$ in (6.2.8) to find the first occurrence of a discontuity;

2. Successive derivatives of $X(t_i,t)$ can be represented by

$$\frac{d^jX(t_i,t)}{dt^j}=X(t_i,t)P_j(M),$$

where the P_j satisfy the recurrence

$$P_0=I,P_j=\frac{dP_{j-1}}{dt}-MP_{j-1},\ j=1,2,\cdots,k.$$

Paralleling the above argument now gives the result that the first k derivatives of $\mathbf{x}(t|n)$ are continuous at t_i provided

$$\mathbf{b}^TP_{j-1}(M)^T\mathbf{h}=0,\ j=1,2,\cdots,k. \tag{6.2.15}$$

If the successive vectors $P_j(M)^T\mathbf{h}, j=0,1,\cdots,m-1$ are linearly independent then at most the first $m-1$ derivatives can be continuous as any vector $\mathbf{b}$ orthogonal to m linearly dependent vectors in R^m must vanish, and this leads to a contradiction in (6.2.15) in the case $m=k$.

Remark 6.2.2 *In general the dependence of M on t would complicate the satisfaction of the conditions (6.2.15). However, they can be satisfied if M has certain structural properties. Returning to the example where a single higher order equation is reduced to a first order system by the standard transformation as discussed in Remark 6.2.1, note that $P_i(M)$ can be written*

$$P_i(M)=\begin{bmatrix}0&\cdots&0&(-1)^iI&0&\cdots&0\\ \vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\ 0&\cdots&0&0&0&\cdots&(-1)^iI\\ A^{(i)}_{(m-i+1)1}&\cdots&\cdots&\cdots&\cdots&\cdots&A^{(i)}_{(m-i+1)m}\\ \vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\ A^{(i)}_{m1}&\cdots&\cdots&\cdots&\cdots&\cdots&A^{(i)}_{mm}\end{bmatrix}$$

where the $A^{(i)}_{jk}$ are functions of the components of the coefficient vector $\mathbf{m}$ and its derivatives. It follows that

$$\begin{aligned}\mathbf{e}_m^TP_i(M)^T\mathbf{e}_1&=0,\ i=1,2,\cdots,m-2,\\ &=(-1)^i,\ i=m-1.\end{aligned}$$

This shows that the choice

$$\mathbf{b}=\mathbf{e}_m,\ \mathbf{h}=\mathbf{e}_1$$

makes $x_1 = \mathbf{h}^T\mathbf{x}$ *together with its first* $2m-2$ *derivatives continuous at the data points. This is a consequence of the definition of* $\mathbf{x}$ *which gives* $x_j = \frac{d^{j-1}x_1}{dt^{j-1}}$, $j = 1, 2, \cdots, m$ *in this case. The discontinuity in the* $(2m-1)$*'st derivative is obtained by evaluating the bracketed term in (6.2.12). The result simplifies using the Kalman filter equations to*

$$x_1^{2m-1}(t_i+) - x_1^{2m-1}(t_i-) = (-1)^m \lambda \left(y_i - \mathbf{h}^T\mathbf{x}(t_i|n)\right).$$

A direct comparison with the variational argument [94] shows that $x_1(t|n)$ *is exactly a generalised smoothing spline provided:*

1. $\lambda = 1/\tau$, *and*
2. *the natural boundary conditions*

$$\begin{aligned}\mathbf{x}(t|n) &= X(t,t_n)\,\mathbf{x}_{n|n},\ t > t_n,\\ &= X(t,t_1)\,\mathbf{x}_{1|n},\ t < t_1.\end{aligned}$$

are satisfied.

This second condition is an immediate consequence of the extrapolation (prediction) step in the Kalman filter formulation .

Example 6.2.1 *Different conditions on* $\mathbf{b}$, $\mathbf{h}$ *are obtained for the example (5.1.1). Recall that* $(\mathbf{h})_3$ *is necessary for a well posed estimation problem. Choose* $\mathbf{h} = \mathbf{e}_3$. *Then the conditions determining a twice continuously differentiable M-spline are*

$$\mathbf{h}^T\mathbf{b} = 0, \mathbf{h}^T M\mathbf{b} = 0.$$

This has the essentially unique solution, independent of β_1, β_2,

$$\mathbf{b} = \mathbf{e}_1.$$

Remark 6.2.3 *An alternative characterisation of* $\mathbf{b}$ *and* $\mathbf{h}$ *can be obtained by considering the eigenvalue decomposition of* $R(t+\delta, t)$ *in (6.2.6) in the limit as* $\delta \to 0$. *Expanding* $X(t+\delta, s)$ *about* $t+\delta$ *gives*

$$R(t+\delta,t) = \lambda \int_t^{t+\delta} \sum_{i,j} \frac{(s-(t+\delta))^{i+j}}{i!j!} P_i(M)\,\mathbf{b}\mathbf{b}^T P_j(M)^T\,ds.$$

Because successive powers of δ *are incommensurable as* $\delta \to 0$, *an application of the Rayleigh Quotient gives:*

1. *The largest eigenvalue of* $R(t+\delta, t)$ *is*

$$\pi_m = \lambda\delta \mathbf{b}^T\mathbf{b} + O\left(\delta^2\right),\ \delta \to 0, \tag{6.2.16}$$

and is associated with an eigenvector which tends to $\mathbf{b}$ *as* $\delta \to 0$*,*

2. *If the orthogonality conditions (6.2.15) are satisfied for* $i = 1, 2, \cdots, m-2$ *then the eigenvector associated with the smallest eigenvalue tends to* $\mathbf{h}$ *as* $\delta \to 0$*. The corresponding Rayleigh Quotient estimate is*

$$\pi_1 = \frac{\lambda}{((m-1)!)^2}\frac{\left(\mathbf{h}^T P_{m-1}(M)\,\mathbf{b}\right)^2}{\mathbf{h}^T\mathbf{h}}\frac{\delta^{2m-1}}{2m-1} + O\left(\delta^{2m}\right). \tag{6.2.17}$$

