An approach to parameter estimation and model selection in differential equations

M.R.Osborne

Mathematical Sciences Institute, Australian National University, ACT 0200, Australia

Summary. Two distinct and essentially independent sources of error occur in the parameter estimation problem – the error due to noisy observations, and the error due to approximation or discretization effects in the computational procedure. These give contributions of different orders of magnitude so the problem is essentially a two grid problem and there is scope for balancing these to minimize computational effort. Here the underlying computing procedures that determine these errors are reviewed. There is considerable structure in the integration of the differential system, and the role of cyclic reduction in unlocking this is discribed. The role of the stochastic effects in the optimization component of the computation is critically important in understanding the success of the Gauss-Newton algorithm, and the importance both of adequate data and of a true model is stressed. If the true model must be sought among a range of competitors then a stochastic embedding technique is suggested that converts under-specified models into "non-physical" consistent models for which the Gauss-Newton algorithm can be used.

1 Introduction

The estimation problem for parameterized systems of differential equations starts with data acquired through observations of system trajectories made in the presence of noise, and it seeks to estimate the parameter values by solving an optimization problem which matches computed solutions of the differential equation to this observed data. Note that the explicit assumption that the data is noisy means that there is an explicit stochastic component to the problem. This has the immediate consequence that two distinct grids are relevant in this problem formulation. These are:

- the grid defined by the observation process; and
- the grid defined by the discretization of the differential system.

The need to distinguish between these is a consequence of the different rates of convergence of the estimates implied by the two different sources of error – the

stochastic component deriving from the measurement process with its characteristic $O\left(n^{-1/2}\right)$ rate, and the rate deriving from the truncation error in the numerical procedure where an $O\left(n^{-2}\right)$ estimate is readily achieved. It follows that there could well be scope for economizing on the work done in integrating the differential system because of the distinctly lower accuracy achievable in the estimates based on the observation grid. Our aim here is to examine the key computational components of the estimation problem (the discretization of the differential equations and the parameter estimation optimization problem) to expose structure that might be used in the economization process.

To define the problem let the differential equation be

$$\frac{d\mathbf{x}}{dt} = \mathbf{w}(t, \mathbf{x}, \boldsymbol{\beta}) \tag{1}$$

where $\mathbf{x}, \mathbf{w} \in R^m,\ \boldsymbol{\beta} \in R^p$. An important case is the equation linear in the state variable $\mathbf{x}$

$$\mathbf{w} = M(t, \boldsymbol{\beta})\mathbf{x} + \mathbf{f}(t). \tag{2}$$

This is not only significant in its own right, but through a process of successive linearization it provides the "enabling technology" needed for the fully nonlinear problem. For this reason most of the discussion here is in terms of the linear equation system. The second component of the problem formulation is the observed data:

$$y_i = \boldsymbol{\phi}^T \mathbf{x}(t_i) + \varepsilon_i,\ i = 1, 2, \cdots, n, \tag{3}$$

where $\boldsymbol{\phi}$ defines the "observation functional" and the observational error is given by ε_i. These errors are assumed to be independent and identically distributed and typically normal.

Typically, two classes of method are considered:

embedding method : Here explicitly computed solution trajectories are compared directly with the observations in an unconstrained optimization procedure. Typically this has two main components:

1. Given trial $\boldsymbol{\beta}$, plus *auxiliary information* $\mathbf{b}$, generate trial solution $\mathbf{x}(t, \boldsymbol{\beta}, \mathbf{b})$.
2. Using trial solution make adjustments to $\boldsymbol{\beta}$ and auxiliary information $\mathbf{b}$ to improve estimate of $\boldsymbol{\beta}$ and $\mathbf{x}(t, \boldsymbol{\beta}, \mathbf{b})$. Measure goodness of fit by

$$F(\boldsymbol{\beta}, \mathbf{b}) = \sum_{i=1}^{n} r_i(t_i, \boldsymbol{\beta}, \mathbf{b})^2, \tag{4}$$

$$r_i = y_i - \boldsymbol{\phi}^T \mathbf{x}(t_i, \boldsymbol{\beta}, \mathbf{b}). \tag{5}$$

To set up the auxiliary information first select boundary matrices B_1, B_2 and guess the variable part of the auxiliary information $\mathbf{b}$. This data is required to permit the solution of the boundary (or initial) value problem

$$B_1\mathbf{x}(0) + B_2\mathbf{x}(1) = \mathbf{b},$$
$$\frac{d\mathbf{x}}{dt} = M(t,\beta)\mathbf{x} + \mathbf{f}(t).$$

It is important to choose B_1, B_2 so that the Green's matrix is nicely bounded. This is always possible if the system possesses a known, well-defined dichotomy [1]. However, this requires significant additional structural information about the differential system. Typically it is required of the imposed conditions that the fast solutions of the differential equation be pinned down at $t = 1$, and slow solutions be similarly pinned down at $t = 0$. Simple shooting corresponds to $B_1 = I,\ B_2 = 0$. It requires the initial value problem to be stable.
This formulation leads to the nonlinear least squares problem:

$$\min_{\mathbf{b},\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{\phi}^T \mathbf{x}(t_i, \boldsymbol{\beta}, \mathbf{b}))^2. \tag{6}$$

