NUMERICAL ANALYSIS AND MATHEMATICAL MODELLING
BANACH CENTER PUBLICATIONS, VOLUME 24
PWN POLISH SCIENTIFIC PUBLISHERS
WARSAW 1990

TRANSFORMATION METHODS FOR DIFFERENTIAL-ALGEBRAIC EQUATIONS

EBERHARD GRIEPENTROG

Section of Mathematics, EMA University,
Greifswald, German Democratic Republic

In order to get first informations about differential-algebraic equations (DAEs) in R^m we deal with the linear system

$$Ax'(t)+Bx(t) = q(t) \tag{1}$$

with constant coefficients. The solution behaviour of this system is characterized by the so-called *matrix pencil* (A, B). The uniqueness of the solutions for given initial conditions requires the regularity of the pencil, i.e. $\det(zA+B) \not\equiv 0$ $(z \in C)$. For regular pencils a decomposition (cf. [2])

$$A = E\,\mathrm{diag}(I_r, J)F, \qquad B = E\,\mathrm{diag}(W, I_{m-r})F \tag{2}$$

with nonsingular matrices E and F and a nilpotent matrix J is possible. The smallest integer l with $J^l = 0$ is called the *index of the pencil:* $l = \mathrm{ind}(A, B)$. The integers r and l are independent of the special choice of E and F. For nonsingular matrices A the matrix J does not occur, then $\mathrm{ind}(A, B) = 0$ per definition. Obviously, we get $J = 0$ iff $\mathrm{ind}(A, B) = 1$. Moreover $\mathrm{ind}(A, B) = \mathrm{ind}\,(A^T, B^T)$ holds ("T" means transposition), and for nonsingular matrices G and H

$$\mathrm{ind}(GA, GB) = \mathrm{ind}(AH, BH) = \mathrm{ind}(A, B).$$

In [4] the following theorem is proved; since the proof is rather difficult, we will do without it here:

THEOREM 1. *If R and Q are projectors along* $\mathrm{im}(A)$ *and onto* $\ker(A)$ *respectively and the pencil (A, B) is regular, then*

$$\mathrm{ind}(A+RB, B) = \mathrm{ind}(A+BQ, B) = \mathrm{ind}(A, B)-1.$$

In particular, $\mathrm{ind}(A, B) = 1$ *causes $A+RB$ and $A+BQ$ to be nonsingular.*

Remark. Because of

$$\operatorname{ind}(A+RB,\, B) = \operatorname{ind}\big((A^T+B^TR^T)^T,\, (B^T)^T\big) = \operatorname{ind}(A^T+B^TQ,\, B^T),$$

where $Q = R^T$ is a projection onto $\ker(A^T)$, we had to verify only one of the relations

$$\operatorname{ind}(A+RB,\, B) = \operatorname{ind}(A,\, B)-1, \qquad \operatorname{ind}(A+BQ,\, B) = \operatorname{ind}(A,\, B)-1.$$

As a corollary of Theorem 1 the following statement holds.

THEOREM 2. *The pencil (A, B) is regular with the index k, iff the sequences $A_0 := A$ and for $i = 0, \dots, k-1$*
R_i projector along $\operatorname{im}(A_i)$, $A_{i+1} := A_i+R_iB$ or
Q_i projector onto $\ker(A_i)$, $A_{i+1} := A_i+BQ_i$ produce a nonsingular matrix A_k.

Proof. We have to prove only the regularity of the pencil, if A_k is nonsingular. Following [1] we derive from the assumption $\det(zA_0+B) = 0$ according to

$$\det(zA_{i+1}+B) = \det(zA_i+zR_iB+B) = \det(I+zR_i)\det(zA_i+B)$$

or

$$\det(zA_{i+1}+B) = \det(zA_i+B)\det(I+zQ_i)$$

a contradiction to $\det(zA_k+B) \not\equiv 0$. ∎

Starting from (1), we obtain by differentiation of $R_0Bx(t) = R_0q(t)$ for a differentiable $R_0q(t)$ the equation $A_1x'(t)+Bx(t) = q_1(t)$ with $q_1(t) := q(t)+(R_0q(t))'$. Continuation of the process yields with $q_{i+1}(t) := q_i(t)+(R_iq_i(t))'$ after $k-1$ steps the equation

$$A_kx'(t)+Bx(t) = q_k(t) \tag{3}$$

provided that all $R_iq_i(t)$ $(i = 1, \dots, k-1)$ are differentiable, too.