This is an upper bound because, while the eigenvector asymptotes to $\mathbf{h}$*, the asymptotically negligible components can still contribute to the exact eigenvalue.*

Example 6.2.2 *The simplest non trivial case corresponds to the cubic spline . Here the differential equation data is*

$$M = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},\ X(t,0) = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}.$$

The covariance matrix R *is given by*

$$\begin{aligned} \frac{1}{\lambda}R(\delta,0) &= \int_0^\delta \begin{bmatrix} 1 & \delta - s \\ & 1 \end{bmatrix} \mathbf{e}_2\mathbf{e}_2^T \begin{bmatrix} 1 & \\ \delta - s & 1 \end{bmatrix} ds, \\ &= \int_0^\delta \begin{bmatrix} \delta - s \\ 1 \end{bmatrix} \begin{bmatrix} \delta - s & 1 \end{bmatrix} ds, \\ &= \begin{bmatrix} \delta^3/3 & \delta^2/2 \\ \delta^2/2 & \delta \end{bmatrix}. \end{aligned}$$

The Rayleigh Quotient estimate of the smallest eigenvalue using $\mathbf{h} = \mathbf{e}_1$ *as test vector is* $\pi_1 = \delta^3/3$*. The characteristic equation is*

$$\lambda^2 - \left(\delta + \delta^3/3\right)\lambda + \delta^4/12 = 0.$$

The resulting eigenvalues are $\lambda = \delta + O(\delta^3)$*, and* $\lambda = \delta^3/12 + O(\delta^5)$*. The Rayleigh Quotient gets the order right but is a strict overestimate for the smallest eigenvalue.*

$\alpha = 1$	
$n = 11$	$D_i = \{8.3 - 5, 1.0 - 1\}$
$n = 51$	$D_i = \{6.7 - 7, 2.0 - 2\}$
$\alpha = 1, \beta = 2$	
$n = 11$	$D_i = \{9.9 - 13, 1.4 - 8, 8.3 - 5, 1.0 - 1\}$
$n = 51$	$D_i = \{0.0, 4.4 - 12, 6.7 - 7, 2.0 - 2\}$

Table 6.1: Tension splines provide a potentially unstable system

Example 6.2.3 *[84] Tension smoothing splines . Here the spline is constructed using exponentials rather than polynomials. This corresponds to an example with potentially unstable dynamics. For one and two parameter splines we have*

$$M = \begin{bmatrix} 0 & 1 \\ \alpha^2 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 & 0 & 0 \\ \alpha^2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \beta^2 & 0 \end{bmatrix}.$$

Smoothness is maximized by the choice $\mathbf{h} = \mathbf{e}_1$, $\mathbf{b} = \mathbf{e}_m$. *Again the covariances are illconditioned leading to very small elements in the component blocks* D_i. *These particular tension spline examples are not very unstable, but instability does not appear to be the major driver for the small elements in* D_i *in the rank revealing factorization of the* R_i. *The following table gives the computed diagonal elements produced by this decomposition for two different meshes on* $0 \leq t \leq 1$. *They are the same for each block of* D.

Example 6.2.4 *[84] Similar behaviour is shown in the case of the stable example provided by the simple chemical reaction* $A \to B \to C$ *with rates* k_1 *and* k_2. *Here the differential equation is*

$$\frac{d}{dt}\begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} -k_1 & 0 & 0 \\ k_1 & -k_2 & 0 \\ 0 & k_2 & 0 \end{bmatrix}\begin{bmatrix} A \\ B \\ C \end{bmatrix}.$$

It was seen in the previous chapter that well-posedness of the estimation problem requires $(\mathbf{h})_3 \neq 0$. *Maximum smoothness of the g-spline is achieved with* $\mathbf{b} = \mathbf{e}_1$, $\mathbf{h} = \mathbf{e}_3$. *Table 6.2 gives the diagonal elements of the rank revealing Cholesky for the same meshes as above. Again small elements are produced.*

$k_1 = 1,\ k_2 = 2$	
$n = 11$	$D_i = \{5.5 - 8, 6.8 - 5, 9.1 - 2\}$
$n = 51$	$D_i = \{1.8 - 11, 6.4 - 7, 2.0 - 2\}$

Table 6.2: A stable example from chemical kinetics

6.2.2 Smoothing spline computation

Wahba makes the assumption of a diffuse prior in order to show that the prediction given by her model does determine a smoothing spline. Here it follows from (1.4.3), and (1.4.4) that this assumption leads to the generalised least squares problem

$$\min_{\mathbf{x}} \left\{ \mathbf{r}_1^T R^{-1} \mathbf{r}_1 + \mathbf{r}_2^T V^{-1} \mathbf{r}_2 \right\}, \tag{6.2.18}$$

where

$$\begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} = \begin{bmatrix} -X_1 & I & & & & & \\ & -X_2 & I & & & & \\ & & \cdots & \cdots & \cdots & & \\ & & & & & -X_{n-1} & I \\ \mathbf{h}_1^T & & & & & & \\ & \mathbf{h}_2^T & & & & & \\ & & \cdots & \cdots & \cdots & & \\ & & & & & & \mathbf{h}_n^T \end{bmatrix} \mathbf{x} - \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix} \tag{6.2.19}$$

$$R = \sigma^2 \operatorname{diag}\{R_1, R_2, \cdots, R_{n-1}\},\ R_i = R(t_{i+1}, t_i),\ V = \sigma^2 I.$$

The values of the spline at the knots t_i are obtained from the predicted state variables $\mathbf{x}_{i|n}$ by forming $\eta_i = \mathbf{h}^T \mathbf{x}_{i|n}$. Interpolation between the knots is obtained by using the interpolation smoother (1.4.15). This formulation has the attractive feature that dependence on an initial vector has disappeared. In this sense each piece of data has equal weight. It also means that numerical solution is not dependent upon an a priori given elimination order so that there is complete freedom in seeking stable computational algorithms.This point could be important if the deterministic component of the differential equation is not stable as an initial value problem. Note that the form of scaling used permits σ^2 to be factored out of (6.2.19).