It is recommended that the correction to $\boldsymbol{\beta},\ \mathbf{b}$ be computed using the Gauss-Newton or Scoring method [2]. This requires the integration of the variational equations:

$$\begin{cases} \frac{d\boldsymbol{\Delta}_\beta}{dt} = M\boldsymbol{\Delta}_\beta + \nabla_\beta M\mathbf{x}, \\ B_1\boldsymbol{\Delta}_\beta(0) + B_2\boldsymbol{\Delta}_\beta(1) = 0, \end{cases}$$
$$\begin{cases} \frac{d\boldsymbol{\Delta}_b}{dt} = M\boldsymbol{\Delta}_b, \\ B_1\boldsymbol{\Delta}_b(0) + B_2\boldsymbol{\Delta}_b(1) = I, \end{cases}$$

where

$$\boldsymbol{\Delta}_\beta = \frac{\partial \mathbf{x}}{\partial \boldsymbol{\beta}},\ \boldsymbol{\Delta}_b = \frac{\partial \mathbf{x}}{\partial \mathbf{b}}.$$

simultaneous method : In this approach the system of differential equations is imposed as explicit constraints on the optimization problem (6) [3]. The resulting equality constrained mathematical program typically is solved by a variant of sequential quadratic programming. For the linear differential equation the mathematical program can be formulated as

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{\phi}^T \mathbf{x}(t_i, \boldsymbol{\beta}))^2, \tag{7}$$

subject to the equality constraints

$$\mathbf{x}_{i+1} - X_i(t_{i+1}, t_i)\mathbf{x}_i = \mathbf{v}_i,\ i = 1, 2, \cdots, n-1, \tag{8}$$

where $X_i(t, t_i)$ is the fundamental matrix, and $\mathbf{v}_i(t)$ is the particular integral for equation (2) given by

$$\frac{dX_i}{dt} = MX_i,\ X_i(t_i, t_i) = I, \tag{9}$$

$$\mathbf{v}_i = \int_{t_i}^{t_{i+1}} X_i(t_{i+1}, u)\mathbf{f}(u)du. \tag{10}$$

In practice, the differential equation constraints would be replaced by an appropriate discretization. In this formulation the additional information comes from the Lagrange multipliers, but now the dimension of the constraint system grows with n. This can be avoided to some extent at least by the use of a coarser solution grid to generate data which can be interpolated linearly to obtain values to compare with the observed data. However, the differential system has only m degrees of freedom. This suggests a more compact specification could be available.

This brief sketch of the algorithmic possibilities suggests a number of significant problems including the determination of suitable boundary matrices $B_1,\ B_2$, finding problem formulations that might aid the selection of appropriate integration grids, and reducing the degrees of freedom information in the mathematical programming formulation. A possible approach to these problems is considered in the next section which is based around the possible forms of cyclic reduction in this context. It is shown that, in certain circumstances at least, an optimal reduction in the degrees of freedom in the constraint system is possible. The applicability of the Gauss-Newton algorithm has been indicated above. A general treatment is sketched in the following section (including for constrained problems). Part of the context required here is that the model for the system trajectory be correctly specified. To rescue the Gauss-Newton method when this is not the case introduces the interesting class of problems involving model selection. One possible approach, involving a stochastic embedding of the tentative model equations, is sketched in the final section.

2 Cyclic reduction

Cyclic reduction [4] is an elimination scheme applied to the block bidiagonal recurrence

$$A_i^0\mathbf{x}_{i+1} + B_i^0\mathbf{x}_i = \mathbf{c}_i^0 \tag{11}$$

which combines adjacent rows using techniques such as partial pivoting or orthogonal reduction as follows:

$$\begin{bmatrix} B_{i-1}^0 & A_{i-1}^0 & 0 & \mathbf{c}_{i-1}^0 \\ 0 & B_i^0 & A_i^0 & \mathbf{c}_i^0 \end{bmatrix} \rightarrow \begin{bmatrix} B_{i/2}^1 & 0 & A_{i/2}^1 & \mathbf{c}_{i/2}^1 \\ V_i^1 & -I & W_i^1 & \mathbf{w}_i^1 \end{bmatrix}. \tag{12}$$

The procedure can be applied recursively to give

Interpolation equations

$$\mathbf{x}_t = V_t\mathbf{x}(0) + W_t\mathbf{x}(1) + \mathbf{w}_t, \tag{13}$$

Constraint equation

$$G_1^k \mathbf{x}(0) + G_2^k \mathbf{x}(1) = \mathbf{c}_1^k. \tag{14}$$

The process is simplest if $n = 2^k$, but this restriction is not necessary in the bidiagonal case considered here [5]. The resulting equations (13), (14) are intrinsic properties of the differential equation system in the sense that they do not depend on the boundary conditions.

The constraint equation (14) gives immediate information concerning the choice of the boundary matrices in the embedding approach. It is required to choose $B_1,\ B_2$ to ensure that

$$\begin{bmatrix} B_1 & B_2 \\ G_1^k & G_2^k \end{bmatrix}$$

has a 'nicely' bounded inverse for then $\mathbf{x}(0),\ \mathbf{x}(1)$ can be found stably and the remaining values filled in using the interpolation equations. Thus G_1^k, G_2^k must reflect the dichotomy properties of the ODE system.

The interpolation equations (13) produced by the recursive cyclic reduction transformations allow the reformulation of the equality constrained estimation problem with minimum degrees of freedom (minimum numbers of equality constraints):

$$\min_{\beta} \sum_{t=t_i, i=1}^{n} \left(y_t - \boldsymbol{\phi}^T \left(V_t \mathbf{x}(0) + W_t \mathbf{x}(1) + \mathbf{w}_t \right) \right)^2 \tag{15}$$

subject to the constraints

$$G_1^k \mathbf{x}(0) + G_2^k \mathbf{x}(1) = \mathbf{c}_1^k. \tag{16}$$

This reduces the Lagrangian form of the problem to solving an optimization problem involving a fixed number m of equality constraints.