THEOREM 3. *If the $R_iq_i(t)$ $(i = 0, \dots, k-1)$ are differentiable and $\operatorname{ind}(A, B) = k$, then the DAE (1) is equivalent to the explicit system*

$$x'(t) = A_k^{-1}(q_k(t)-Bx(t)) \tag{4}$$

with the constraints $R_i(Bx(t_0)-q_i(t_0)) = 0$ $(i = 0, \dots, k-1)$ for the initial conditions $x(t_0) = x_0$.

Proof. We have to show that $A_{i+1}x'(t)+Bx(t) = q_{i+1}(t)$ and $R_i(Bx(t_0)-q_i(t_0)) = 0$ imply $A_ix'(t)+Bx(t) = q_i(t)$. This fact becomes clear by application of the following lemma to $R(t) := R_i$, $y(t) := A_ix'(t)+Bx(t)-q_i(t)$. ∎

LEMMA 4. *Let $R(t)$ be an arbitrary projector function and let $R(t)y(t)$ be differentiable. Then $y(t) \equiv 0$ iff $y(t)+\big(R(t)y(t)\big)' \equiv 0$ and $R(t_0)y(t_0) = 0$.*

Proof. Due to

$$R'(t) = \big(R(t)^2\big)' = R'(t)R(t) + R(t)R'(t)$$

we have

$$R(t)R'(t) = R'(t)\big(I - R(t)\big).$$

Hence,

$$\begin{aligned} 0 &= R(t)\big(y(t) + (R(t)y(t))'\big) \\ &= R(t)y(t) + R(t)y'(t) + R(t)R'(t)y(t) \\ &= R(t)y(t) + R(t)y'(t) + R'(t)\big(I - R(t)\big)y(t) \\ &= \big(R(t)y(t)\big)' + \big(I - R'(t)\big)R(t)y(t). \end{aligned}$$

This is a homogeneous linear ordinary differential equation for $R(t)y(t)$ delivering $R(t)y(t) \equiv 0$ by virtue of $R(t_0)y(t_0) = 0$. Consequently, $\big(R(t)y(t)\big)' \equiv 0$ and $y(t) \equiv 0$. ∎

Index-1-systems are called *transferable*; if $Rq(t)$ is differentiable, they are equivalent to the explicit ordinary differential equations

$$x'(t) = (A + RB)^{-1}\{q(t) + (Rq)'(t) - Bx(t)\} \tag{5}$$

with the restrictions $R\{Bx(t_0) - q(t_0)\} = 0$ for the initial vectors $x(t_0) = x_0$. The differentiability requirement for $Rq(t)$ can be avoided, if the index reduction by means of the projector Q is used. For $P := I - Q$, $u(t) := Px(t)$, $v(t) := Qx(t)$ we obtain from (1) $(AQ = 0!)$

$$Au'(t) + B\{u(t) + Qv(t)\} = q(t).$$

Because of

$$\begin{aligned} (A + BQ)^{-1}A &= (A + BQ)^{-1}AP = (A + BQ)^{-1}(A + BQ)P = P, \\ (A + BQ)^{-1}BQ &= I - P = Q, \end{aligned}$$

by multiplication with $(A + BQ)^{-1}$

$$Pu'(t) + (A + BQ)^{-1}Bu(t) + Qv(t) = (A + BQ)^{-1}q(t)$$

arises. Hence, we get by transformation with P and Q respectively

$$u' = P^2u' = P(A + BQ)^{-1}\{q(t) - Bu(t)\}, \qquad v = Q(A + BQ)^{-1}\{q(t) - Bu(t)\}.$$

