Numerical Linear Algebra with Applications, Vol. 4(6), 443–458 (1997)

Rank-deficient Matrices as a Computational Tool

W. Govaerts* and B. Sijnave

Department of Applied Mathematics and Computer Science, University of Gent, Krijgslaan 281 – S9, B9000 Gent, Belgium

Rank-deficient matrices arise naturally in many applications. Detecting rank changes and computing parameter values for which a matrix has a prescribed (low) rank deficiency is a fundamental task in computing least squares and minimum norm solutions to systems of linear equations.

We describe an approach that originates from numerical continuation and bifurcation theory but has a wider applicability. It uses only linear solves with a bordered extension of the rank-deficient matrix and the transpose of that extension. We discuss the basic methods and their application in fundamental problems such as minimization and in more advanced problems in non-linear analysis. We present extensive numerical evidence in instructive test cases as well as in a chemical model (one-dimensional PDE) and a biological model (using the software package CONTENT for dynamical systems). © 1997 by John Wiley & Sons, Ltd.

Numer. Linear Algebra Appl., Vol. 4, 443–458 (1997)
(No. of Figures: 3 No. of Tables: 9 No. of Refs: 15)

KEY WORDS rank deficiency; bordered matrices

1. Introduction

In many applications we deal with parameter-dependent matrices $A(\alpha)$, $\alpha \in \mathbb{R}^k$, where α is a set of k parameters. Singularity (in the case of square matrices) or rank deficiency of such matrices is not necessarily a nuisance. In minimization problems it is precisely what we expect to find. In applications like continuation and bifurcation most interesting phenomena appear when certain matrices are singular or even have a higher rank deficiency. Computation of such matrices (i.e., computation of the parameter values α for which $A(\alpha)$ exhibits the prescribed rank deficiency) is therefore part of the problem.

We recall that manifolds of matrices with prescribed Jordan forms (rank deficiency is part of this setting) were studied extensively in [14] and [4]. However, the results obtained there are far from numerical applicability.

* Correspondence to W. Govaerts, Department of Applied Mathematics and Computer Science, University of Gent, Krijgslaan 281 – S9, B9000 Gent, Belgium.

Contract grant sponsor: Fund for Scientific Research (F.W.O. Belgium)

CCC 1070–5325/97/060443–16 $17.50

Received 26 November 1996
Revised 22 April 1997

Defining functions, based on bordered extensions of the matrix, were proposed in [7] for a limited (but important) class of rank deficiencies and Jordan forms; the origin of this method is in numerical continuation and bifurcation theory, see [8,9,2]; a survey is given in [3]. We will apply and discuss these ideas in practical numerical examples, including their introduction into software. The basic idea is to use only solutions with linear systems of the form

$$\begin{pmatrix} A(\alpha) & B \\ C^{\mathrm{T}} & D \end{pmatrix} \text{ or } \begin{pmatrix} A(\alpha) & B \\ C^{\mathrm{T}} & D \end{pmatrix}^{\mathrm{T}}$$

where B, C, D are bordering matrices.

In Section 2 we discuss the general method to detect rank deficiencies and to compute the parameter values associated with a certain rank defect. This is illustrated in Section 3 by a set of numerical tests in the case of an artificial but instructive example.

In Section 4 we show how to solve the linear minimum norm least squares problem using a small number of solutions with a fixed bordered extension of the given matrix and its transpose; we compare the method with the standard LAPACK routines. Numerical comparisons are given in Section 5.

In Section 6 we discuss a more advanced application : detection of a singularity of high codimension in a non-linear problem. This singularity is characterized by the rank defect of the Jacobian matrix of the set of equations that defines the problem.

In Section 7 we discuss another application : detection and numerical continuation of a path of solutions to a non-linear problem where the Jacobian matrix has a conjugate pair of complex eigenvalues with real part zero.

2. Computation of the matrix rank : the basic idea.

In the space of $n_1 \times n_2$ matrices we consider the matrices A with rank r ($r \leq \min(n_1, n_2)$), i.e., with rank defect $k = \min(n_1, n_2) - r$. It is known (and proved again in [7]) that they form a manifold with dimension $n_1 n_2 - (n_1 - r)(n_2 - r) = r(n_1 + n_2) - r^2$. Mathematically, this means that locally they look like a subspace of $\mathbb{R}^{r(n_1+n_2)-r^2}$, though the global structure of the manifold may be quite complicated. Perhaps more importantly, the manifold is locally defined by $(n_1 - r) \times (n_2 - r)$ scalar conditions which together form a system of equations with full linear rank.

Unfortunately, there are usually no global systems that characterize such manifolds, but in [7] local systems were obtained, using bordered matrices. The method requires a bordering of A with m_1 additional rows and m_2 additional columns to obtain an extension of the form

$$M = \begin{pmatrix} A & B \\ C^{\mathrm{T}} & D \end{pmatrix}$$

with B in $\mathbb{R}^{n_1 \times m_2}$, C in $\mathbb{R}^{n_2 \times m_1}$ and D in $\mathbb{R}^{m_1 \times m_2}$. The extension has to be square, hence $n_1 + m_1 = n_2 + m_2$. It has to be non-singular also, i.e. $m_1 \geq n_2 - r$, or, equivalently, $m_2 \geq n_1 - r$.

Matrices $V \in \mathbb{R}^{n_2 \times m_1}$ and $G \in \mathbb{R}^{m_2 \times m_1}$ are then defined by the system

$$M \begin{pmatrix} V \\ G \end{pmatrix} = I_{1:n_1+m_1, n_1+1:n_1+m_1} \tag{2.1}$$

where the right-hand side of (2.1) contains the m_1 rightmost columns of the n_1+m_1 identity matrix. The basic result proved in [7], Proposition 3.2, is that A and G have the same rank defect, i.e., the rank r_1 of G is given by $r_1 = m_2 - n_1 + r = m_1 - n_2 + r$. If in particular, m_1, m_2 are chosen minimally, i.e. $m_1 = n_2 - r, m_2 = n_1 - r$, then $r_1 = 0$ and so G is the zero matrix. In fact, the equations $G = 0$ then form (locally) a regular system of $(n_1 - r) \times (n_2 - r)$ defining equations for the manifold of matrices with rank r.