An information filter

The use of the information filter to solve (6.2.18), (6.2.19) is considered in [81], [82]. This approach has the advantage of using orthogonal transformations to factorize the matrix. These can be expected to lead to numerically

stable computations in the case that the dynamics equation has a deterministic component possessing a nontrivial dichotomy [44]. However, an initial transformation of the dynamics equations by L^{-1} where $LL^T = R$ is required. Here the estimates (6.2.17), (6.2.16), and the numerical results for examples 6.2.3 and 6.2.4 are significant because they show that R can be extremely illconditioned. That this can cause problems is exemplified in [99]. Note that the information filter described in subsection 2.3.5 (especially equation (2.3.62)) uses an ordering of the design matrix different to that in (6.2.19). This has advantages in reducing additional fill during the factorization. An alternative strategy to control sparsity is suggested for the application of V-invariant methods.

Recursive filter algorithms

The methods suggested in [59], [113], and [111] use the Kalman filter in its recursive or initial value form. To understand the connections between them recall that there is a formal equivalence between the solution of a generalised least squares problem for a (constant) parameter vector $\mathbf{x}$ and the limiting case of a prediction problem where the random solution vector is subject to a diffuse prior obtained by letting $R^{-1} \to 0$ in (1.3.25). Both possibilities have been used in smoothing spline calculations. For example, [59] and [113] initialise the Kalman filter using a diffuse prior. The form of initialisation using asymptotic expansions to cope with the unbounded terms arrising from the diffuse prior discussed in subsubsection 2.3.5 is used in [59]. Here the asymptotic expansion for the interpolation gain for $i < m$ is used in the calculation of the dependence of the state variables on all the data. An alternative approach makes use of a block calculation based on (6.2.18), (6.2.19) with $n = m$, the first value of n for which the matrix has full column rank, to obtain $\mathbf{x}_{m|m}$, $S_{m|m}$ in order to initialise the filter at the step $t_m \to t_{m+1}$. The factorization of (6.2.19) permits the smoothing component of the algorithm to be carried out as a back substitution in the manner discussed for the Paige and Saunders information filter in Chapter 2. This approach is basically the same as that advocated in [113].

The alternative approach using initialisation of the Kalman filter by means of a constant vector characterizes the algorithm of [111]. Here the data is represented in a form which makes clear the role of the kernel of the differential operator in the modelling process. This is:

$$y_i = \mathbf{h}^T X(t_i, t_1)\,\mathbf{d} + \mathbf{h}^T \mathbf{z}_i + \varepsilon_i,$$

where $\mathbf{d}$ is the parameter vector, the stochastic term $\mathbf{z}_i$ satisfies the dynamics equation (6.2.4), and a consistent choice is $\mathbf{z}_i = \sigma\sqrt{\lambda}\int_{t_1}^{t_i} X(t_i, s)\,\mathbf{b}\frac{dw}{ds}ds$

in which case $\mathbf{z}_1 = 0$, $\mathcal{V}\{\mathbf{z}_1\} = 0$, and $\mathbf{d}$ is the estimate of $\mathbf{x}(t_1)$. There are two novel features in this approach. The first is in the interpretation of the innovation sequence $\widetilde{y}_i$, $i = 1, 2, \cdots, n$. These are constructed by the Kalman filter to be independent (orthogonal) using the projection theorem. As $\mathcal{V}\{\widetilde{y}_i\}$ is computed at the same time (1.4.8) the innovations can be scaled to have unit variance. Now recall Remark 1.3.3 concerning the determination of the parameter vector in the mixed model problem . The generalised least squares problem considered there is identical with the least squares problem with data $L^{-1}[A, \mathbf{b}]$ where $LL^T = V + BR_{22}B^T$, and $L^{-1}[A\mathbf{d} - \mathbf{b}]$ is the vector of iid residuals which are identified with the scaled innovations $\widetilde{y}_{i+1}/\left(\mathbf{h}^T S_{i+1|i}\mathbf{h} + \sigma^2\right)^{\frac{1}{2}}$ in the mixed model problem solved by the Kalman filter in the current context. It follows that, for given observation vector $\mathbf{v}$, the Kalman filter, as a linear operation on the data, produces the result $L^{-1}\mathbf{v}$, where $\mathbf{v}$ is in turn the observation vector with components y_i, $i = 1, 2, \cdots, n$, and the columns of the design matrix $\mathbf{h}^T X(t_i, t_1)\mathbf{e}_j$, $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, m$. The estimate $\widehat{\mathbf{d}}$ of $\mathbf{d}$ for given λ can now be obtained by solving the least squares problem in which σ^2 cancels out conveniently. The second feature connects these calculations with a form of log likelihood. Using the independence of the innovations this can be written

$$\mathcal{L}_W\left(\mathbf{d}, \sigma^2, \lambda\right) = -\frac{1}{2}\left\{\begin{array}{c} \frac{\left(y_1 - \mathbf{h}^T\mathbf{d}\right)^2}{\sigma^2} + \log\sigma^2 \\ +\sum_{i=1}^{n-1}\left\{\begin{array}{c} \frac{\widetilde{y}_{i+1}^2}{\mathbf{h}^T S_{i+1|i}\mathbf{h} + \sigma^2} \\ +\log\left(\mathbf{h}^T S_{i+1|i}\mathbf{h} + \sigma^2\right) \end{array}\right\} \end{array}\right\} + \text{const.} \tag{6.2.20}$$