Certain properties are an immediate consequence of the basic process:

- Boundary conditions on the interpolation equations (13) follow directly:

$$\begin{array}{l} V(0) = I,\ V(1) = 0, \\ W(0) = 0,\ W(1) = I, \\ \mathbf{w}(0) = 0,\ \mathbf{w}(1) = 0. \end{array} \tag{17}$$

- $V_t, W_t, \mathbf{w}_t, G_1, G_2, c$ are not uniquely defined by the cyclic reduction process. Let C_t be the transformation that combines adjacent block rows. Then there is an equivalence class of transformations:

$$C_t \leftarrow \begin{bmatrix} R_1(t) & 0 \\ R_{21}(t) & R_2(t) \end{bmatrix} C_t \tag{18}$$

  that preserve the basic structure in the elimination tableau (12). Freedom in the interpolation is in $R_2^{-1} R_{21}$. Freedom in the constraint is in R_1.

- Relationships in the constraint equation can also be identified

$$\left(G_2^k\right)^{-1} G_1^k = X(1,0),\ \left(G_2^k\right)^{-1} \mathbf{c}_1^k = \mathbf{v}_1.$$

The simplest transformation introducing a zero in the cyclic reduction step is given by

$$C = \begin{bmatrix} I & X(t_{i+1},t_i) \\ I & -X(t_{i+1},t_i) \end{bmatrix}. \tag{19}$$

Assume $\delta = t_{i+1} - t_i$ is small. Substitute for the state variable using the interpolation equations, apply the transformation, and expand in powers of δ. Equating leading terms gives second order differential systems for the interpolation equation quantities (note this means the boundary condition (17) can be satisfied):

$$\begin{aligned} &\frac{d^2}{dt^2}\left(X^{-1}\left\{\begin{matrix} V \\ W \end{matrix}\right.\right) = 0, \\ &\Rightarrow V = X(t,0)(1-t),\ W = X(t,1)t. \end{aligned}$$

Other possibilities can be found by fixing

$$R_2^{-1} R_{21} = S_1 = \delta S + O(\delta^2). \tag{20}$$

Substituting this into $C \leftarrow RC$ and repeating the calculation gives for V (W is similar)

$$\frac{d^2V}{dt^2} + 2\left(S - M\right)\frac{dV}{dt} + \left(M^2 - 2SM - \frac{dM}{dt}\right)V = 0. \tag{21}$$

The interesting reduction is based on the use of orthogonal transformations as we know that bidiagonal systems with coupled boundary conditions provides a practical example of the potential instability of partial pivoting. Orthogonal transformation requires

$$C^T R^T R C = I$$

Substituting and expanding in powers of δ gives

$$S = \frac{M + M^T}{2} \tag{22}$$

Substituting in the general equation (21) gives (here order is important)

$$\left(\frac{d}{dt} + M^T\right)\left(\frac{d}{dt} - M\right)\left\{\begin{matrix} V \\ W \end{matrix}\right. = 0$$

The general equation (21) corresponds to the first order system (write Y for either V, W)

$$\frac{d}{dt}\begin{bmatrix} Y \\ Z \end{bmatrix} = N\begin{bmatrix} Y \\ Z \end{bmatrix},\ N = \begin{bmatrix} M & I \\ & -(2S - M) \end{bmatrix}. \tag{23}$$

To see where the constraint equation (14) fits in write the particular integral equation in form

$$\frac{d}{dt}\begin{bmatrix}\mathbf{w}\\ \mathbf{z}\end{bmatrix} = N\begin{bmatrix}\mathbf{w}\\ \mathbf{z}\end{bmatrix} + \begin{bmatrix}\mathbf{f}\\ 0\end{bmatrix}.$$

The interpolation equation (13) gives $\mathbf{x}$ as a combination of solutions of the higher order $(2m \times 2m)$ first order systems for V, W, $\mathbf{w}$. The function of the constraint equation is to remove unwanted components of the expanded fundamental matrix. It is required that

$$\begin{aligned} 0 &= \frac{d\mathbf{x}}{dt} - M\mathbf{x} - \mathbf{f}, \\ &= \left(\frac{dV}{dt} - MV\right)\mathbf{x}(t_1) + \left(\frac{dW}{dt} - MW\right)\mathbf{x}(t_n) \\ &+ \frac{d\mathbf{w}}{dt} - M\mathbf{w} - \mathbf{f}, \\ &= Z_V(t)\mathbf{x}(t_1) + Z_W(t)\mathbf{x}(t_n) + \mathbf{z}(t). \end{aligned} \tag{24}$$

Although the constraint must hold for every t there is really only one condition here because the equations for Z_V, Z_W, $\mathbf{z}$ are homogeneous so the constraint equation determined for $t = t_2$ is obtained from that for $t = t_1$ by multiplying by $Z(t_2, t_1)$ where Z is a fundamental matrix for the system

$$\frac{dZ}{dt} = -\left(2S - M\right)Z.$$

3 Optimization methodology

When the model is known then the Gauss-Newton or scoring method appears the method of choice for the embedding methods, and good reasons for this are presented which are a consequence of the stochastic setting. Similar approximations appear to work well in sequential quadratic programming applied to the simultaneous class of methods. Results from Zengfeng Li's thesis [6] are summarized.