For numerical purposes it is important to note that the derivatives of G can be derived fairly easily from the derivatives of A. Indeed if we define the matrix $W \in \mathbb{R}^{n_1 \times m_2}$ by solving

$$\begin{pmatrix} W^{\mathrm{T}} & G \end{pmatrix} M = I_{n_2+1:n_2+m_2, 1:n_2+m_2} \tag{2.2}$$

then we have $G_z = -W^{\mathrm{T}} A_z V$ for any variable z on which A depends (cf [7]; this idea goes back to [8] and [9]).

The choice of B, C, D does not (in principle) matter provided that M is non-singular. This is generically the case if the dimensions satisfy the given requirements. The word 'generically' can be given a precise mathematical meaning in the space of all matrices but this is not very relevant for numerical purposes. In practice, it means that we can pick any random numbers to fill up the bordering vectors, though this is not necessarily the best way to choose B, C, D. Since the choice is rather problem dependent, we will discuss it in the applications.

To get the basic idea right, we first consider the case of a square matrix $A(\alpha)$ $(n_1 = n_2 = n)$, with rank defect 1 where $\alpha \in \mathbb{R}^k$ is a vector of parameters. We construct a one-bordered extension of this matrix and solve the system

$$\begin{pmatrix} A(\alpha) & b \\ c^{\mathrm{T}} & d \end{pmatrix} \begin{pmatrix} v(\alpha) \\ g(\alpha) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

where b, c and d are chosen such that the bordered matrix

$$\begin{pmatrix} A(\alpha) & b \\ c^{\mathrm{T}} & d \end{pmatrix} \tag{2.3}$$

is non-singular. Then it follows that $A(\alpha)$ is singular if and only if

$$g(\alpha) = 0 \tag{2.4}$$

holds. Also, under an appropriate transversality condition (2.4) has α as a regular solution, i.e., the Jacobian matrix of (2.4) has full rank at the solution point. Then the solution value of α can be detected by monitoring sign changes of $g(\alpha)$.

If the square matrix $A(\alpha)$ has rank defect at most 2, we construct a non-singular two-bordered extension and again solve the system

$$\begin{pmatrix} A(\alpha) & B \\ C^{\mathrm{T}} & D \end{pmatrix} \begin{pmatrix} V(\alpha) \\ G(\alpha) \end{pmatrix} = \begin{pmatrix} 0 \\ I_2 \end{pmatrix}$$

where I_2 denotes the 2×2 identity matrix. Now $A(\alpha)$ is singular iff $G(\alpha)$ is singular which means that

$$\det\,(G(\alpha)) = 0 \tag{2.5}$$

Also, $A(\alpha)$ has rank defect 2 iff

$$G(\alpha) = 0 \tag{2.6}$$

As before, these phenomena can generically be detected by monitoring sign changes and the critical parameters can be computed from the equations (2.5) and (2.6) respectively.

The choice of the borders so far has been undecided. If no a priori information is available, then we may just use randomly generated vectors which are reasonably scaled. We will do this in the examples in Sections 3 and 5. If approximations to the left and right singular spaces are available, then we can do better. First consider the case of rank defect 1. Proposition 2.2 in [7] shows that an optimal choice in the sense of a minimal spectral condition number of (2.3) is $b = \psi, c = \phi, d = 0$ where ψ (respectively, ϕ) is a left (respectively, right) singular vector of A scaled so that $\|\phi\|$ and $\|\psi\|$ lie between the first and $(n-1)^{\text{th}}$ singular values of A. So if approximations $\tilde{\psi}$ to ψ and $\tilde{\phi}$ to ϕ are available, then the choices

$$b = \tilde{\psi}, \quad c = \tilde{\phi} \text{ and } d = 0$$

are appropriate. Similarly, in the case of a matrix with rank defect 2, one should choose

$$B = \tilde{\Psi}, \quad C = \tilde{\Phi} \text{ and } D = 0$$

where $\tilde{\Psi}$ denotes an approximation to a scaled orthogonal base of the left singular space of the matrix and $\tilde{\Phi}$ an approximation to a scaled orthogonal base of the right singular space. The scaling should ideally be so that the norms of the columns of B and C are between the first and $(n-2)^{\text{th}}$ singular values of A.

3. Example 1: numerical detection of rank defect

As a first example, we construct an artificial problem for illustrating our method in the cases of a singular matrix and a matrix with rank defect 2. Consider the matrix

$$A(\alpha) = M_L \cdot \left(\begin{array}{c|cc} M_0 & & \\ \hline & \lambda_1 & 0 \\ & 0 & \lambda_2 \end{array}\right) \cdot M_R \tag{3.1}$$

with $\alpha = (\lambda_1, \lambda_2) \in \mathbb{R}^2$ and zeros in the blank spaces. In this construction, M_L and M_R are products of five (different) Householder matrices of dimension n and M_0 is a Householder matrix of dimension $n-2$. We recall that a Householder matrix of dimension n has the form

$$H_n = I_n - \frac{2uu^{\mathrm{T}}}{u^{\mathrm{T}}u}$$

where I_n denotes the identity matrix of dimension n and $u \in \mathbb{R}^n$. The components of all the Householder vectors were generated uniformly random in $[-0.5, 0.5]$.

From (3.1), it is clear that if one of the two parameters in α is zero then the matrix $A(\alpha)$ is singular. It is also obvious that $A(\alpha)$ has rank defect 2 for the choice $(\lambda_1, \lambda_2) = (0, 0)$.