The method used to compute $\widehat{\mathbf{d}}$ follows directly from this by maximizing $\mathcal{L}_W$. This leads to the same least squares problem as the above argument. To see this note that the quadratic form in the $\widetilde{y}_i$ must have the form $\mathbf{v}^T L^{-T} L^{-1}\mathbf{v}$ where $v_i = y_i - \mathbf{h}^T X(t_i, t_1)\mathbf{d}$ as these are the inhomogeneous terms driving the filter, and the lower triangular form of L follows immediately from the initial value form of the filter. Because the filter is linear it suffices to evaluate the contribution from $\mathbf{y}$ by setting $\mathbf{d} = 0$. The corresponding design matrix is then obtained by differentiating the filter with respect to each component of $\mathbf{d}_i$. The dependence on all the data is obtained by using the smoothing algorithm to predict the $\mathbf{z}_{i|n}$.

A V-invariant approach

V-invariant methods offer a possible way to avoid the inversion of illconditioned lower triangular matrices required by the information filter formu-

lation while also avoiding any potential for instability in the standard recurrence forms of the filter algorithms. The factorization proceeds by steps which transform successive diagonal blocks. If the basic transformation is applied to (6.2.19) then there is accumulating fill. This is illustrated for the first few steps in (6.2.21) below:

$$\begin{bmatrix} -X_1 & I & \\ & -X_2 & I \\ \vdots & \vdots & \vdots \\ \mathbf{h}_1^T & & \\ & \mathbf{h}_2^T & \\ & & \mathbf{h}_3^T \end{bmatrix} \rightarrow \begin{bmatrix} U_1 & W_1 & \\ & -X_2 & I \\ \vdots & \vdots & \vdots \\ & \mathbf{z}_{11}^T & \\ & \mathbf{h}_2^T & \\ & & \mathbf{h}_3^T \end{bmatrix} \rightarrow \begin{bmatrix} U_1 & W_1 & \\ & U_2 & W_2 \\ \vdots & \vdots & \vdots \\ & & \mathbf{z}_{21}^T \\ & & \mathbf{z}_{22}^T \\ & & \mathbf{h}_3^T \end{bmatrix} \tag{6.2.21}$$

It is convenient to control this fill after m block factorization steps by an appropriate use of orthogonal transformations which have the effect in practice of reducing the fill to a total of $m+1$ rows [84]. For example the step corresponding to $i = m$ can be organised as follows:

$$\begin{bmatrix} \cdots & \cdots & \cdots \\ U_m & W_m & \\ & -X_{m+1} & I \\ \cdots & \cdots & \cdots \\ & \mathbf{z}_{m1}^T & \\ & \mathbf{z}_{m2}^T & \\ & \vdots & \\ & \mathbf{z}_{mm}^T & \\ & \mathbf{h}_{m+1}^T & \end{bmatrix} \rightarrow \begin{bmatrix} \cdots & \cdots & \cdots \\ U_m & W_m & \\ & -X_{m+1} & I \\ \cdots & \cdots & \cdots \\ & Z_m & \\ & 0 & \end{bmatrix}$$

where $Z_m : R^m \rightarrow R^m$ is upper triangular, and V-invariance is preserved by orthogonal transformations applied to the observation equations which have covariance $V = \sigma^2 I$. This orthogonal transformation step can be broken up into two stages $i \geq m$

$$\begin{bmatrix} \mathbf{z}_{i1}^T \\ \vdots \\ \mathbf{z}_{im}^T \\ \mathbf{h}_{i+1}^T \end{bmatrix} \rightarrow \begin{bmatrix} Z_i^1 \\ \mathbf{h}_{i+1}^T \end{bmatrix} \rightarrow \begin{bmatrix} Z_i \\ 0 \end{bmatrix},$$

where $Z_i^1 : R^m \rightarrow R^m$ is upper triangular. This structuring has the advantage that the predicted values $\mathbf{x}_{i+1|i}$, $S_{i+1|i} = \sigma^2 \left(Z_i^1\right)^{-1} \left(Z_i^1\right)^{-T}$, together with $\mathbf{x}_{i+1|i+1}$, $S_{i+1|i+1} = \sigma^2 Z_i^{-1} Z_i^{-T}$, $i = m, m+1, \cdots, n-1$ are readily available.

To set up the V-invariant transformations for the i'th block of the design a rank revealing Cholesky factorization of R_i is made:

$$P_i R_i P_i^T = \lambda L_i D_i L_i^T.$$

Associated with this is the transformation of the current block:

$$\left[\begin{array}{cc} -X_i & I \end{array}\right] \to \left[\begin{array}{cc} -L_i^{-1} P_i X_i & L_i^{-1} P_i \end{array}\right] = \overline{D}_i.$$

This is followed by the reordering of D_i to ensure the elements are nondecreasing using the permutation matrix P_o:

$$D_i \to P_o D_i P_o^T = \overline{D}_i.$$

This permutation transformation is also applied to the current block. The elements of the weight matrix that enter into the V-invariant transformation at this step are $\left[\begin{array}{cc} \overline{D}_i & \\ & I_k \end{array}\right]$ where $k = \min(i, m+1)$, and typically these are correctly ordered for stable transformations. This follows from Remark 6.2.3 which shows that the largest eigenvalue of $R_i = O(\lambda\delta)$ and no permutation is required unless this is (significantly) greater than 1. Thus no permutation is suggested unless $\lambda \geq O(\delta^{-1})$. However, if the model is accurate then it is expected that an optimal λ will also be small.