Scoring [2] is a generalisation of the Gauss-Newton algorithm. It is based on two main ideas:

Maximum likelihood for parameter estimation This starts with:

- *events*: $\mathbf{y}_t \in R^m$, $t \in T$,
- *probability density*: $f(\mathbf{y}_t, \boldsymbol{\eta}(\mathbf{x}, t))$,
- *exact model*: $\boldsymbol{\eta}(\mathbf{x}, t) : R^p \times T \to R^q$ (the parameter and covariate information).

The parameter estimates $\mathbf{x}$ are computed by minimizing the negative of the likelihood

$$\mathbf{x}_T = \arg\min \mathcal{K}_T(\mathbf{x}), \tag{25}$$

$$\mathcal{K}_T(\mathbf{x}) = -\sum_{t\in T} \mathcal{L}_t, \tag{26}$$

$$\mathcal{L}_t = \log f(\mathbf{y}_t, \boldsymbol{\eta}(\mathbf{x}, t)). \tag{27}$$

The context which generalises that typically assumed for least squares involves:

- independent events and an appropriately structured sampling regime,
- $n = |T| \gg m = \dim\ \mathbf{x}$, and
- a model with suitable analytic properties.

Newton's method for function minimization This computes:

$$\mathcal{J} = \nabla^2\mathcal{K}(\mathbf{x}), \tag{28}$$

$$\mathbf{h} = -\mathcal{J}^{-1}(\mathbf{x})\nabla\mathcal{K}(\mathbf{x})^T, \tag{29}$$

$$\mathbf{x} \to \mathbf{x} + \mathbf{h}. \tag{30}$$

Advantages:

1 It has a fast rate of ultimate convergence to $\hat{\mathbf{x}} \ni \nabla\mathcal{K}(\hat{\mathbf{x}}) = 0$ provided $\mathcal{J}(\hat{\mathbf{x}})$ is nonsingular.
2 It has good transformation invariance properties.

Disadvantages:

1 Convergence is local and could be, at least in theory, just to a stationary point.
2 The method requires $\nabla^2\mathcal{K}(\mathbf{x})$. In the past it has often been regarded as uneconomical or inconvenient to compute this.

Scoring aims to maintain or even improve on the advantages of Newton's method while avoiding the disadvantages. The key step is the replacement of $\nabla^2\mathcal{K}(\mathbf{x})$ by its expectation. This gives the modified iteration the basic (full step) form:

$$\mathcal{I} = \frac{1}{n}\mathcal{E}\left\{\nabla^2\mathcal{K}(\mathbf{x})\right\}, \tag{31}$$

$$\mathbf{h} = -\mathcal{I}(\mathbf{x})^{-1}\frac{1}{n}\nabla\mathcal{K}(\mathbf{x})^T, \tag{32}$$

$$\mathbf{x} \to \mathbf{x} + \mathbf{h}. \tag{33}$$

In [2] it is shown that the scoring iteration gives consistent estimates $\hat{\mathbf{x}}_n$ which tend to the true parameter vector $\mathbf{x}^*$ in an appropriate stochastic sense as the number of observations increases without limit provided

- the sampling procedure is structured appropriately, and
- the assumed model is correct.

The important identity

$$\mathcal{E}\{\nabla^2\mathcal{K}(\mathbf{x})\} = \mathcal{E}\{\nabla\mathcal{K}(\mathbf{x})^T\nabla\mathcal{K}(\mathbf{x})\} \tag{34}$$

shows that $\mathcal{I}$ is generically positive definite under reasonable modelling assumptions. In [2] it is shown that $\lim_{n\to\infty}\mathcal{I}_n$ is essentially a bounded Gram matrix. A second consequence of the disappearance of second derivatives in $\mathcal{I}_n$ is that the modified algorithm has even better transformation invariance properties. If $\mathcal{I}$ has to be estimated because integration is difficult then the law of large numbers can help.

$$\begin{aligned}\frac{1}{n}\mathcal{E}\{\nabla^2\mathcal{K}_n\} &= \frac{1}{n}\sum_i \mathcal{E}\{\nabla\mathcal{L}_i^T\nabla\mathcal{L}_i\}\\ &= -\frac{1}{n}\sum_i\left(\nabla\mathcal{L}_i^T\nabla\mathcal{L}_i - \mathcal{E}\{\nabla\mathcal{L}_i^T\nabla\mathcal{L}_i\}\right)\\ &\quad +\frac{1}{n}\sum_i \nabla\mathcal{L}_i^T\nabla\mathcal{L}_i\\ &\to \frac{1}{n}\sum_i \nabla\mathcal{L}_i^T\nabla\mathcal{L}_i,\ n\to\infty.\end{aligned}$$

As $\mathcal{I}_n$ is positive (semi) definite so $\nabla\mathcal{K}_n\mathbf{h} < (=)0$. This last property ensures that the scoring step is necessarily downhill for minimizing $\mathcal{K}_n$ when $\mathcal{I}_n$ is nonsingular, and that $\mathcal{K}_n$ is a suitable function to use in a linesearch step to stabilize the iteration. This has the consequences:

1 $\frac{\nabla\mathcal{K}\mathbf{h}}{\|\nabla\mathcal{K}\|\|\mathbf{h}\|} < -\frac{1}{\mathrm{cond}\mathcal{I}}$,
2 Limit points of the iteration are stationary points of $\mathcal{K}$.
3 A full step will be acceptable in the line search eventually provided n is large enough.