We first solve the system

$$\begin{pmatrix} A(\alpha) & b \\ c^{\mathrm{T}} & d \end{pmatrix} \begin{pmatrix} v(\alpha) \\ g(\alpha) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{3.2}$$

Table 1

λ_1	λ_2	$\|v\|_2$	g
0.000	0.002	197.5266	-7.7936D-14
0.000	0.001	197.5266	-1.3082D-13
0.000	0.000	204.9146	8.2212D-14
0.000	-0.001	197.5266	3.6757D-13
0.000	-0.002	197.5266	-4.7704D-13

where $n = 100$, $m = 1$ and with borders chosen uniformly random in $[-0.5, 0.5]$ and scaled so that $\max_{i,j}(|a_{i,j}|) = \max_i(\max(|b_i|, |c_i|, |d|))$.

The following table of results (Table 1) was obtained by computations in Fortran double precision on a SUN workstation; the linear systems were solved using the LAPACK routines DGETRF and DGETRS (which amounts to Gaussian elimination with partial pivoting).

We see indeed that $g(\alpha) \approx 0$ if $A(\alpha)$ is singular, which is the case since $\lambda_1 = 0$. It is perhaps more surprising that $\|v\|_2$ does not tend to infinity if λ_2 tends to zero. The reason is that the system (3.2) is generically solvable even if $\lambda_1 = \lambda_2 = 0$. Suppose ϕ_1 and ϕ_2 are two linearly independent zero-vectors of $A(\alpha)$, then

$$\begin{pmatrix} \beta\phi_1 + \gamma\phi_2 \\ 0 \end{pmatrix}$$

is a solution to (3.2) if

$$\beta(c^{\mathrm{T}}\phi_1) + \gamma(c^{\mathrm{T}}\phi_2) = 1$$

Since c was chosen randomly, it is unlikely that c will be orthogonal to both ϕ_1 and ϕ_2, so the equation generically has a solution in β, γ. It is known from standard treatments of Gaussian elimination with partial pivoting (e.g., [5], Chapter 4) that in such cases the computed solution to (3.2) has only a moderate growth.

On the other hand, if we solve the system

$$\begin{pmatrix} A(\alpha) & b \\ c^{\mathrm{T}} & d \end{pmatrix} \begin{pmatrix} v(\alpha) \\ g(\alpha) \end{pmatrix} = \begin{pmatrix} e \\ f \end{pmatrix} \tag{3.3}$$

with an RHS that is generated uniformly random in $[-0.5, 0.5]$, then we expect that the size of v will grow to $u^{-1}\| \begin{pmatrix} e & f \end{pmatrix}^{\mathrm{T}} \| \, \|M^{-1}(\alpha)\|$, where M is the matrix used in (3.3) and u denotes the machine precision. Experimentally we obtain the results in Table 2.

Though we do not recommend using singular bordered extensions of $A(\alpha)$ in any systematic way, their accidental appearance is unavoidable in some computations and should be monitored. In a case like in Table 2 it can be detected fairly easily by the growth of the computed solution to (3.3); the normal conclusion is that one should switch to a bordered extension with more rows and columns.

When solving the system with two additional rows and columns to border $A(\alpha)$, i.e.,

$$\begin{pmatrix} A(\alpha) & B \\ C^{\mathrm{T}} & D \end{pmatrix} \begin{pmatrix} V(\alpha) \\ G(\alpha) \end{pmatrix} = \begin{pmatrix} 0 \\ I_2 \end{pmatrix}$$

one expects that $|G| = 0$ if $A(\alpha)$ is singular and that $G = 0$ if $A(\alpha)$ has rank defect 2, i.e., if $(\lambda_1, \lambda_2) = (0, 0)$. This is confirmed by Table 3 where g_{11}, g_{12}, g_{21} and g_{22} denote the

Table 2

λ_1	λ_2	$\|v\|_2$	g
0.000	2.0D-03	14472.2990	75.6895
0.000	1.0D-03	14287.6101	75.6895
0.000	1.0D-04	11864.2915	75.6895
0.000	1.0D-05	53537.4841	75.6895
0.000	1.0D-06	605017.1447	75.6895
0.000	0.0D+00	2.1734D+16	39.2714
0.000	-1.0D-03	15050.3637	75.6895
0.000	-2.0D-03	14853.8306	75.6895

Table 3

λ_1	λ_2	g_{11}	g_{12}	g_{21}	g_{22}	$\|G\|$
0.0	0.002	-4.58D-03	3.37D-03	-1.34D-02	9.91D-03	1.15D-19
0.0	0.001	-2.43D-03	1.79D-03	-7.15D-03	5.27D-03	4.62D-18
0.0	0.0	7.77D-16	5.42D-16	7.40D-16	8.53D-16	2.62D-31
0.0	-0.001	2.79D-03	-2.05D-03	8.19D-03	-6.04D-03	-5.92D-18
0.0	-0.002	6.01D-03	-4.43D-03	1.76D-02	-1.30D-02	-1.33D-17

elements of G. The bordering matrices B, C, D were again generated uniformly random in $[-0.5, 0.5]$ and scaled so that the maximal absolute value of the entries of A is also the maximal absolute value of the entries of the union of B, C, D.

4. Minimum-norm least-squares solutions

Suppose that $A \in \mathbb{R}^{n_1 \times n_2}$ and a vector $b \in \mathbb{R}^{n_1}$ are given. We want to compute the minimum-norm least squares solution to the problem

$$Ax = b \tag{4.1}$$

with $x \in \mathbb{R}^{n_2}$.