The Reinsch algorithm

An alternative approach leads to a generalisation of the classic Reinsch algorithm [94] for computing smoothing splines. The starting point is a generalisation of divided differences that is used to remove the dependence of the systematic component in $\mathbf{x}(t)$ defined by (6.2.4) from the observation equations. To do this consider the system of equations

$$\sum_{j=1}^{m+1} \Delta_{ij}^m \mathbf{h}^T X(t_{i+j-1}, t_i) = 0, \ \sum_{i=1}^{m} \left|\Delta_{ij}^m\right| = 1.$$

This system defines an operation Δ_i^m on the data. It is well defined essentially under the same conditions that are required to ensure that the estimation problem has a well determined solution.

Example 6.2.5 *Let* $\mathbf{m} = 0$, $\mathbf{h} = \mathbf{e}_1$, $m = 2$. *Then*

$$X(t, t_i) = \left[\begin{array}{cc} 1 & t - t_i \\ & 1 \end{array}\right],$$

and the equations determining the Δ_{ij}^m *are*

$$\begin{aligned}\Delta_{i1}^m + \Delta_{i2}^m + \Delta_{i3}^m &= 0,\\ (t_{i+1} - t_i)\,\Delta_{i2}^m + (t_{i+2} - t_i)\,\Delta_{i3}^m &= 0,\\ \Delta_{i3}^m &= \gamma,\end{aligned}$$

where γ *is to be chosen to satisfy the scaling condition. The result is*

$$\Delta_{i1}^m = \frac{(t_{i+2} - t_{i+1})}{2\,(t_{i+2} - t_i)}, \Delta_{i2}^m = -\frac{1}{2}, \Delta_{13}^m = \frac{(t_{i+1} - t_i)}{2\,(t_{i+2} - t_i)}\ .$$

It will be recognised as a scaled form of the second divided difference.

This generalised divided difference operation gives

$$z_i = \Delta_i^m y_i = \Delta_i^m \mathbf{h}^T \mathbf{u}_i + \Delta_i^m \varepsilon_i,\ i = 1, 2, \cdots, n - m.$$

It is now required to determine the best prediction of ε given the data. This is computed using (1.3.16). Subtracting this from the observed data gives the corresponding estimate for η:

$$\eta = \mathbf{y} - \mathcal{C}\left\{\varepsilon, \mathbf{z}\right\} \mathcal{V}\left\{\mathbf{z}\right\}^{-1} \mathbf{z}.$$

This is just the vector of values of the spline at the knot points. To compute these quantities let

$$J = \begin{bmatrix} \Delta_{11}^m & \cdots & \Delta_{1(m+1)}^m & & & \\ & \Delta_{21}^m & \cdots & & & \\ & & \cdots & \cdots & & \\ & & & \Delta_{(n-m)1}^m & \cdots & \Delta_{(n-m)(m+1)}^m \end{bmatrix}.$$

Then

$$\begin{aligned}\mathcal{C}\left\{\varepsilon, \mathbf{z}\right\} &= \sigma^2 J^T,\\ \mathcal{V}\left\{\mathbf{z}\right\} &= \sigma^2 J J^T + \sigma^2 \lambda \mathcal{V}\left\{\mathbf{q}\right\},\end{aligned}$$

where

$$q_i = \sum_{j=2}^{m+1} \Delta_{ij}^m \mathbf{h}^T \int_{t_i}^{t_{i+j-1}} X\left(t_{i+j-1}, s\right) \mathbf{b} dw,\ i = 1, 2, \cdots, n - m. \qquad (6.2.22)$$

There can be a contribution to $\mathcal{E}\left\{\mathbf{q}\mathbf{q}^T\right\}$ only when the intervals of integration in (6.2.22) overlap because otherwise the random walk contributions are independent. It follows that $\mathcal{V}\left\{\mathbf{q}\right\}$ is a $2m - 1$ banded matrix.

Use of the divided difference like operator Δ_i^m raises questions about the conditioning of this computation as a result of the cancellation which occurs consequent on the use of the high order differences. Presumably this is the behaviour which corresponds with the small eigenvalues of the covariance matrices in the Kalman filter approach.

Bibliography

[1] H. D. I. Abarbanel, D. R. Creveling, R. Farsian, and M. Kostuk, *Dynamical State and Parameter Estimation*, SIAM J. Applied Dynamical Systems **8** (2009), 1341–1381.

[2] B. D. O. Anderson and J. B. Moore, *Optimal filtering*, Prentice-Hall, 1979.

[3] C. F. Ansley and R. Kohn, *A geometrical derivation of the fixed interval smoothing algorithm*, Biometrica **69** (1982), 486–487.

[4] ———, *Estimation, filtering, and smoothing in state space models with incompletely specified initial conditions*, The Annals of Statistics **13** (1985), 1286–1316.

[5] U. Ascher and M. R. Osborne, *A note on solving nonlinear equations and the " natural criterion function"*, J. Opt. Theory and Applic. **55** (1988), 147–152.

[6] U. M. Ascher, R. M. M. Mattheij, and R. D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, SIAM, Philadelphia, 1988.

[7] G. J. Bierman, *Factorization methods for discrete sequential estimation*, Academic Press, New York, 1977.

[8] P. Billingsley, *Probability and measure*, John Wiley & Sons, New York, 1979.

[9] Å. Björck, *Solving linear least squares problems by Gram-Schmidt orthogonalization*, BIT **7** (1967), 1–22.

[10] Å. Björck and C. C. Paige, *Loss and recapture of orthogonality in the modified Gram-Schmidt algorithm*, SIAM J. Matrix Anal. Appl. **13** (1992), 176–190.

[11] Åke Björck, *Numerical methods for least squares problems*, SIAM, 1996.