The final point in favour of scoring is that the rate of convergence to a consistent estimate is very satisfactory. Once a full step is acceptable in the linesearch then the iteration can be written as the fixed point iteration:

$$\mathbf{x}_{i+1} = \mathbf{F}(\mathbf{x}_i);\quad \mathbf{F}(\mathbf{x}) = \mathbf{x} - \mathcal{I}_n(\mathbf{x})^{-1}\frac{1}{n}\nabla\mathcal{K}_n(\mathbf{x})^T$$

Here $\hat{\mathbf{x}}_n$ is a point of attraction provided the spectral radius

$$\varpi(\nabla\mathbf{F}(\hat{\mathbf{x}}_n)) < 1$$

As $\nabla\mathcal{K}_n(\hat{\mathbf{x}}_n) = 0$ it follows that

$$\begin{aligned}\nabla\mathbf{F}(\hat{\mathbf{x}}_n) &= I - \mathcal{I}_n(\hat{\mathbf{x}}_n)^{-1}\frac{1}{n}\nabla^2\mathcal{K}_n(\hat{\mathbf{x}}_n)\\ &= (\mathcal{I}_n(\hat{\mathbf{x}}_n))^{-1}\left((\mathcal{I}_n(\hat{\mathbf{x}}_n) - \frac{1}{n}\nabla^2\mathcal{K}_n(\hat{\mathbf{x}}_n))\right)\\ &= \nabla\mathbf{F}(\mathbf{x}^*) + \mathcal{O}(\|\hat{\mathbf{x}}_n - \mathbf{x}^*\|),\ \text{a.s.},\ n\to\infty\end{aligned}$$

But $\nabla\mathbf{F}(\mathbf{x}^*) = o(1),\ n\to\infty$ where $\mathbf{x}^*$ is the true vector of parameters using the strong law of large numbers

$$\Rightarrow \varpi(\nabla \mathbf{F}(\hat{\mathbf{x}}_n)) \to 0,\ n \to \infty$$

showing an arbitrary fast rate of (first order) convergence provided the effective sample size is large enough. Note that consistency of the estimate is used explicitly here.

4 Extension to constrained problems

For simplicity consider the linearly constrained problem [7]:

$$\min_{\mathbf{x}} \mathcal{K}_n;\ C\mathbf{x} = \mathbf{d},\ C : R^p \to R^m,\ \mathrm{rank}(C) = m. \tag{35}$$

The necessary conditions for a minimum of (35) give

$$\nabla \mathcal{K}_n = \boldsymbol{\lambda}^T C \tag{36}$$

where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers. The limiting form as $n \to \infty$ follows by an application of the law of large numbers to

$$\frac{1}{n}\left\{\nabla \mathcal{K}_n - \mathcal{E}\{\nabla \mathcal{K}_n\}\right\} + \frac{1}{n}\mathcal{E}\{\nabla \mathcal{K}_n\} = (\boldsymbol{\lambda}/n)^T C$$

The left hand side has the limiting form

$$-\int_0^1 \mathcal{E}\{\nabla \mathcal{L}(\mathbf{y}, \mathbf{x}, t)\} d\omega(t),$$

where ω is a limiting weight function characterizing the design of the observation process. Thus the limiting system is

$$-\int_0^1 \mathcal{E}\{\nabla \mathcal{L}(\mathbf{y}, \mathbf{x}, t)\} d\omega(t) = \boldsymbol{\lambda}^{*T} C \tag{37}$$

$$C\mathbf{x} = \mathbf{d} \tag{38}$$

where $\boldsymbol{\lambda}^* = \lim_{n\to\infty} \boldsymbol{\lambda}/n$. This has the solution

$$\mathbf{x} = \mathbf{x}^*,\ \boldsymbol{\lambda}^* = 0.$$

Thus if there is a fixed finite number of constraints then the associated Lagrange multipliers asymptote to zero. In particular, the limiting (correctly scaled) multipliers associated with (15), (16) are zero.

An equality constrained sequential quadratic programming approach to constrained problems is now sketched. The target problem here is the simultaneous method. Needed results are straightforward if cyclic reduction can be applied directly (that is rather than used to simplify a current linearization), but this is not assumed. Let the problem have the form:

$$\min_{\mathbf{x}} \mathcal{K}(\mathbf{x});\ \mathbf{c}(\mathbf{x}) = 0.$$

We introduce the Lagrangian

$$l(\mathbf{x}, \boldsymbol{\lambda}) = \mathcal{K}(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{c}(\mathbf{x}).$$

Let B_k be an approximation to $\nabla^2_{\mathbf{x}} l(\mathbf{x}_k, \boldsymbol{\lambda}_k)$ and solve linear subproblem

$$\begin{aligned} &\min_{\mathbf{d} \in S} \nabla \mathcal{K}(\mathbf{x})\mathbf{d} + \frac{1}{2}\mathbf{d}^T B_k \mathbf{d} \\ &S = \{\mathbf{d};\ \mathbf{c}(\mathbf{x}_k) + A(\mathbf{x}_k)\mathbf{d} = 0\} \end{aligned}$$

(typically cyclic reduction would have an application in the constraint reduction step). To make progress take the guarded step

$$\mathbf{x}_{k+1} := \mathbf{x}_k + \gamma \mathbf{d}_k.$$

Here Li implemented the Byrd and Omojokum trust region strategy following [8]. The current iteration is completed by updating the Lagrange multiplier vector $\boldsymbol{\lambda}$ using

$$\boldsymbol{\lambda}_{k+1} = -A_k^+ (\nabla \mathcal{K}_k^T + B_k \mathbf{d}_k))$$

The Gauss-Newton approximation to B_k (possibly guarded to avoid too large changes) involves:

i ignoring the term $\sum_{i=1}^{n} r_i \frac{\partial^2 r_i}{\partial x_j \partial x_k}$, and this is justified as in the unconstrained case;

ii ignoring also the term $\sum_{i=1}^{n-1} \boldsymbol{\lambda}_i^T \frac{\partial^2 \mathbf{c}_i}{\partial x_j \partial x_k}$.