Consequently, the original equation (1) is transferable into the more general state variable system

$$\begin{aligned} u'(t) &= P(A + BQ)^{-1}\{q(t) - Bu(t)\}, \\ x(t) &= Q(A + BQ)^{-1}\{q(t) - Bu(t)\} + u(t). \end{aligned} \tag{6}$$

The system (6) shows that only the component $u(t) = Px(t)$ occurs differentiated; $x(t)$ is not differentiable, if $Q(A+BQ)^{-1}q(t)$ has this property. This fact is very important, because in many practical problems the source function $q(t)$ is piecewise continuous only, for instance an impulse function in electronic circuits. For that reason in [5] a function $x(t)$ is called *solution* of (1), if $x(t)$ and $(Px)'(t)$ are continuous and satisfy the equation

$$A(Px)'(t)+Bx(t) = q(t). \tag{7}$$

The continuation of this transformation process for linear higher index systems with constant coefficients is possible, but unfortunately rather complicated.

There are many methods to compute the projectors R and Q which are necessary for the suggested transformations. By means of a completely pivoted Gauss decomposition we obtain

$$S_1AS_2 = \begin{pmatrix} L_1 & 0 \\ L_2 & 0 \end{pmatrix} \begin{pmatrix} U_1 & U_2 \\ 0 & 0 \end{pmatrix} \tag{8}$$

with nonsingular triangular matrices L_1 and U_1. S_1 and S_2 are permutation matrices, i.e. $S_i^{-1} = S_i^T$ holds $(i = 1, 2)$. Obviously the matrices $\left(\begin{smallmatrix} 0 & 0 \\ M & I \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 0 & W \\ 0 & I \end{smallmatrix}\right)$ with $M := -L_2L_1^{-1}$ and $W := -U_1^{-1}U_2$ are projectors along $\operatorname{im}(S_1AS_2)$ and onto $\ker(S_1AS_2)$ respectively. Consequently, we may choose

$$R = S_1^T\begin{pmatrix} 0 & 0 \\ M & I \end{pmatrix}S_1, \qquad Q = S_2\begin{pmatrix} 0 & W \\ 0 & I \end{pmatrix}S_2^T.$$

The Householder transformation is very convenient for the computation of Q. Since Q is invariant with respect to scaling of A we apply the decomposition with column pivoting to the scaled matrix $\tilde{A} = \operatorname{diag}(g_1, \dots, g_m)A$ (g_i = scaling factors). We obtain

$$\tilde{A}S = H\begin{pmatrix} U_1 & U_2 \\ 0 & 0 \end{pmatrix},$$

where S is a permutation matrix, H an orthogonal matrix and U_1 is a nonsingular triangular matrix. Again we get

$$Q = S\begin{pmatrix} 0 & W \\ 0 & I \end{pmatrix}S^T \quad \text{with} \quad W := -U_1^{-1}U_2.$$

Now we consider the linear time-dependent problem

$$A(t)x'(t)+B(t)x(t) = q(t). \tag{9}$$

Then we have for regular pencils $(A(t), B(t))$ the local index $k(t)$. If it is constant, it seems to be reasonable to define $k(t) = k$ as the global index of (9). But simple examples show that the pencil $(A(t), B(t))$ and its index only in the case $\operatorname{ind}(A(t), B(t)) = 1$ characterize the solution behaviour (cf. [5], §1.3). Therefore, Gear and Petzold [3] called (9) to have the *global index* k, if a continuous nonsingular matrix function $E(t)$ and a continuously differentiable nonsingular matrix function $F(t)$ exist, so that scaling of (9) by $E(t)$ and the transformation $x(t) = F(t)y(t)$ lead to the DAE

$$A_0 y'(t) + B_0(t)y(t) = q_0(t),$$

where

$$A_0 := \operatorname{diag}(I_r, J) = E(t)A(t)F(t), \qquad q_0(t) := E(t)q(t),$$

$$B_0(t) := \operatorname{diag}\big(W(t), I_{m-r}\big) = E(t)B(t)F(t) + E(t)A(t)F'(t),$$

hold and J is a constant matrix with $J^{k-1} \neq 0$, $J^k = 0$.

