JOURNAL OF COMPUTATIONAL PHYSICS 30, 137–144 (1979)

On the Removal of the Singularities from the Riccati Method

A. DAVEY

School of Mathematics, University of Newcastle upon Tyne, England

Received December 13, 1977

When the Riccati method is used to solve a difficult linear homogeneous two-point boundary value problem it is frequently necessary to switch between the Riccati matrix and its inverse matrix when a singularity of either is approached. This switching causes a loss of numerical accuracy partly because it is not easy to decide exactly when to switch and it is generally rather a nuisance, especially if it is desired to calculate the eigenfunction as well as the eigenvalue. Herein we point out that these singularities may be removed by considering the differential equations for the numerators and the denominators *separately* of the elements of the Riccati matrix and its inverse. It transpires that this reformulation of the Riccati method is just the compound matrix method advocated by Gilbert and Backus and rediscovered and used by Ng and Reid. We give a brief discussion of some features of the compound matrix method and we explain why it enables the standard shooting method to be used.

1. INTRODUCTION

If we wish to solve a linear homogeneous two-point boundary value problem by a shooting method and if the characteristic values of the differential operator are widely separated then the problem will be difficult in the sense that it may not be possible to use the classical standard shooting method, because the base solutions may lose too much of their linear independence. A method which is commonly used in these circumstances is the Riccati method (see, for example, Scott [4]). If this method is used for a differential problem of order $2n$, say, with n boundary conditions given at each end of the range of integration, say $0 \leqslant x \leqslant 1$, then it is usual to split the dependent variables into two sets $\mathbf{u} = (y_1, y_2, \ldots, y_n)^T$ and $\mathbf{v} = (y_{n+1}, y_{n+2}, \ldots, y_{2n})^T$ where the elements of $\mathbf{u}$ are chosen to be those which are zero when $x = 0$, if the integration is done with x increasing from 0 to 1.

The relationship between $(\mathbf{u}, \mathbf{v})^T$ at some x station and the corresponding value of $(\mathbf{u}, \mathbf{v})^T = (\mathbf{u}_0, \mathbf{v}_0)^T$ when $x = 0$ is of the form

$$\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{F} \end{pmatrix} \begin{pmatrix} \mathbf{u}_0 \\ \mathbf{v}_0 \end{pmatrix}, \tag{1}$$

where $\mathbf{C}$, $\mathbf{D}$, $\mathbf{E}$, $\mathbf{F}$, are $n \times n$ matrices whose elements will be functions of x and the matrix formed by $\mathbf{C}$, $\mathbf{D}$, $\mathbf{E}$, $\mathbf{F}$ on the right-hand side of (1) is the solution matrix of the

0021-9991/79/010137-08$02.00/0

problem. In the space of solutions which satisfy the known initial condition $\mathbf{u} = \mathbf{u}_0 = \mathbf{0}$ when $x = 0$ then

$$\mathbf{u} = \mathbf{D}\mathbf{v}_0 \qquad \text{and} \qquad \mathbf{v} = \mathbf{F}\mathbf{v}_0 . \tag{2}$$

In the Riccati method one reasons that since the $2n$-dimensional vector $(\mathbf{u}, \mathbf{v})^T$ must lie in the n-dimensional vector space spanned by $\mathbf{v}_0$, then at any x station if $\mathbf{v}$ is given then $\mathbf{u}$ will be determined by a linear relation

$$\mathbf{u} = \mathbf{R}\mathbf{v}, \tag{3}$$

and $\mathbf{R}$ is called the Riccati matrix. If $\mathbf{R}$ is nonsingular then $\mathbf{S} = \mathbf{R}^{-1}$ is called the inverse matrix so that

$$\mathbf{v} = \mathbf{S}\mathbf{u}. \tag{4}$$

From (2), (3) we have

$$\mathbf{D}\mathbf{v}_0 = \mathbf{u} = \mathbf{R}\mathbf{v} = \mathbf{R}\mathbf{F}\mathbf{v}_0 , \tag{5}$$

and this must be true for all vectors $\mathbf{v}_0$ so that

$$\mathbf{D} = \mathbf{R}\mathbf{F} \qquad \text{and} \qquad \mathbf{R} = \mathbf{D}\mathbf{F}^{-1}. \tag{6}$$