We first perform some preprocessing work on A (not involving b). Suppose that A has rank $r \leq \min(n_1, n_2)$ and rank defect $k = \min(n_1, n_2) - r$. These are not known in advance (they are part of the computation) but we assume that k is small (say, $k = 1, 2, 3, \ldots$) and that we can border A with m_1 additional rows and m_2 additional columns so that its bordered extension

$$\begin{pmatrix} A & B \\ C^{\mathrm{T}} & D \end{pmatrix} \tag{4.2}$$

is non-singular. This requires $n_1 + m_1 = n_2 + m_2$ and $n_2 - m_1 = n_1 - m_2 \leq r$, so only a lower bound for r is required. We note that $B \in \mathbb{R}^{n_1 \times m_2}$, $C \in \mathbb{R}^{n_2 \times m_1}$, $D \in \mathbb{R}^{m_1 \times m_2}$. Now compute $V \in \mathbb{R}^{n_2 \times m_1}$ and $G \in \mathbb{R}^{m_2 \times m_1}$ by solving

$$\begin{pmatrix} A & B \\ C^{\mathrm{T}} & D \end{pmatrix} \begin{pmatrix} V \\ G \end{pmatrix} = I_{1:n_1+m_1, n_1+1:n_1+m_1} \tag{4.3}$$

Then G has rank $r_1 = m_2 - n_1 + r = m_1 - n_2 + r$ and rank defect $k = \min(m_1, m_2) - r_1 = \min(n_1, n_2) - r$. Since G is a small dense matrix we can compute r_1 easily by standard methods, e.g., by a complete orthogonal factorization

$$Q^{\mathrm{T}} G Z = \begin{pmatrix} T & 0 \\ 0 & 0 \end{pmatrix} \tag{4.4}$$

where Q, Z are orthogonal and T is a non-singular upper triangular $r_1 \times r_1$ matrix (this can be done in LAPACK by a QR decomposition with column pivoting followed by another QR decomposition, cf. LAPACK routines SGELSX or DGELSX). We now construct an orthogonal base of $\mathbb{R}^{m_1}$

$$\{\xi_1, \ldots, \xi_{r_1}, \xi_{r_1+1}, \ldots, \xi_{m_1}\}$$

such that $G\xi_1, G\xi_2, \ldots, G\xi_{r_1}$ span the range of G and $G\xi_{r_1+1} = \ldots = G\xi_{m_1} = 0$. If (4.4) was performed, then the columns of Z form precisely such a base. Now define $\xi^{(1)} = [\xi_1, \ldots, \xi_{r_1}]$ and $\xi^{(2)} = [\xi_{r_1+1}, \ldots, \xi_{m_1}]$. By multiplying (4.3) with $\xi^{(2)}$ on the right, one finds that $V\xi^{(2)}$ has full rank $m_1 - r_1 = n_2 - r$. From (4.3) we also find that $A\,V\xi^{(2)} = 0$, meaning that $V\xi^{(2)}$ spans the right singular space of A. Next we compute $W \in \mathbb{R}^{n_1 \times m_2}$ by solving

$$\begin{pmatrix} W^{\mathrm{T}} & G \end{pmatrix} \begin{pmatrix} A & B \\ C^{\mathrm{T}} & D \end{pmatrix} = I_{n_2+1:n_2+m_2, 1:n_2+m_2} \tag{4.5}$$

and construct an orthogonal base

$$\{\eta_1, \ldots, \eta_{r_1}, \eta_{r_1+1}, \ldots, \eta_{m_2}\}$$

for $\mathbb{R}^{m_2}$ such that $G^{\mathrm{T}}\eta_1, \ldots, G^{\mathrm{T}}\eta_{r_1}$ span the range of G^{T} and $G^{\mathrm{T}}\eta_{r_1+1} = \ldots = G^{\mathrm{T}}\eta_{m_2} = 0$. If (4.4) was performed, then the columns of Q form precisely such a base.

Put $\eta^{(1)} = [\eta_1, \ldots, \eta_{r_1}]$ and $\eta^{(2)} = [\eta_{r_1+1}, \ldots, \eta_{m_2}]$. In the same way as before we find that $W\eta^{(2)}$ has full rank $m_2 - r_1$ and spans the left singular space of A, i.e., the orthogonal complement of $\mathfrak{R}(A)$.

From (4.3) it follows that $A\,V\xi^{(1)} + B\,G\xi^{(1)} = 0$. So $B\,(G\xi^{(1)})$ is in the range of A and we know that the r_1 columns of $G\xi^{(1)}$ are linearly independent. But $B(\mathbb{R}^m)$ must also contain a $n_1 - r$ dimensional space complementary to the range of A since the matrix (4.2) was chosen to be non-singular. Since $r_1 + (n_1 - r) = m_2 - n_1 + r + n_1 - r = m_2$, it follows that if $s \in \mathbb{R}^{m_2}$ is any vector for which $B\,s \in \mathfrak{R}(A)$ holds, then necessarily s is in the span of $G\xi^{(1)}$, i.e., there exists a $t \in \mathbb{R}^{r_1}$ such that $s = G\xi^{(1)}t$.

Now consider again (4.1). We first project b onto the range of A, i.e. we consider

$$b_1 = b - W\eta^{(2)}\zeta \tag{4.6}$$

where $\zeta \in \mathbb{R}^{m_2 - r_1}$ is chosen so that $b - W\eta^{(2)}\zeta \in \mathfrak{R}(A)$, or, equivalently, so that

$$(W\eta^{(2)})^{\mathrm{T}}(W\eta^{(2)})\zeta = (W\eta^{(2)})^{\mathrm{T}} b \tag{4.7}$$

Now we solve

$$\begin{pmatrix} A & B \\ C^{\mathrm{T}} & D \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} b_1 \\ 0 \end{pmatrix} \tag{4.8}$$

from which it follows that $B\,q \in \mathfrak{R}(A)$. So there exists a $t \in \mathbb{R}^{r_1}$ such that $q = G\xi^{(1)}t$.

We find t by solving

$$(G\,\xi^{(1)})^{\mathrm{T}}\,(G\,\xi^{(1)})\,t = (G\,\xi^{(1)})^{\mathrm{T}}\,q \tag{4.9}$$

Now we have $A\,p + B\,G\,\xi^{(1)}\,t = b_1$ and since $B\,G = -A\,V$ (cf. (4.3)), we can also write $A\,(p - V\,\xi^{(1)}\,t) = b_1$. If we put

$$x_1 = p - V\,\xi^{(1)}\,t \tag{4.10}$$

then x_1 is a least squares solution of (4.1). To find a minimum-norm least squares solution we put

$$x = x_1 + (V\,\xi^{(2)})\,\lambda \tag{4.11}$$

and solve

$$(V\,\xi^{(2)})^{\mathrm{T}}\,(V\,\xi^{(2)})\,\lambda = -(V\,\xi^{(2)})^{\mathrm{T}}\,x_1 \tag{4.12}$$

to find λ.