Numerical experience is that this Gauss-Newton style approximation works. It is not too difficult to show that a suitably scaled $\boldsymbol{\lambda}_i \to 0$, but this is not quite enough to justify omitting the constraint contributions for a potentially unbounded number of constraints. However, there is more structure which includes a stochastic differential equation for a limiting multiplier vector:

$$d\boldsymbol{\lambda} = -\nabla_{\mathbf{x}} \mathbf{w}(t, \mathbf{x}, \boldsymbol{\beta})^T \boldsymbol{\lambda} dt + \sigma \boldsymbol{\phi} d\omega.$$

It is worth observing that the modified iteration just gives the unconstrained minimization step. This iteration can be shown to give consistent but not feasible estimates in the case of a fixed finite number of constraints. Here the trust region step serves to force feasibility.

The following example was given in [6] to illustrate the Gauss-Newton strategy. The distinctly unstable equation is due to Matheij [1]. The comparison is between a Newton strategy where the Hessian is computed exactly and the Gauss-Newton strategy summarized above.

$$M(t,\boldsymbol{\beta}) = \begin{bmatrix} 1-\beta_1\cos(\beta_2 t) & 0 & 1+\beta_1\sin(\beta_2 t) \\ 0 & \beta_1 & 0 \\ 1+\beta_1\sin(\beta_2 t) & 0 & 1+\beta_1\cos(\beta_2 t) \end{bmatrix}$$

$$\mathbf{f}(t) = e^t \begin{bmatrix} -1+19(\cos(2t)-\sin(2t)) \\ -18 \\ 1-19(\cos(2t)+\sin(2t)) \end{bmatrix}$$

$$\mathbf{x}(t) = e^t\mathbf{e}$$

The data is chosen in the form $\mathbf{x}(t)+\sigma\mathbf{rnd}$ where **rnd** is a vector of standardized normal variates and $\sigma = 5.,\ 1.,\ .01$. The initial parameter vector has values 20% larger than true - $[19, 2]$. The results are displayed in the following table, and show clearly the reduction in number of iterations as the number of observations is increased for each of the values of σ.

Table 1. Newton and scoring compared

n	Ne	GN	Ne	GN	Ne	GN
2^5+1	15	55	6	11	4	4
2^7+1	16	20	6	10	3	4
$2^{10}+1$	7	13	4	5	3	3

5 Model selection

It all becomes harder if the only information available is that the model is known to lie within a parameterized class of systems. Presumably one should start the searching with the simpler members of this class (the potentially under-specified systems) as an aid to numerical stability. However, the scoring method requires a consistency result and thus looses its justification in this case. A stochastic embedding procedure which produces spline-like fits to the data, and which offers the possibility of overcoming this difficulty, is being studied for systems linear in the state variables.

The smoothing spline $\eta(t)$ is defined by:

$$\min_{\eta} \sum_{i=1}^{n} (y_i - \eta(t_i))^2 + \tau \int_0^1 \left(\frac{d^k\eta}{dt^k}\right)^2 dt. \tag{39}$$

Here the value of τ can be chosen to provide a compromise between data fit and smoothness. An alternative stochastic formulation is given by Wahba [9]:

$$\begin{aligned} \eta(t) &= \mathcal{E}\left\{y(t)|y_1, y_2, \cdots, y_n, \lambda\right\}, \\ \frac{d^k\eta}{dt^k} &= \sigma\sqrt{\lambda}\frac{d\omega}{dt}. \end{aligned} \tag{40}$$

Here $\lambda = 1/\tau$. The consistency result available is $\eta(t) \to \mathcal{E}\{y(t)\}$, $n \to \infty$ provided λ is chosen appropriately. For our purposes a key step is the generalisation to more general differential operators (g-splines)

$$\min_{\eta} \sum_{i=1}^{n} (y_i - \eta(t_i))^2 + \tau \int_0^1 (\mathcal{M}_k \eta)^2 \, dt. \tag{41}$$

As τ gets large the minimizing η is forced to the null space of $\mathcal{M}_k$. This suggests choosing $\mathcal{M}_k$ to provide the possibility of identifying a linear model for the underlying signal.

An alternative approach has been considered by Wecker, Ansley, and Kohn ([10], [11] for example). They write $\mathcal{M}_k$ in first order system form

$$\frac{d\mathbf{x}}{dt} = M_k \mathbf{x}$$

giving the stochastic form corresponding to (40) (here $\mathbf{b} = \mathbf{e}_k$)

$$\begin{aligned} \eta(t) &= \mathcal{E}\left\{x_1(t) | y_1, y_2, \cdots, y_n, \lambda\right\}, \\ d\mathbf{x} &= M_k \mathbf{x} dt + \sigma\sqrt{\lambda}\mathbf{b} d\omega. \end{aligned} \tag{42}$$