Hence we see that $\mathbf{R}$ is essentially the ratio of two matrices and in particular

$$\det \mathbf{R} = \det \mathbf{D}/\det \mathbf{F}, \tag{7}$$

so that $\mathbf{R}$ has singularities at those values of x for which $\det \mathbf{F} = 0$ on the path of integration. As the integration of the equations for the elements of $\mathbf{R}$ proceeds it is therefore necessary to avoid the zeros of $\det \mathbf{F}$, and this is usually done by switching to the inverse matrix $\mathbf{S}$. One then switches back to $\mathbf{R}$ if one approaches a zero of $\det \mathbf{D}$ and so on repeatedly until $x = 1$ is reached.

These singularities are the main disadvantage of the Riccati method, which is otherwise a neat and efficient method for solving difficult linear differential eigenvalue problems. In Section 2 we examine the elements of $\mathbf{R}$ and $\mathbf{S}$ in detail and we find that the elements of each matrix have a common denominator. This suggests the possibility of removing the singularities from the Riccati method by using the differential equations satisfied by the numerators and the denominators of the elements of $\mathbf{R}$ and $\mathbf{S}$ *separately*, instead of using the usual differential equations for the elements of $\mathbf{R}$ and $\mathbf{S}$. We investigate this possibility, and the necessary closure of the new differential system, and we find that the new formulation contains no singularities.

2. The Removal of the Singularities

We will illustrate the ideas involved by considering a simple, but sufficiently general, example of a single fourth-order linear differential equation

$$L\phi \equiv \phi'''' - a_1\phi''' - a_2\phi'' - a_3\phi' - a_4\phi = 0, \tag{8}$$

where a $'$ denotes differentiation with respect to x and a_1, a_2, a_3, a_4 are functions of x. We suppose that the initial conditions are $\phi = \phi' = 0$ when $x = 0$ and that

the range of integration is $0 \leqslant x \leqslant 1$, there will also be two boundary conditions at $x = 1$ so that we will have an eigenvalue problem. The particular form of the boundary conditions at $x = 1$ need not concern us at this stage.

Now let ϕ_1, ϕ_2 be any two linearly independent solutions of (8) both of which satisfy the known initial conditions $\phi = \phi' = 0$ at $x = 0$. Then if λ, μ are arbitrary constants the most general solution f of (8) which satisfies the boundary conditions at $x = 0$ will be of the form

$$f = \lambda\phi_1 + \mu\phi_2 , \tag{9}$$

so that also

$$f' = \lambda\phi_1' + \mu\phi_2' ,$$

and

$$f'' = \lambda\phi_1'' + \mu\phi_2'' ,$$

$$f''' = \lambda\phi_1''' + \mu\phi_2''' .$$

The usual Riccati formulation for this problem, with $\phi = \phi' = 0$ when $x = 0$, is of the form

$$\begin{pmatrix} f \\ f' \end{pmatrix} = \begin{pmatrix} r_1 & r_2 \\ r_3 & r_4 \end{pmatrix} \begin{pmatrix} f'' \\ f''' \end{pmatrix}, \tag{10}$$

so that using (9) we have

$$\begin{pmatrix} \phi_1 & \phi_2 \\ \phi_1' & \phi_2' \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \begin{pmatrix} r_1 & r_2 \\ r_3 & r_4 \end{pmatrix} \begin{pmatrix} \phi_1'' & \phi_2'' \\ \phi_1''' & \phi_2''' \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \end{pmatrix}. \tag{11}$$

Since (11) holds for all values of λ, μ it follows that

$$\begin{pmatrix} \phi_1 & \phi_2 \\ \phi_1' & \phi_2' \end{pmatrix} = \begin{pmatrix} r_1 & r_2 \\ r_3 & r_4 \end{pmatrix} \begin{pmatrix} \phi_1'' & \phi_2'' \\ \phi_1''' & \phi_2''' \end{pmatrix},$$

and hence

$$\begin{pmatrix} r_1 & r_2 \\ r_3 & r_4 \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 \\ \phi_1' & \phi_2' \end{pmatrix} \begin{pmatrix} \phi_1'' & \phi_2'' \\ \phi_1''' & \phi_2''' \end{pmatrix}^{-1}. \tag{12}$$