[12] H. G. Bock, *Recent advances in parameter identification techniques in ODE*, Numerical Treatment of Inverse Problems in Differential and Integral Equations (P. Deuflhard and E. Hairer, eds.), Birkhäuser, 1983, pp. 95–121.

[13] H. G. Bock, E. Eich, and J. P. Schlöder, *Numerical solution of constrained least squares boundary value problems in differential algebraic equations*, Numerical Teatment of Differential Equations (K. Strehmel, ed.), Teubner, 1988, pp. 269–280.

[14] H. G. Bock, Ekaterina Kostina, and J. P. Schlöder, *On the role of natural level functions to achieve global convergence for damped Newton methods*, System Modelling and Optimization Methods, Theory and Applications (M. J. P. Powell and S. Sholtes, eds.), Kluwer Academic Publishers, 2000.

[15] _____, *Numerical methods for parameter estimation in nonlinear differential algebraic equations*, GAMM Mitteilungen, 2007, in press.

[16] H.G. Bock, *Numerical treatment of inverse problems in chemical reaction kinetics*, Modelling of Chemical Reaction Systems (K.H. Ebert, P. Deuflhard, and W. Jäger, eds.), Springer Series in Chemical Physics, vol. 18, Springer, Heidelberg, 1981, pp. 102–125.

[17] S. Broadfoot, *Selection of boundary matrices in the embedding method*, 2011, Honours thesis in Mathematics, Australian National University.

[18] J. C. Butcher, *Numerical methods for ordinary differential equations*, John Wiley & Sons, 2003.

[19] C. K. Chui and G. Chen, *Kalman filtering*, Springer-Verlag, 1987, with Real-Time Applications.

[20] R. M. Corless and N. Fillion, *Graduate Introduction to Numerical Methods*, 2012, Vol 1, in preparation.

[21] G. Dahlquist, *Convergence and stability in the numerical integration of ordinary differential equations*, Math. Scand. **4** (1956), 33–53.

[22] F. R. de Hoog and R. M. M. Mattheij, *On Dichotomy and Well-conditioning in BVP*, SIAM J. Numer. Anal. **24** (1987), 89–105.

[23] P. Deuflhard, *A modified Newton method for the solution of ill conditioned systems of nonlinear equations with application to multiple shooting*, Num. Math. **22** (1974), 289–315.

[24] ———, *Newton methods for nonlinear problems*, Springer-Verlag, Berlin Heidelberg, 2004.

[25] P. Deuflhard and G. Heindl, *Affine invariant convergence theorems for Newton's method and extensions to related methods*, SIAM J. Numer. Anal. **16** (1979), 1–10.

[26] D. B. Duncan and S. D. Horn, *Linear dynamic recursive estimation from the viewpoint of regression analysis*, J. Amer. Statist. Assoc. **67** (1972), 816–821.

[27] R. England and R. M. M. Mattheij, *Boundary value problems and dichotomic stability*, SIAM J. Numer. Anal. **25** (1988), 1037–1054.

[28] J. Fan and I. Gijbels, *Local polynomial modelling and its applications*, Chapman and Hall, 1996.

[29] A. V. Fiacco and G. P. McCormick, *Nonlinear programming: Sequential unconstrained minimization techniques*, John Wiley & Sons, 1968.

[30] R. Fletcher, *Practical methods of optimization: Unconstrained optimization*, vol. 1, John Wiley & Sons, Chichester, 1980.

[31] ———, *Practical methods of optimization: Constrained optimization*, vol. 2, John Wiley & Sons, Chichester, 1981.

[32] L. Fox, *The numerical solution of two-point boundary value problems in ordinary differential equations*, Oxford University Press, Oxford, 1957.

[33] P. E. Gill, G. H. Golub, W. Murray, and M. A. Saunders, *Methods for modifying matrix factorizations*, Math. Comp. **28** (1974), 505–535.

[34] V. P. Godambe and C. C. Heyde, *Quasi-likelihood and optimal estimation*, Int. Statist. Rev. **55** (1987), 231–244.

[35] G. H. Golub, *Numerical methods for solving least squares problems*, Num. Math. **7** (1965), 206–216.

[36] G. H. Golub and S. G. Nash, *Nonorthogonal analysis of variance using a generalised conjugate-gradient algorithm*, J. Amer. Stat. Assoc. **77** (1982), no. 377, 109–116.

[37] G. H. Golub and V. Pereyra, *The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate*, SIAM J. Num. Anal. **10** (1973), 413–432.

[38] ———, *Separable nonlinear least squares: the variable projection method and its applications*, Inverse Problems **19** (2003), R1–R26.

[39] G. H. Golub and C. F. Van Loan, *Matrix computations*, Johns Hopkins University Press, Baltimore, 1996, third edition.

[40] G. H. Golub and J. H. Wilkinson, *Iterative refinement of least squares solutions*, Num. Math. **9** (1966), 189–198.

[41] M. Gulliksson and P. Wedin, *Modifying the QR-decomposition to constrained and weighted linear least squares*, SIAM J. Matrix Anal. Appl. **13** (1992), no. 4, 1298–1313.

[42] E. J. Hannan and M. Deistler, *The statistical theory of linear systems*, John Wiley & Sons, New York, 1988.

[43] P. C. Hansen and P. Y. Yalamov, *Computing symmetric rank-revealing decompositions via triangular factorization*, SIMAX **23** (2001), no. 2, 443–458.

[44] M. Hegland and M. R. Osborne, *Wrap-around partitioning for block bidiagonal linear systems*, IMA J. Numer. Anal. **18** (1998), no. 3, 373–383.

[45] C. C. Heyde, *Fixed sample and asymptotic optimality for classes of estimating functions*, Contemp. Math. **89** (1988), 241–247.