This invites generalisations to the more general observation data: $y_i = \boldsymbol{\phi}^T \mathbf{x}(t_i)$, to general linear systems characterized by the matrix $M(t, \boldsymbol{\beta})$, and to more general smoothness controls characterized by the steering vector $\mathbf{b}$ [12], [13]. Let $X(t, \xi)$ be a fundamental matrix of the deterministic equation. Then variation of parameters gives the relations

$$\mathbf{x}_{i+1} = X(t_{i+1}, t_i)\mathbf{x}_i + \sigma\sqrt{\lambda}\mathbf{u}_i, \tag{43}$$

$$\mathbf{u}_i = \int_{t_i}^{t_{i+1}} X(t_{i+1}, s)\mathbf{b} d\omega(s), \tag{44}$$

$$\sim N(0, \sigma^2 R(t_{i+1}, t_i)), \tag{45}$$

$$R(t_{i+1}, t_i) = \lambda \int_{t_i}^{t_{i+1}} X(t_{i+1}, s)\mathbf{b}\mathbf{b}^T X(t_{i+1}, s)^T ds. \tag{46}$$

The system is now in a form suitable for computing the conditional expectation $\mathbf{x}(t|n)$ using the Kalman filter and interpolation smoother. The filter is a forward recursion for $\mathbf{x}_{i|i} = \mathcal{E}\left\{\mathbf{x}(t_i) | y_1, y_2, \cdots, y_i, \lambda\right\}$, and $\sigma^2 S_{i|i}$, the corresponding covariance. The interpolation smoother gives the dependence on all the data. If $t_i \le t \le t_{i+1}$ then:

$$\mathbf{x}_(t|n) = X(t, t_i)\mathbf{x}_{i|i} + A(t, t_i)\left(\mathbf{x}_{i+1|n} - \mathbf{x}_{i+1|i}\right), \tag{47}$$

$$A(t, t_i) = \left\{X(t, t_i) S_{i|i} X_i + \mathit{\Gamma}(t, t_i)\right\} S_{i+1|i}^{-1}, \tag{48}$$

$$\mathit{\Gamma}(t, t_i) = R(t, t_i) X(t_{i+1}, t)^T. \tag{49}$$

Two choices of the initial condition $\mathbf{x}_{1|0}$ are possible – either to choose it as constant or to assume the diffuse prior ($\mathbf{x}_{1|0} = 0,\ S_{1|0} \uparrow \infty$). Both possibilities have been considered for computing smoothing splines. The filter is an initial value process involving a possibly unstable deterministic component. However, it does involve a correction step to take account of the new data. Still stability is a legitimate question. In this connection note that there is a "multiple shooting" form which has proved convenient in the case of the diffuse prior [14]:

$$\min_{\mathbf{x}} \left\{ \mathbf{r}_1^T V^{-1} \mathbf{r}_1 + \mathbf{r}_2^T R^{-1} \mathbf{r}_2 \right\}, \tag{50}$$

$$\begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\phi}_1^T & & & & \\ & \boldsymbol{\phi}_2^T & & & \\ & \cdots & \cdots & & \\ & & & & \boldsymbol{\phi}_n^T \\ -X_1 & I & & & \\ & -X_2 & I & & \\ & & \cdots & \cdots & \\ & & & X_{n-1} & I \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix}, \tag{51}$$

where $V = \sigma^2 I,\ R = \sigma^2 \mathrm{diag}\{R_1, R_2, \cdots, R_{n-1}\},\ R_i = R(t_{i+1}, t_i)$. Here the conditioning of R is a potential source of problems.

Interaction between the choices of $\boldsymbol{\phi},\ \mathbf{b}$ turns out to determine the smoothness properties of the estimated state variable $\mathbf{x}(t|n)$ and thus provides useful information on how to choose them. Differentiating the interpolation smoother gives

$$\begin{aligned} \frac{d\mathbf{x}(t|n)}{dt} &= M\mathbf{x}(t|n) + \mathbf{b}\mathbf{b}^T X(t_{i+1}, t)^T \mathbf{v}, \\ \mathbf{v} &= S_{i+1|i}^{-1}(\mathbf{x}_{i+1|n} - \mathbf{x}_{i+1|i}). \end{aligned} \tag{52}$$

It follows that if the deterministic equation has smooth solutions then the state can fail to be smooth only at the points t_i. The interesting term in (52) is that involving $\mathbf{b}\mathbf{b}^T X(t_{i+1}, t)^T$. The calculations need the derivatives of $X(t_i, t)$:

$$\begin{aligned} \frac{d^j X(t_i, t)}{dt^j} &= X(t_i, t) P_j(M), \\ P_0 = I,\ P_j &= \frac{dP_{j-1}}{dt} - MP_{j-1},\ j = 1, 2, \cdots. \end{aligned} \tag{53}$$

Smoothness of $\frac{d^j \mathbf{x}(t|n)}{dt^j}$ at t_i requires [13]

$$\boldsymbol{\phi}^T P_{j-1}(M)\mathbf{b} = 0,\ j = 1, 2, \cdots. \tag{54}$$

Limits to smoothness follow necessarily if the $P_{j-1}(M)^T \boldsymbol{\phi}$ are linearly independent for $j < k$. For example, the standard smoothing spline results follow by setting

$$\boldsymbol{\phi} = \mathbf{e}_1,\ \mathbf{b} = \mathbf{e}_k,$$

$$M = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \cdots & \cdots & \\ & & & 0 & 1 \\ -m_k & -m_{k-1} & \cdots & \cdots & -m_1 \end{bmatrix}$$

The smoothness results have implications for the conditioning of $R(t_{i+1}, t_i)$ and hence for possible difficulties in methods in which $R(t_{i+1}, t_i)$ appears explicitly. If δ is small then Taylor expansion gives

$$R(t+\delta, t) = \int_t^{t+\delta} \sum_{i,j} \frac{(s-(t+\delta))^{i+j}}{i!j!} P_i(M)\mathbf{b}\mathbf{b}^T P_j(M)^T ds \qquad (55)$$

Here the Rayleigh quotient gives eigenvalue information as $\delta \to 0$.