Now recall that ϕ_1, ϕ_2 were *any* two linearly independent solutions of (8) satisfying the conditions at $x = 0$ so that (12) tells us that $\mathbf{R}$ is invariant with respect to which pair of such solutions is chosen. In a moment we will solve (12) for r_1, r_2, r_3, r_4 but before we do this it is convenient to define six new quantities

$$\begin{aligned} y_1 &= \phi_1\phi_2' - \phi_1'\phi_2 , \\ y_2 &= \phi_1\phi_2'' - \phi_1''\phi_2 , \\ y_3 &= \phi_1\phi_2''' - \phi_1'''\phi_2 , \\ y_4 &= \phi_1'\phi_2'' - \phi_1''\phi_2' , \\ y_5 &= \phi_1'\phi_2''' - \phi_1'''\phi_2' , \\ y_6 &= \phi_1''\phi_2''' - \phi_1'''\phi_2'' , \end{aligned} \tag{13}$$

and we note the Monge identity

$$y_2 y_5 = y_3 y_4 + y_1 y_6 , \tag{14}$$

which is evident from the Laplace expansion of

$$\begin{vmatrix} \phi_1 & \phi_2 & 0 & 0 \\ \phi_1' & \phi_2' & \phi_1' & \phi_2' \\ \phi_1'' & \phi_2'' & \phi_1'' & \phi_2'' \\ \phi_1''' & \phi_2''' & \phi_1''' & \phi_2''' \end{vmatrix}$$

based on the 2×2 minors of the first two columns.

We may now conveniently write down the solution of (12) for r_1, r_2, r_3, r_4 as

$$r_1 = \frac{y_3}{y_6}, \qquad r_2 = \frac{-y_2}{y_6}, \qquad r_3 = \frac{y_5}{y_6}, \qquad r_4 = \frac{-y_4}{y_6}. \tag{15}$$

Notice in particular that each r has the same denominator

$$y_6 = \phi_1'' \phi_2''' - \phi_1''' \phi_2'' ,$$

and from (9) this will be zero when f'', f''' vanish simultaneously on the path of integration. That is when the eigenvalue problem (8) whose boundary conditions at $x = 1$ are complementary to those at $x = 0$ has characteristic lengths on the range of integration. Moreover we may readily verify that the elements s_1, s_2, s_3, s_4 of the inverse matrix **S** are given by

$$s_1 = \frac{-y_4}{y_1}, \qquad s_2 = \frac{y_2}{y_1}, \qquad s_3 = \frac{-y_5}{y_1}, \qquad s_4 = \frac{y_3}{y_1}, \tag{16}$$

so that they also have a common denominator, namely y_1.

Thus, as expected, we see that the singularities of the Riccati method are the zeros of y_6 and y_1 and it is just these zeros which we wish to avoid. Since the numerators in both (15), (16) involve y_2, y_3, y_4, y_5 only it seems natural to derive the differential system satisfied by $y_1, y_2, y_3, y_4, y_5, y_6$ instead of deriving the usual Riccati equations for r_1, r_2, r_3, r_4 from (8), (10), and (10) differentiated. Hence the logical way to remove the singularities from the Riccati method leads directly to the compound matrix method as advocated by Gilbert and Backus [1] and used by Ng and Reid [3].

The six differential equations for y_1, y_2, y_3, y_4, y_5, y_6 may be found by differentiating (13) and using (8) and they are the *linear* equations

$$\begin{aligned} y_1' &= y_2 , \\ y_2' &= y_3 + y_4 , \\ y_3' &= a_3 y_1 + a_2 y_2 + a_1 y_3 + y_5 , \\ y_4' &= y_5 , \\ y_5' &= -a_4 y_1 + a_2 y_4 + a_1 y_5 + y_6 , \\ y_6' &= -a_4 y_2 - a_3 y_4 + a_1 y_6 . \end{aligned} \tag{17}$$

The initial conditions $\phi = \phi' = 0$ at $x = 0$ now become $y_1 = y_2 = y_3 = y_4 = y_5 = 0$ and we may set $y_6 = 1$ because the system is linear. More generally if any two of ϕ, ϕ', ϕ'', ϕ''' are zero when $x = 0$ then five of $y_1, y_2, y_3, y_4, y_5, y_6$ will be zero and the remaining one can be set equal to unity as a normalizing condition.