For all constant projectors R_0 along $\operatorname{im}(A_0)$ and Q_0 onto $\ker(A_0)$ we obtain

$$R_0 \operatorname{diag}\big(W(t), I_{m-r}\big) = R_0 \operatorname{diag}(0, I_{m-r}),$$

$$\operatorname{diag}\big(W(t), I_{m-r}\big) Q_0 = \operatorname{diag}(0, I_{m-r}) Q_0.$$

Therefore the matrices

$$A_1 := \operatorname{diag}(I_r, J) + R_0 \operatorname{diag}\big(W(t), I_{m-r}\big),$$

$$A_1 := \operatorname{diag}(I_r, J) + \operatorname{diag}\big(W(t), I_{m-r}\big) Q_0$$

are constant again. Consequently, the global index k means that the sequences

(11) $\quad A_{i+1} := A_i + R_i \operatorname{diag}\big(W(t), I_{m-r}\big)$ or $A_i := A_{i+1} + \operatorname{diag}\big(W(t), I_{m-r}\big) Q_i$

with arbitrary constant projectors R_i along $\operatorname{im}(A_i)$ and Q_i onto $\ker(A_i)$ are ending with nonsingular matrices A_k. Then we obtain from $A_i y'(t) + B_0(t)y(t) = q_i(t)$ by multiplication with R_i

$$R_i B_0(t) y'(t) = \big(R_i q_i(t)\big)',$$

since $R_i B_0(t) y(t) = R_i q_i(t)$ with a constant matrix $R_i B_0(t)$. Consequently,

$$A_{i+1} y'(t) + B_0(t) y(t) = q_{i+1}(t) \tag{12}$$

with $A_{i+1} := A_i + R_i B_0(t)$, $q_{i+1}(t) := q_i(t) + \big(R_i q_i(t)\big)'$.

Finally, if we suppose the differentiability of all $R_i q_i(t)$ for $i = 0, \ldots, k-1$, equation (9) is equivalent to the explicit ordinary differential equation

$$y'(t) = A_k^{-1}\{q_k(t) - B_0(t)y(t)\} \tag{13}$$

with the restrictions

$$R_i\{B_0(t_0)y(t_0) - q_i(t_0)\} = 0 \qquad (i = 0, \ldots, k-1) \tag{14}$$

for the initial vector $y(t_0)$.

We suggest a more general definition of the global index k, which is applicable to nonlinear problems, too. Let us consider the DAE

$$f\big(x'(t), x(t), t\big) = q(t) \tag{15}$$

with $f(y, \dot{x}, t)$: $R^m \times R^m \times [t_0, T] \to R^m$. We assume f to have Lipschitz-continuous Jacobians $F_y(y, x, t)$, $F_x(y, x, t)$ and $f_t'(y, x, t)$. If $q(t)$ is differentiable too, then we can write

$$F_y\big(y(t), x(t), t\big)y'(t) + F_x\big(y(t), x(t), t\big)y(t) + f_t'\big(y(t), x(t), t\big) = q'(t)$$
$$x'(t) - y(t) = 0.$$

So we obtain a system

$$A\big(z(t), t\big)z'(t) + g\big(z(t), t\big) = 0 \tag{16}$$

with

$$z(t) = \begin{pmatrix} y(t) \\ x(t) \end{pmatrix}, \quad A(z, t) := \begin{pmatrix} F_y(y, x, t) & 0 \\ 0 & I \end{pmatrix},$$
$$g(z, t) := \begin{pmatrix} F_x(y, x, t)y + f_t'(y, x, t) - q'(t) \\ -y \end{pmatrix}.$$

For the equation (16) we perform sequences $\{A_i\}_{i=0}^k$, $\{g_i\}_{i=0}^k$ by the definitions $A_0(z, t) := A(z, t)$, $g_0(z, t) := g(z, t)$ and for $i = 0, \ldots, k-1$: $R_i(z, t)$ projector along $\operatorname{im}\big(A_i(z, t)\big)$

$$\begin{aligned} A_{i+1}(z, t) &:= A_i(z, t) + \frac{\partial}{\partial z}\{R_i(z, t)g_i(z, t)\}, \\ g_{i+1}(z, t) &:= g_i(z, t) + \frac{\partial}{\partial t}\{R_i(z, t)g_i(z, t)\}. \end{aligned} \tag{17}$$

provided that the occurring derivatives exist and the rank of each $A_{i+1}(z, t)$ is constant.