[46] N. J. Higham, *Accuracy and stability of numerical algorithms*, SIAM, 2002, Second Edition.

[47] R. R. Hocking, *The analysis of linear models*, Brooks/Cole, 1984.

[48] A. S. Householder, *Unitary triangularization of a nonsymmetric matrix*, J. ACM **6** (1958), 339–342.

[49] ———, *The theory of matrices in numerical analysis*, Blaisdell Pub. Co., New York, 1964.

[50] M. Jamshidian and R. I. Jennrich, *Nonorthogonal analysis of variance using gradient methods*, J. Amer. Stat. Assoc. **83** (1988), no. 402, 483–489.

[51] L. S. Jennings and M. R. Osborne, *A direct error analysis for least squares*, Num. Math. **22** (1974), 325–332.

[52] R. I. Jennrich, *Asymptotic properties of nonlinear least squares estimators*, Ann. Math. Statist. **40** (1969), 633–643.

[53] Krisorn Jittorntrum, *Accelerated convergence for the Powell/Hestenes multiplier method*, Math. Prog. **18** (1980), 197–214.

[54] M. H. Kahn, M. S. Mackisack, M. R. Osborne, and G. K. Smyth, *On the consistency of Prony's method and related algorithms*, J. Computational and Graphical Statistics **1** (1992), 329–350.

[55] T. Kailath, *Lectures on Weiner and Kalman filtering*, Springer-Verlag, Wien-New York, 1981.

[56] L. Kaufman, *Variable projection method for solving separable nonlinear least squares problems*, BIT **15** (1975), 49–57.

[57] M. G. Kendall and A. Stuart, *The advanced theory of statistics*, vol. 2: Inference and Relationship, Charles Griffin and Company Limited, London, 1967.

[58] Peter E. Kloeden and Eckhard Platen, *Numerical solution of stochastic differential equations*, Springer-Verlag, Berlin Heidelberg, 1992.

[59] R. Kohn and C. Ansley, *A new algorithm for spline smoothing based on smoothing a stochastic process*, J. Sci. Statist. Comput. **8** (1987), 33–48.

[60] R. Kohn and C. F. Ansley, *On the smoothness properties of the best linear unbiased estimate of a stochastic process observed with noise*, Ann. Statist. **11** (1983), 1011–1017.

[61] M. Lalee, J. Nocedal, and T. Plantenga, *On the implementation of an algorithm for large-scale equality constrained optimization*, SIAM J. Optim. **8** (1998), no. 3, 682–706.

[62] K. Levenberg, *A method for the solution of certain nonlinear problems in least squares*, Quart. Appl. Math. **2** (1944), 164–168.

[63] Z. Li, M. R. Osborne, and T. Prvan, *Parameter estimation of ordinary differential equations*, IMA J. Numer. Anal. **25** (2005), 264–285.

[64] Z. F. Li, M. R. Osborne, and Tania Prvan, *Numerical algorithms for constrained maximum likelihood estimation*, ANZIAM J. **45** (2003), 91–114.

[65] J. W. Longley, *Least squares computations using orthogonalization methods*, Marcel Decker, 1984.

[66] D. G. Luenberger, *Optimization by vector space methods*, John Wiley & Sons, New York, 1969.

[67] D. W. Marquardt, *An algorithm for least squares estimation of nonlinear parameters*, J. Soc. Indust. Appl. Math. **11** (1963), 431–441.

[68] B. McCabe and A. Tremayne, *Elements of modern asymptotic theory with statistical applications*, Manchester University Press, 1993.

[69] P. McCullagh and J. A. Nelder, *Generalised linear models*, 2nd edition ed., Chapman and Hall, 1989.

[70] A. J. Miller, *Subset selection in regression*, Chapman and Hall, 1990.

[71] J. J. Moré, *The Levenberg-Marquardt algorithm: Implementation and theory*, Numerical Analysis. Proceedings, Dundee 1977 (G. A. Watson, ed.), Springer-Verlag, 1978, Lecture Notes in Mathematics No. 630, pp. 105–116.

[72] D. D. Morrison, *Methods for nonlinear least squares problems and convergence proofs*, JPL Seminar Proceedings, Space Technology Laboratory Inc., 1960.

[73] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer Verlag, 1999.

[74] M. R. Osborne, *A method for finite-difference approximation to ordinary differential equations*, The Computer Journal **7** (1964), 58–64.

[75] ———, *On shooting methods for boundary value problems*, J. Math. Analysis and Applic. **27** (1969), 417–433.

[76] ———, *Nonlinear least squares - the Levenberg algorithm revisited*, J. Aust. Math. Soc., Series B **19** (1977), 343–357.

[77] ———, *Fisher's Method of Scoring*, Int. Stat. Rev. **86** (1992), 271–286.

[78] ———, *Solving least squares problems on parallel vector processors*, Numerical Analysis (D. F. Griffiths and G. A. Watson, eds.), World Scientific, 1996, A. R. Mitchell 75'th Birthday Volume, pp. 208–224.

[79] ———, *Simplicial algorithms for minimizing polyhedral functions*, Cambridge University Press, 2001.

[80] ———, *Separable least squares, variable projection, and the Gauss-Newton algorithm*, ETNA **28** (2007), 1–15.

[81] M. R. Osborne and Tania Prvan, *On algorithms for generalised smoothing splines*, J. Austral. Math. Soc. Ser. B **29** (1988), 322–341.

[82] ———, *Smoothness and conditioning in generalised smoothing spline calculations*, J. Austral. Math. Soc. Ser. B **30** (1988), 43–56.

[83] ———, *What is the covariance analogue of the Paige and Saunders information filter*, SIAM J. Sci. Stat. Computing **12** (1991), 1324–1331.

[84] M. R. Osborne and I. Söderkvist, *V-invariant methods, generalised least squares problems, and the Kalman filter*, ANZIAM J. **45(E)** (2004), C232–C247.