The main computational work in this algorithm consists of m_1 solves with the matrix (4.2) in (4.3), m_2 solves with its transpose in (4.5) and one more direct solve for every right-hand side vector in (4.8). The other computations are in the factorization (4.4) and in (4.6), (4.7), (4.9), (4.10), (4.11) and (4.12). If A is a large dense and nearly square matrix, say $n_1 \simeq n_2 \simeq n$ and m_1, m_2 are small compared with n then the bulk of the work is in the direct factorization of the matrix (4.2) and the number of flops (addition plus multiplication and some index computations) has order $\frac{1}{3}n^3$ ([5], Chapter 4.2). The computational work for an orthogonal factorization of A has order $\frac{2}{3}n^3$, i.e. double the previous amount ([5], Chapter 6.2). So the present method actually requires less computational work than the standard algorithm. Of course the accuracy of the method will depend strongly on the condition of the matrix in (4.2). So we feel that it can be recommended mainly in cases where a complete orthogonal factorization is not possible but linear systems can be solved (e.g., in the case of large sparse matrices).

5. Example 2: a minimum-norm least-squares solution

As an illustration of the method described in the previous section, we consider the calculation of the minimum-norm least-squares solution to $A\,x = b$, where $A = A(\alpha)$ is defined in (3.1) and b is generated randomly in $\mathbb{R}^n$. So $n_1 = n_2 = n$, $m_1 = m_2 = m$. We used $n = 50, m = 2$ in our tests. The entries of the bordering matrices B, C, D were generated uniformly random in $[-0.5, 0.5]$ and scaled as in Section 2. For the parameters we chose $\alpha = (0.002, 0.0)$, i.e. the matrix A has rank defect $k = 1$.

By solving the bordered system

$$\begin{pmatrix} A & B \\ C^{\mathrm{T}} & D \end{pmatrix}\begin{pmatrix} V \\ G \end{pmatrix} = \begin{pmatrix} 0 \\ I_2 \end{pmatrix}$$

one gets

$$G = \begin{pmatrix} -0.0046035 & 0.00812417 \\ 0.00275985 & -0.00487055 \end{pmatrix}$$

and $|G| = 1.62292 \times 10^{-18}$, so the 2×2 matrix G is singular, as expected. The vectors

$\xi_1, \xi_2, \eta_1, \eta_2$ can easily be found directly and we obtain

$$\xi_1 = \begin{pmatrix} -0.4929952 \\ 0.8700320 \end{pmatrix}, \xi_2 = \begin{pmatrix} 0.8700320 \\ 0.4929952 \end{pmatrix}$$

$$\eta_1 = \begin{pmatrix} 0.8576771 \\ -0.5141887 \end{pmatrix}, \eta_2 = \begin{pmatrix} -0.5141887 \\ -0.8576771 \end{pmatrix}$$

Further we find

$$\begin{aligned} (W\eta_2)^{\mathrm{T}}(b - (W\eta_2\zeta)) &= -3.469447 \times 10^{-18} \\ (W\eta_1)^{\mathrm{T}}(b - (W\eta_2\zeta)) &= 6.227227 \end{aligned} \tag{5.1}$$

From (5.1) it is clear that $(b - (W\eta_2\zeta)) \in \Re(A)$.

We also compared our method with the result by means of the LAPACK routine DGELSX. The Euclidean norm of the LAPACK solution x_L is :

$$\|x_L\|_2 = 1458.1355415651$$

For the solution by bordered matrices x_B we got

$$\|x_B - x_L\| = 2.61396049 \times 10^{-10}$$

so that the relative difference is

$$\|x_B - x_L\|/\|x_L\| = 1.79209598 \times 10^{-13}$$

which illustrates the accuracy of the solution by bordered matrix methods.

6. Example 3: singularities in a non-linear problem

As a more advanced application we consider the Brusselator (cf. [13]) where we have a reaction–diffusion equation governed by

$$\left\{ \begin{aligned} \frac{\partial X}{\partial t} &= \frac{D_X}{L^2}\frac{\partial^2 X}{\partial z^2} - ((B+1)X - X^2Y - A(A_0, L, D_A, z)) \\ \frac{\partial Y}{\partial t} &= \frac{D_Y}{L^2}\frac{\partial^2 Y}{\partial z^2} - (X^2Y - BX) \end{aligned} \right. \tag{6.1}$$

with $z \in [0, 1]$, $t \in [0, +\infty[$, L is the length of the one-dimensional reactor, D_X the diffusion coefficient of X, D_Y the diffusion coefficient of Y and

$$A(A_0, L, D_A, z) = \frac{A_0}{1 + \exp(P)}\exp(Pz) + \frac{A_0\exp(P)}{1 + \exp(P)}\exp(-Pz)$$

with $P = \frac{L}{\sqrt{D_A}}$.

We impose the Dirichlet boundary conditions $X(0) = X(1) = A_0$, $Y(0) = Y(1) = B/A_0$. The equilibrium equations are now discretized by a mesh of 42 equidistant internal

Table 4

$u_{1+4i}, i = 0, \ldots, 10$	$u_{2+4i}, i = 0, \ldots, 10$	$u_{3+4i}, i = 0, \ldots, 9$	$u_{4+4i}, i = 0, \ldots, 9$
4.4185838410825	1.7166251083831	4.2918010162073	1.7678240768052
4.1635037412020	1.8191140088368	4.0334866676875	1.8704237908616
3.9019060321546	1.9216053060640	3.7692536952841	1.9724388384093
3.6363193117997	2.0226412639883	3.5041414041340	2.0718768282817
3.3739496256918	2.1197699918002	3.2471017096471	2.1659195737336
3.1250193208394	2.2099132724050	3.0091271501162	2.2513416156210
2.9007991077070	2.2898104945305	2.8013144898589	2.3249516397705
2.7118256992968	2.3564306696349	2.6333377206709	2.3839526300582
2.5666982996089	2.4072652111517	2.5125968082830	2.4261600313354
2.4715691891117	2.4404725090697	2.4440061677227	2.4500808897998
2.4301620782259	2.4549049694102		

points of [0, 1], resulting in 84 state variables $u(i), i = 1, \ldots, 84$ with

$$u(i) = \begin{cases} \text{concentration of } X \text{ for } i \text{ odd} \\ \text{concentration of } Y \text{ for } i \text{ even} \end{cases}$$

The discretization uses a Numerov method, i.e. in the grid point z_j the values of the non-linear functions $f(X_j, Y_j, z_j)$ that appear in (6.1) are replaced by

$$\frac{1}{12}(f(X_{j-1}, Y_{j-1}, z_{j-1}) + 10f(X_j, Y_j, z_j) + f(X_{j+1}, Y_{j+1}, z_{j+1}))$$

to obtain a higher order of accuracy.