DEFINITION. We call k *global index* of (16), if k is the smallest integer for which a sequence (17) ending with a nonsingular $A_k(z, t)$ exists.

With this definition the following theorem holds.

THEOREM 5. *If* (16) *has the global index k, then it is equivalent to the explicit ordinary differential equation*

$$z'(t) = -A_k\big(z(t), t\big)^{-1} g_k\big(z(t), t\big) \tag{18}$$

with the initial restrictions

$$R_i\big(z(t_0), t_0\big)g_i\big(z(t_0), t_0\big) = 0 \quad (i = 0, \ldots, k-1).$$

Proof. The statement follows immediately from the application of Lemma 4 to the functions

$$y(t) := A_i\big(z(t), t\big)z'(t) + g_i\big(z(t), t\big), \quad R(t) := R_i\big(z(t), t\big),$$

because of

$$y(t)+\big(R(t)y(t)\big)' = A_{i+1}\big(z(t),\, t\big)z'(t)+q_{i+1}\big(z(t),\, t\big)$$

using

$$\big(R(t)y(t)\big)' = \frac{\partial}{\partial z}\big\{R_i\big(z(t),\, t\big)q_i\big(z(t),\, t\big)\big\}z'(t)+\frac{\partial}{\partial t}\big\{R_i\big(z(t),\, t\big)q_i\big(z(t),\, t\big)\big\}.$$

In [4] we have proved, that the suggested definition of the global index generalizes the definition of Gear and Petzold in the case of a differentiable matrix function $E(t)$. ∎

As an example we consider the motion equations of a multibody system

$$\begin{aligned} v'+f(v,\, u,\, w,\, t) &= 0, \\ u'+q(v,\, u,\, t) &= 0, \\ h(u,\, t) &= 0, \end{aligned}$$

with a nonsingular matrix function $Z = h'_u g'_v f'_w$. Then we obtain

$$A_0 = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad R_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{bmatrix}, \quad R_0 q_0 = \begin{bmatrix} 0 \\ 0 \\ h \end{bmatrix},$$

$$A_1 = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & h'_u & 0 \end{bmatrix}, \quad R_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -h'_u & I \end{bmatrix}, \quad R_1 q_1 = \begin{bmatrix} 0 \\ 0 \\ -h'_u g + s_1 \end{bmatrix},$$

$$A_2 = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ -h'_u g'_v & S_1 & 0 \end{bmatrix}, \quad R_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ h'_u g'_v & -S_1 & I \end{bmatrix}, \quad R_2 q_2 = \begin{bmatrix} 0 \\ 0 \\ h'_u g'_v f + s_2 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ S_2 & S_3 & Z \end{bmatrix},$$

i.e. det (A_3) = det $(Z) \neq 0$. Thus we have verified the well-known property of (19) to be of the index 3. In particular, the mathematical pendulum modelled by $u,\ v \in R^2$,

$$u'-v = 0, \qquad Mv'+g\begin{pmatrix} 0 \\ 1 \end{pmatrix}-2\lambda u = 0, \qquad \|u\|^2 = l^2$$

(u = position, v = velocity, g = gravitation constant, λ = Lagrange parameter, l = length) yields $f'_\lambda = 2u$, $g'_v = I_2$, $h'_u = 2u^T$, that is, $Z = (-4l^2) \neq 0$.