1 Largest eigenvalue of $R(t_{i+1}, t_i)$:

$$\pi_k = \lambda\delta\mathbf{b}^T\mathbf{b} + O(\delta^2).$$

The corresponding eigenvector $\to \mathbf{b}$.

2 Smallest eigenvalue of $R(t_{i+1}, t_i)$: If the orthogonality conditions

$$\mathbf{b}^T P_{j-1}(M)^T \boldsymbol{\phi} = 0,\ j = 1, 2, \cdots, k-1,$$

are satisfied then the eigenvector associated with the smallest eigenvalue $\to \boldsymbol{\phi}$. The corresponding Rayleigh quotient is

$$\pi_1 = \frac{\lambda}{((k-1)!)^2} \frac{(\mathbf{b}^T P_{k-1}(M)^T \boldsymbol{\phi})^2}{\boldsymbol{\phi}^T\boldsymbol{\phi}} \frac{\delta^{2k-1}}{2k-1} + O(\delta^{2k}).$$

This is an upper bound for the smallest eigenvalue.

Two main approaches have been used for computing parameter estimates – generalised cross validation (GCV) [15],and generalised maximum likelihood (GML) [10],[11].The latter involves a "likelihood" approach. It takes its starting point from the observation that the innovations $\zeta_i = y_i - \boldsymbol{\phi}^T \mathbf{x}_{i|i-1}$ are independent, normally distributed with variance $\sigma^2\mathcal{V}_i$ where $\mathcal{V}_i = (1 + \boldsymbol{\phi}^T S_{i|i-1}\boldsymbol{\phi})$. The idea is to minimize

$$\sum_i{}' \left\{ \log\sigma^2 + \log\mathcal{V}_i + \frac{\zeta_i^2}{\sigma^2\mathcal{V}_i} \right\}.$$

Minimizing with respect to σ^2 gives (here $N \le n$ and the summation limits depend on the form of the initial conditions):

$$\hat{\sigma}^2 = \frac{1}{N} \sum_i{}' \frac{\zeta_i^2}{\mathcal{V}_i}.$$

Substituting back gives the concentrated likelihood

$$GML = \sum_i{}' \log \mathcal{V}_i + N \log \left(\sum_i{}' \frac{\zeta_i^2}{\mathcal{V}_i} \right).$$

The alternative uses generalised cross validation. The ojective function is

$$GCV = \frac{\sum_{i=1}^n \left(y_i - \boldsymbol{\phi}^T \mathbf{x}_{i|n} \right)^2 / n}{\{\text{trace}\{I - T\}/n\}^2},$$

were T is the influence matrix mapping observations y_i into the estimated signal $\boldsymbol{\phi}^T \mathbf{x}_{i|n}$. Advantages are claimed for its use in estimating λ. Problem is in finding an implementation that can be used for parameter estimation which requires less than $O(n^2)$ cost. In contrast, GML is relatively easy to calculate with $O(n)$ cost.

References

1. Ascher, U., Mattheij, R., Russell, R.: Numerical Solution of Boundary Value Problems for Ordinary Differential Equations. SIAM, Philadelphia (1995)
2. Osborne, M.: Fisher's method of scoring. Int. Stat. Rev. **86** (1992) 271–286
3. Tjoa, I., Biegler, L.: Simultaneous solution and optimization strategies for parameter estimation of differential-algebraic systems. Ind. Eng. Chem. Res. **30** (1991) 376–385
4. Osborne, M.: Cyclic reduction, dichotomy, and the estimation of differential equations. J. Comp. and Appl. Math. **86** (1997) 271–286
5. Hegland, M., Osborne, M.R.: Wrap-around partitioning for block bidiagonal linear systems. IMA J. Numer. Anal. **18** (1998) 373 – 383
6. Li, Z.: Parameter Estimation of Ordinary Differential Equations. PhD thesis, School of Mathematical Sciences, Australian National University (2000)
7. Osborne, M.: Scoring with constraints. ANZIAM J. **42** (2000) 9–25
8. Lalee, M., Nocedal, J., Plantenga, T.: On the implementation of an algorithm for large-scale equality constrained optimization. SIAM J. Optim. **8** (1998) 682 – 706
9. Craven, P., Wahba, G.: Smoothing noisy data with spline functions. Numer. Math. **31** (1979) 377–403
10. Wecker, W., Ansley, C.F.: The signal extraction approach to nonlinear regression and spline smoothing. J. Amer. Statist. Assoc. **78** (1983) 81–89
11. Ansley, C.F., Kohn, R.: Estimation, filtering, and smoothing in state space models with incompletely specified initial conditions. The Annals of Statistics **13** (1985) 1286–1316
12. Osborne, M.R., Prvan, T.: On algorithms for generalised smoothing splines. J. Austral. Math. Soc. B **29** (1988) 322–341
13. Osborne, M.R., Prvan, T.: Smoothness and conditioning in generalised smoothing spline calculations. J. Austral. Math. Soc. B **30** (1988) 43–56

14. Paige, C.C., Saunders, M.A.: Least squares estimation of discrete linear dynamic systems using orthogonal transformations. SIAM J. Numer. Anal. **14** (1977) 180–193
15. Wahba, G.: A comparison of GCV and GML for choosing the smoothing parameter in the generalised spline smoothing problem. Ann. Statist. **13** (1985) 1378–1402