In this way we get a non-linear system

$$F(u, D_X, D_Y, A_0, L, D_A, B) = 0 \tag{6.2}$$

that describes the equilibrium solutions of the reaction equation for the six parameters D_X, D_Y, A_0, L, D_A, B.

The solution set of (6.2) is an extremely complicated geometric object that requires advanced analytical methods even to just describe its local structure. We will discover a point where the Jacobian matrix F_u has rank defect 2. In general, i.e. in a problem with no symmetry, this is a codimension four case and requires four free parameters. In practice it is not possible to find such points directly; one computes singular points of increasing codimension by freeing successively more parameters. For singularities with a distinguished bifurcation parameter we refer to [6]. However, in a problem with $\mathbb{Z}_2$-symmetry (such as the Brusselator) a Jacobian with rank defect 2 is a codimension two phenomenon.

For the present purposes only the last stage is important. Thereby we start from a point with parameter values

$$(D_X, D_Y, A_0, L, D_A, B) = (0.0016, 0.008, 4.5444340712, 0.1353905795,\\ 0.0209709146, 7.5687934072)$$

This point is in the symmetric solution space of (6.2), i.e., $u_{84-2i+1} = u_{2i-1}$ and $u_{84-2i+2} = u_{2i}$ for $i = 1, \ldots, 21$. So it is sufficient to give $u_j, j = 1, \ldots, 21$, see Table 4.

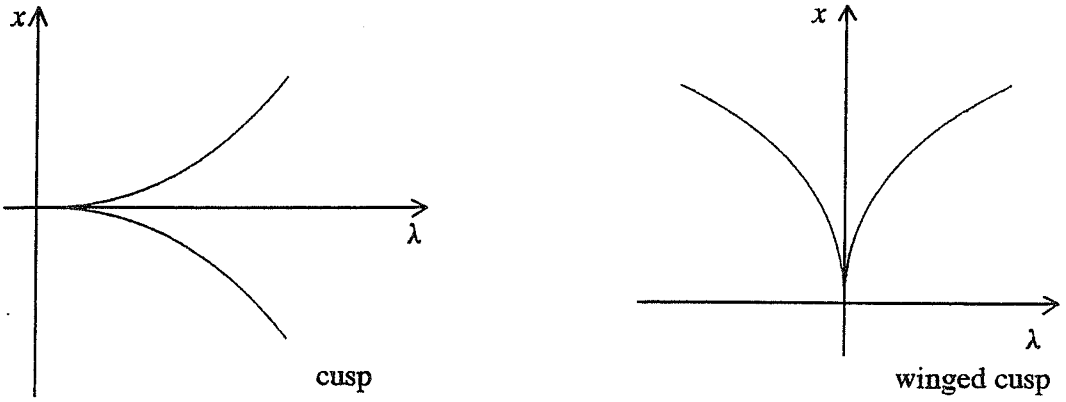

Figure 1

We computed a branch of cusp points in (u, L)-space through this point. For a precise definition of cusp points we refer to the literature, e.g., [6]. Loosely speaking, a cusp point is a solution point to (6.2) where the solution set (with fixed D_X, D_Y, A_0, D_A, B) locally looks like that of $x^2 - \lambda^3 = 0$ in $\mathbb{R}^2$. In Figure 1 a cusp point in (λ, x)-space is shown.

Computationally one expresses two requirements. First, the solution set to (6.2) has a singular point, i.e., the Jacobian $[F_u, F_L]$ does not have full rank. We recall that the matrices with rank 83 form a manifold with codimension $(84 - 83) \times (85 - 83) = 2$ in the space of the matrices with dimension 84×85, so two scalar conditions are needed. Second, the two solution curves through the point must have a common tangent (otherwise, we have generically a simple bifurcation point instead of a cusp). This condition involves second-order derivatives of F.

Our starting point was found to be a cusp in previous computations. In fact it is a winged cusp in (L, u)-space, i.e., the common tangent is orthogonal to the L axis, cf. [6]. Figure 1 also shows a winged cusp in a two - dimensional (λ, x)-space.

With D_X, D_Y fixed and A_0, L, D_A, B free, we compute by numerical continuation a curve of cusp points in (u, L)-space. In all points of this curve the Jacobian F_u is singular. It turns out that on the curve there is a point where the Jacobian has rank defect 2. The effective continuation of the curve was performed using a doubly bordered Jacobian matrix; an attempt by using single borders failed because of the presence of the rank defect 2 point.

In the sequel we describe side computations that were done in every computed point without influencing the continuation itself.

By solving the system

$$\left(\begin{array}{c|c} F_u & b \\ \hline c^{\mathrm{T}} & d \end{array}\right)\begin{pmatrix} v \\ g \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

with b, c and d fixed scaled random vectors and F_u the Jacobian along the curve, we get the values shown in Table 5 in four points of the curve.

In this table, the first row contains the values at the starting point of our curve. The second row represents the values in the continuation point just before the critical point (rank defect 2). The third row gives the situation just after the critical point and the last row contains the values some points behind the critical point.