Hence an important feature of this new differential problem is that it is an *initial* value problem which is well conditioned, because a subdominant solution is not sought, whereas the original differential problem was a two-point boundary value problem. Moreover the original problem might have been so *stiff* that it could not be solved by the standard shooting method; we discuss this point further in Section 3.

If the boundary condition at $x = 1$ is, say, $\phi' = \phi''' = 0$ then this means that it must be possible to choose λ, μ so that both

$$\lambda\phi_1' + \mu\phi_2' = 0 \qquad \text{and} \qquad \lambda\phi_1''' + \mu\phi_2''' = 0,$$

when $x = 1$ and thus the outer boundary condition is $y_5 = 0$. More general boundary conditions at $x = 1$ may require a linear combination of the y_i to be zero there.

Ng and Reid [3] have used this method, together with a Newton–Raphson iteration procedure, to calculate eigenvalues and eigenfunctions of the Orr–Sommerfeld equation for plane Poiseuille flow. We have also used the method to obtain eigenvalues and eigenfunctions of the Orr–Sommerfeld equation for a variety of basic velocity profiles, and in particular for the strange "inviscid" eigenmode of the Blasius boundary-layer profile for values of the Reynolds number up to more than 10^6.

3. Some Features of the Compound Matrix Method

Many linear ordinary differential systems which are eigenvalue problems arise from partial differential equations of mathematical physics which are dominated by the Laplace operator. Consequently the characteristic values of ordinary differential systems frequently occur in oppositely signed pairs $\pm\alpha$, $\pm\beta$,..., and this is the situation which we shall consider in this section. What we say below is only strictly valid for differential systems with constant coefficients, however, it is also the essence of the success of the compound matrix method for differential systems whose coefficients are functions of x provided that the characteristic values are relatively unchanging over the range of integration.

Consider a fourth-order problem whose operator has characteristic values $\pm\alpha$, $\pm\beta$ with $\beta \gg \alpha$ so that the problem will be difficult to solve using the standard shooting method. Two general solutions will be

$$\phi_1 = c_1 e^{-\alpha x} + d_1 e^{\alpha x} + f_1 e^{-\beta x} + g_1 e^{\beta x},$$
$$\phi_2 = c_2 e^{-\alpha x} + d_2 e^{\alpha x} + f_2 e^{-\beta x} + g_2 e^{\beta x},$$

and as the integration proceeds both will be dominated by the term with exponent β. The standard shooting method fails because the number of unknown initial conditions

exceeds the number of terms which contribute *numerically* to the solution as x increases up to the second boundary point.

However, if we were to write down $\phi_1\phi_2' - \phi_1'\phi_2$, from the above expressions, we would find that it contains terms with exponents 0, 0, $\pm\alpha$ $\pm\beta$, with either pairing allowed. The terms with exponents $\pm 2\alpha$, $\pm 2\beta$ cancel and the compound differential system satisfied by $\phi_1\phi_2' - \phi_1'\phi_2$ contains two dominating terms with exponents $\beta \pm \alpha$. Hence if two initial conditions for the compound system were unknown then the standard shooting method could be used. We have seen in Section 2 though that there will be only one unknown initial condition and so there is no difficulty at all in using the standard shooting method to solve the compound system.

For a sixth-order problem whose operator has characteristic values $\pm\alpha$, $\pm\beta$, $\pm\gamma$ with $\gamma \gg \beta \gg \alpha$ the general solution of the corresponding compound differential system contains dominating terms with exponents γ (twice) and $\gamma \pm \alpha \pm \beta$, either pairing allowed. As in the case of the fourth-order problem, the compound system again has only one unknown initial condition and so there is no difficulty in using the standard shooting method.