[85] C. C. Paige, *Computer solution and perturbation analysis of generalised linear least squares problems*, Math. Comp. **33** (1979), no. 145, 171–183.

[86] ———, *Fast numerically stable computations for generalised linear least squares problems*, SIAM J. Numer. Anal **16** (1979), no. 1, 165–171.

[87] C. C. Paige and M. A. Saunders, *Least squares estimation of discrete linear dynamic systems using orthogonal transformations*, SIAM J. Numer. Anal. **14** (1977), 180–193.

[88] H. D. Patterson and R. Thompson, *Recovery of interblock information when block sizes are unequal*, Biometrika **58** (1971), 545–554.

[89] H. Pohjanpalo, *System identifiability based on the power series expansion of the solution*, Math. Biosciences **41** (1978), 21–33.

[90] M. J. D. Powell and J. K. Reid, *On applying Householder transformations to linear least squares problems*, IFIP Congress, North Holland Publishing Company, 1969, pp. 122–126.

[91] R. Prony, *Essai expérimental et analytique: Sur les lois de la dilatabilité de fluides elastique et sur celles de la force expansive de la vapeur de l'alkool à différentes températures*, Journal de l'École Polytechnique **1** (1795), 24–76.

[92] B. J. Quinn and E. J. Hannan, *The estimation and tracking of frequency*, Cambridge University Press, Cambridge, United Kingdom, 2001, `ftp://uiarchive.cso.uiuc.edu/pub/etext/gutenberg/;http://www.loc.gov/catdir/description/cam021/00051944.html;http://www.loc.gov/catdir/toc/cam027/00051944.html`.

[93] J. O. Ramsay, G. Hooker, C. Campbell, and C. Cao, *Parameter estimation for differential equations: a generalised smoothing approach*, J. Royal Statistical Society: Series B (Statistical Methodology) **69** (2007), no. 5, 741–796.

[94] C. H. Reinsch, *Smoothing by spline functions*, Numer. Math. **10** (1967), 177–183.

[95] B. D. Ripley, *Computer generation of random variables: a tutorial*, Int. Stat. Rev. **51** (1983), 301–319.

[96] A. Ruhe and P.Å. Wedin, *Algorithms for separable nonlinear least squares problems*, SIAM Review **22** (1980), 318–337.

[97] A. Sedoglovic, *A probabalistic algorithm to test local algebraic observability in polynomial time*, J. Symbolic Comp. **33** (2002), 735–755.

[98] K. S. Sen and J. M. Singer, *Large sample methods in statistics*, Chapman and Hall, 1993.

[99] I. Söderkvist, *An algorithm for Kalman filtering and smoothing*, Computational Techniques and Applications: CTAC95 (Robert L. May and Alan K. Easton, eds.), World Scientific Publishing Co., 1996, pp. 709–716.

[100] I. Söderkvist, *On algorithms for generalized least squares problems with ill-conditioned covariance matrices*, Computational Statistics **11** (1996), 303–313.

[101] Gustav Soderlind, *The Logarithmic Norm. History and Modern Theory*, (2006).

[102] G. W. Stewart, *The effect of rounding error on an algorithm for downdating a Choleski factorization*, JIMA **23** (1979), 203–213.

[103] R. A. Thisted, *Elements of statistical computing*, Chapman and Hall, 1988.

[104] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society, Series B **58** (1996), no. 1, 267–288.

[105] I. Tjoa and L. T. Biegler, *Simultaneous solution and optimization strategies for parameter estimation of differential-algebraic systems*, Ind. Eng. Chem. Res. **30** (1991), 376–385.

[106] H. W. Turnbull and A. C. Aitken, *An introduction to the theory of canonical matrices*, Blackie, London and Glasgow, 1932.

[107] A. Wächter and L.T. Biegler, *On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming*, Mathematical Programming **106** (2006), 25–57.

[108] Grace Wahba, *Improper priors, spline smoothing, and the problem of guarding against model errors in regression*, J. Roy. Statist. Soc. B **40** (1978), 364–372.

[109] ———, *Spline models for observational data*, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1990.

[110] A. Wald, *Note on the consistency of the maximum likelihood estimate*, Ann. Math. Statist. **20** (1949), 595–601.

[111] W. Wecker and C. F. Ansley, *The signal extraction approach to nonlinear regression and spline smoothing*, J. Amer. Statist. Assoc. **78** (1983), 81–89.

[112] R. W. M. Wedderburn, *Quasi-likelihood functions, generalised linear models, and the Gauss-Newton method*, Biometrika **61** (1974), 439–447.

[113] H. L. Weinert, R. H. Byrd, and G. S. Sidhu, *A stochastic framework for recursive computation of spline functions: Part II, smoothing splines*, J. Optim. Theory Appl. **30** (1980), 255–268.

[114] H. White, *Maximum likelihood estimation of misspecified models*, Econometrika **50** (1982), 1–25.

[115] J. H. Wilkinson, *The algebraic eigenvalue problem*, O.U.P., Oxford, 1965.

[116] ———, *Error analysis of transformations based on the use of matrices of the form* $I - 2uu^H$, Error in Digital Computation (New York) (Louis B. Rall, ed.), John Wiley & Sons, 1965, volume 2, pp. 77–101.

[117] S. J. Wright, *Stable parallel algorithms for two point boundary value problems*, SISSC **13** (1992), 742–764.

[118] ———, *A collection of problems for which gaussian elimination with partial pivoting is unstable*, SISSC (1993), 231–238.

[119] S. L. Zeger, K-Y. Liang, and P. S. Albert, *Models for longitudinal data: A generalised estimating equation approach*, Biometrics **44** (1988), 1049–1060.

Index