Table 5

$\|v\|_2$	g
0.598148242	-6.385896D-18
0.576804133	1.579223D-10
0.576934851	1.708340D-09
0.573945687	6.521427D-11

Table 6

g_{11}	g_{12}	g_{21}	g_{22}	$\|G\|$
-3.2176D-02	2.4127D-02	3.4067D-02	-2.5546D-02	-1.0842D-19
-2.0394D-05	1.5299D-05	2.1340D-05	-1.6009D-05	-3.6705D-15
3.0905D-05	-2.3189D-05	-3.2338D-05	2.4266D-05	6.0176D-14
3.6961D-05	-2.7781D-05	-3.8665D-05	2.9062D-05	2.7455D-15

The starting point was obtained by an attempt to locate the winged cusp point accurately; this explains why $|g|$ is very small indeed at that point.

We see that indeed $g \approx 0$ all along the curve as we suspected. If we border F_u with two additional rows and columns (again with fixed scaled random entries) and solve

$$\left(\begin{array}{c|c} F_u & B \\ \hline C^{\mathrm{T}} & D \end{array}\right)\begin{pmatrix} V \\ G \end{pmatrix} = \begin{pmatrix} 0 \\ I_2 \end{pmatrix}$$

with F_u the Jacobian along the curve, we now find (for the same points as in Table 5) the values in Table 6.

In this table we see that at the critical point all components of G change sign indicating a matrix with rank defect 2.

It is easier to detect a sign change than to decide whether or not something is (approximately) zero. To compare with a more standard approach we computed the singular values $\sigma_i (i = 1, \ldots, 84)$ of the Jacobian along the curve to see whether the evolution of the singular values gives sufficient indication of the rank change. The greatest singular value (σ_1) and the least three singular values (σ_{82}, σ_{83} and σ_{84}) for the points in Table 5 are shown in Table 7. The underlined value in the third row is small which is an indication of a matrix with rank defect 2. Of course, one could argue that there is not really that great a difference between this value and the corresponding values in the other rows, as the order of magnitude is concerned. This is understandable because σ_{83} attains a minimum in a point with rank defect 2 but does not change sign.

Table 7. Singular values of the Jacobian F_u

σ_1	σ_{82}	σ_{83}	σ_{84}
4.0085	1.6980D-02	1.5737D-03	1.3542D-16
4.0173	1.3237D-02	2.5265D-06	4.6110D-10
4.0173	1.3227D-02	$\underline{9.2097\text{D-}07}$	9.9833D-10
4.2676	1.6967D-02	2.8932D-05	3.5534D-13

Table 8. Singular values of one-bordered extension of F_u

σ_1	σ_{83}	σ_{84}	σ_{85}
6.0497	2.0228D-02	1.4421D-02	1.5705D-03
6.0548	2.1483D-02	1.2946D-02	2.5166D-06
6.0548	2.1498D-02	1.2935D-02	9.1737D-07
6.1023	2.4734D-02	1.5803D-02	2.8766D-05

Table 9. Singular values of two-bordered extension of F_u

σ_1	σ_{84}	σ_{85}	σ_{86}
7.9220	1.6685D-02	1.2167D-02	1.6160D-03
7.9286	1.2208D-02	9.1958D-03	2.3174D-03
7.9286	1.2200D-02	9.1867D-03	2.3205D-03
7.9888	1.7830D-02	4.7102D-03	2.1894D-03

If we compute the singular values of the one-bordered extension of the Jacobian we get (again with the points from Table 5) the values in Table 8.

Indeed, at the critical point the smallest singular value apparently goes through a minimum, with the same remark as before concerning the sensitivity of the SVD to rank changes. To be complete, the singular values of the two-bordered extension of the Jacobian are presented in Table 9.

In Table 9 none of the singular values is particularly small near the critical point. This suggests that the Jacobian F_u has rank defect at most two at the critical point.

7. Example 4: detection and computation of Hopf points

A Hopf point of a dynamical system is an equilibrium solution where the Jacobian matrix has a conjugate pair of complex eigenvalues with real part zero. A Bogdanov–Takens point is a point where it has a zero eigenvalue with algebraic multiplicity two and geometric multiplicity one. These cases have important dynamic implications [11] and can be computed using bordered matrices. We illustrate the basic idea using the software package CONTENT developed at CWI (Amsterdam) by Kuznetsov and Levitin [12]. This has the advantage of making the method, in principle, applicable to any system. The model is the following eutrophication model introduced in [1] and numerically studied in [15] :

$$\left\{\begin{array}{rcl} \dot{x}_1 & = & x_1(0.2(\lambda_1 - x_1 - x_2) - 0.445x_3 - 4) \\ \dot{x}_2 & = & -0.0455x_2x_3 + 4x_1 \\ \dot{x}_3 & = & \lambda_2(10 - x_3) - 2.67x_3(0.445x_1 + 0.0455x_2) \end{array}\right.$$

We detect and numerically path-follow the same Hopf point by two techniques. We first use the default method of CONTENT (present state of development !) and then discuss our implementation using a bordered matrix.

We start from the point

$$(x_1, x_2, x_3, \lambda_1, \lambda_2) = (0.2359621, 4.53947, 4.56968, 34.94297, 0.7)$$

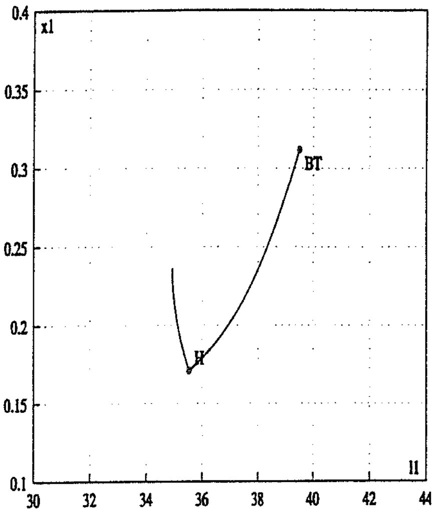

Figure 2

which happens to be a limit point with respect to λ_1. We continue the curve of equilibria with fixed $\lambda_2 = 0.7$ to find a Hopf point H with

$$(x_1, x_2, x_3, \lambda_1, \lambda_2) = (0.1708848, 2.621508, 5.730609, 35.543, 0.7)$$

Then we free λ_2 and use H as an initial point for the computation of a curve of Hopf points. On this curve we find a Bogdanov–Takens point BT with

$$(x_1, x_2, x_3, \lambda_1, \lambda_2) = (0.31177, 4.016539, 6.823618, 39.51084, 1.843967)$$

A projection on the (λ_1, x_1)-space, produced by CONTENT, is given in Figure 2. We will now discuss an implementation that uses bordered biproduct matrices.