More generally suppose that the differential system to be solved is

$$\mathbf{y}' = \mathbf{A}\mathbf{y}, \tag{18}$$

where $\mathbf{A}$ is a $2n \times 2n$ matrix with eigenvalues λ_i where $i = 1, 2, \ldots, 2n$. Now let $\mathbf{Y}$ be the solution matrix for (18) so that

$$\mathbf{Y}' = \mathbf{A}\mathbf{Y}. \tag{19}$$

Also let

$$\mathbf{Z} \equiv \underset{n}{M}\,\mathbf{Y} \tag{20}$$

be the nth minor compound of $\mathbf{Y}$, and let $\mathbf{Z}$ be the solution matrix of the corresponding compound differential system, say

$$\mathbf{Z}' = \mathbf{B}\mathbf{Z}. \tag{21}$$

Then it follows from the formal solutions of (19), (21) that for *all* values of x

$$e^{\mathbf{B}x} = \underset{n}{M}\, e^{\mathbf{A}x}, \tag{22}$$

and so the mapping from $\mathbf{A}$ to $\mathbf{B}$ is a *linear* mapping and hence every element of $\mathbf{B}$ is a linear combination of the elements of $\mathbf{A}$. By taking x to be small, (22) gives

$$\underset{n}{M}\,\mathbf{I} + \mathbf{B}x = \underset{n}{M}\,(\mathbf{I} + \mathbf{A}x),$$

so that changing x to η

$$\mathbf{B} = \lim_{\eta \to 0} \left\{ \frac{\underset{n}{M}\,(\mathbf{I} + \mathbf{A}\eta) - \underset{n}{M}\,\mathbf{I}}{\eta} \right\}, \tag{23}$$

and each element of the right-hand side of (23) is clearly a linear function of the elements of **A**. This result was derived by London [2], it is also true when the elements of **A** depend upon x.

Moreover, if we put $x = 1$ in (22) then

$$\mathbf{B} = \log_n(M\, e^{\mathbf{A}}), \tag{24}$$

and hence **B** has eigenvalues

$$\lambda_{i_1} + \lambda_{i_2} + \cdots + \lambda_{i_n}, \qquad \text{where} \qquad 1 \leqslant i_1 < i_2 < \cdots < i_n \leqslant 2n.$$

If the eigenvalues of **A** occur in oppositely signed pairs, and if one eigenvalue of **A** dominates the general solution of (18) as the integration proceeds, then the general solution of (21) will be dominated by $\binom{2n-2}{n-1}$ terms. Again (21) will have only one unknown initial condition and so the compound matrix method easily ensures that the standard shooting method can be used to integrate (21), since a subdominant solution is not sought.

4. Concluding Remarks

We have shown in Section 2 that for a fourth-order two-point boundary value problem with an equal number of boundary conditions at each end of the range of integration then an attempt to remove the singularities from the Riccati method leads naturally to the sixth-order linear initial value problem as described by Ng and Reid [3]. Although we have restricted our attention to a simple example it can also be shown that for the general problem an attempt to remove the singularities from the Riccati method will lead directly to the same compound matrix method.

The brilliance of the compound matrix method is that it converts even a *stiff* two-point boundary value problem into an initial value problem which may be solved by standard shooting techniques. Perhaps the only worrying feature of this excellent method is that for a differential problem of order $2n$, with n boundary conditions at each end of the range of integration, the linear compound differential system will in general be of order $\binom{2n}{n}$ whereas the nonlinear Riccati formulation would have a differential order of only n^2 (or $4n^2$ if the generalized Riccati transformation [4] is used). These orders are comparable when $n = 2$ or 3 but as n increases from 4 upwards the order of the compound matrix method becomes much larger than that of the Riccati method.

Acknowledgment

I am grateful to Professor W. H. Reid for introducing me to the compound matrix method and for making his paper with Professor B. S. Ng available prior to publication.

References

1. F. Gilbert and G. E. Backus, *Geophysics* **31** (1966), 326.
2. D. London, *Linear and Multilinear Algebra* **4** (1976), 179.
3. B. S. Ng and W. H. Reid, *J. Computational Phys.* **30** (1979), 125–136.
4. M. R. Scott, *J. Computational Phys.* **12** (1973), 334.