The biproduct matrix $2A \odot I_n$ is constructed as

$$(2A \odot I_n)_{(i,j),(k,l)} = \begin{cases} -a_{il} & \text{if } k = j \\ a_{ik} & \text{if } k \neq i \text{ and } l = j \\ a_{ii} + a_{jj} & \text{if } k = i \text{ and } l = j \\ a_{jl} & \text{if } k = i \text{ and } l \neq j \\ -a_{jk} & \text{if } l = i \\ 0 & \text{else} \end{cases}$$

where $A = (a_{i,j})$ is a square matrix of dimension n. In this construction $i > j$ and $k > l$, so the biproduct matrix $2A \odot I_n$ has dimension $\frac{n(n-1)}{2} \times \frac{n(n-1)}{2}$. See [10] for details. The main property is that the eigenvalues of $2A \odot I_n$ are the sums of pairs of eigenvalues of A (cf. [10]). In particular, $2A \odot I_n$ is singular iff A has two eigenvalues with sum zero. The method originally implemented in CONTENT uses the determinant of the biproduct matrix as an indicator of Hopf points. A possible disadvantage of this choice is that there is no simple way to compute derivatives of the determinant function, except by finite differences.

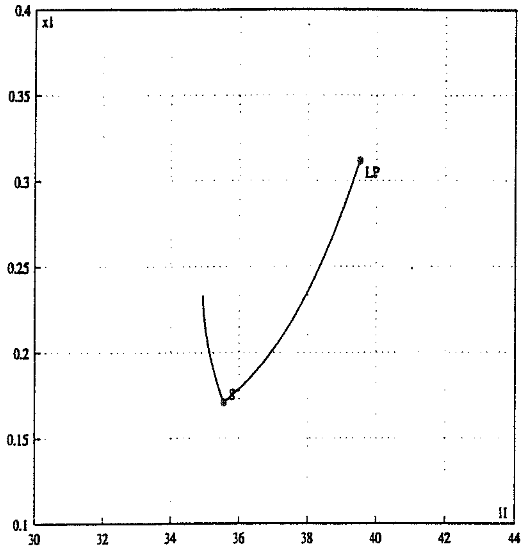

Figure 3

Instead of this, we solve the system

$$\left(\begin{array}{c|c} 2A \odot I_n & b \\ \hline c^{\mathrm{T}} & 0 \end{array}\right)\begin{pmatrix} v \\ g \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

where A is the Jacobian matrix F_u of the system and with b and c random vectors. We then use the resulting g as detection function for Hopf points as well as defining function for their continuation. We declared g as a user-defined function (a facility provided in CONTENT) and got Figure 3.

Figure 3 was also drawn by CONTENT. We note that on a curve of Hopf points a BT-point can also be interpreted as a limit point since the Jacobian matrix is singular in a BT point. This explains the difference in notation between Figure 2 and Figure 3.

REFERENCES

1. J. M. Anderson. Eutrophication of lakes. In D. L. Meadows and H. Donella, editors, *Toward Global Equilibrium*, Wright-Allen Press, Cambridge, MA, 1973.
2. W.-J. Beyn. Defining equations for singular solutions and numerical applications. In T. Küpper, H. Mittelman and H. Weber, editors, *Numerical methods for bifurcation problems*, pages 42–56, Birkhäuser, 1984.
3. W.- J. Beyn. Numerical Methods for dynamical systems. In W. Light, editor, *Advances in Numerical Analysis*, Vol I : *Nonlinear Partial Differential Equations and Dynamical Systems*, pages 175–236, Clarendon Press, Oxford, 1991.
4. J. W. Demmel and A. Edelman. The dimension of matrices (matrix pencils) with given Jordan (Kronecker) canonical form. *Lin. Alg. Appl.*, 230, 61–87, 1995.

5. G. H. Golub and C. F. Van Loan. *Matrix Computations*. John Hopkins, 1983.
6. M. Golubitsky and D. G. Schaeffer. *Singularities and Groups in Bifurcation Theory*, Vol I, Applied Mathematical Sciences 51, Springer Verlag, 1985.
7. W. Govaerts. Defining functions for manifolds of matrices. *Lin. Alg. Appl.*, in press.
8. A. Griewank and G. W. Reddien. Characterization and computation of generalized turning points. *SIAM J. Numer. Anal.*, 21, 176–185, 1984.
9. A. Griewank and G. W. Reddien. Computation of cusp singularities for operator equations and their discretizations. *J. Comput. Appl. Maths.*, 26, 133–153, 1989.
10. J. Guckenheimer, M. Myers and B. Sturmfels. Computing Hopf bifurcations I. *SIAM J. Numer. Anal.*, 43, 1–21, 1997.
11. Yu. A. Kuznetsov. *Elements of Applied Bifurcation Theory*. Applied Mathematical Sciences, Vol. 112, Springer Verlag, 1995.
12. Yu. A. Kuznetsov and V. V. Levitin. CONTENT: a multiplatform environment for analyzing dynamical systems. Dynamical Systems Laboratory, Centrum voor Wiskunde en Informatica, 1996 (software under development).
13. D. Roose and B. De Dier. Numerical determination of an emanating branch of Hopf bifurcation points in a two-parameter problem. *SIAM J. Sci. Stat. Comput.*, 10, 671–685, 1989.
14. W. C. Waterhouse. The codimension of singular matrix pairs. *Lin. Alg. Appl.*, 57, 227–245, 1984.
15. B. Werner. Computation of Hopf bifurcation with bordered matrices *SIAM J. Numer. Anal.*, 33, 435–455,1996.