The application of the transformation concept using the projectors Q to the equation (15) is practicable only for a constant or a time-variable $\ker(F_y)$; it is impossible for $\ker(F_y)$ depending on x or y. In the latter case the rotation of the solution space $C_N^1 := \{x\colon\ x \in C,\ P(y, x, t)x \in C^1\}$ depends on the solution itself.

Let $Q(t) = I - P(t)$ be a projector onto $N(t) := \ker(F_y(y, x, t))$; then the equation (15) is called transferable (in the neighbourhood of a solution or globally) if

$$G(y, x, t) := F_y(y, x, t) + F_x(y, x, t)Q(t)$$

has a bounded inverse. Then the theorem about implicit functions and Hadamard's theorem give

THEOREM 6. *Each solution $x(t)$ of* (15) *fulfills* (20) *and vice versa:*

$$\begin{aligned} u'(t) &= P'(t)u(t) + P(t)\big(I + P'(t)\big)w\big(u(t), t\big), \\ u(t_0) &= P(t_0)x(t_0), \\ x(t) &= u(t) + Q(t)w\big(u(t), t\big), \end{aligned} \tag{20}$$

where $w(u, t)$ solves uniquely $f\big(w, u + Q(t)w, t\big) = q(t)$.

The proof is given in [5], it is based on the fact that the Jacobian of $f\big(w, u + Q(t)w, t\big)$ with respect to w has a bounded inverse F_w^{-1} by virtue of $F_w(y, x, t) = G(y, x, t)$. As in the linear case with constant coefficients the solution concept is generalized here. A solution is any continuous function $x(t)$, whose component $P(t)x(t)$ is continuously differentiable and which satisfies the equation

$$f\big((Px)'(t) - P'(t)x(t), x(t), t\big) = q(t). \tag{21}$$

Again there are no differentiability requirements for $q(t)$ in this transformation concept. This is a very important advantage of the method, as we remarked above. But unfortunately the continuation of the transformation process for higher index systems is very difficult and only possible under additional restricting assumptions about $f(y, x, t)$ (cf. [6], [7]).

Remark. For quasi-linear DAEs, i.e. for $f(y, x, t) = A(x, t)y + g(x, t)$ and $q(t) = 0$ we obtain

$$G(y, x, t) = A(x, t) + \frac{\partial}{\partial x}\{A(x, t)y + g(x, t)\}Q(t).$$

If $G(y, x, t)$ is nonsingular, the same is valid for

$$\begin{aligned} H(y, x, t) &= A(x, t) + R(x, t)\frac{\partial}{\partial x}\{A(x, t)y + g(x, t)\} \\ &= A(x, t) + \frac{\partial}{\partial x}\{R(x, t)g(x, t)\} - \frac{\partial R}{\partial x}(x, t)\{A(x, t)y + g(x, t)\}. \end{aligned}$$

Because of $A(x, t)y+g(x, t) = 0$ for $y(t) = x'(t)$ we get $H(y, x, t) = A_1(x, t)$, i.e. the transferability implies the index-1-property in our sense.

References

[1] Burmeister, Private communication.
[2] F. R. Gantmacher, *Theory of Matrices*, vol. II, Chelsea, New York 1960.
[3] C. W. Gear, L. R. Petzold, *Differential-algebraic systems and matrix pencils*, in *Matrix pencils*, Lecture Notes in Math. 973, Berlin–Heidelberg–New York 1983, 75–89.
[4] E. Griepentrog, *Transformations of differential-algebraic equations*, in *Numerical Treatment of Differential Equations*, Proceedings of the fourth seminar held in Halle 1987; BSB Teubner-Verlag, Leipzig, to appear.
[5] E. Griepentrog, R. März, *Differential-algebraic equations and their numerical treatment*, Leipzig, BSB Teubner 1986.
[6] R. März, *Index-2 differential-algebraic equations*, Humboldt-Univ., Sektion Mathematik, Preprint Nr. 135, Berlin 1987.
[7] —, *Some new results concerning index-3 differential-algebraic equations*, Humboldt-Univ., Sektion Mathematik, Preprint Nr. 141, Berlin 1987.

Presented to the Semester
Numerical Analysis and Mathematical Modelling
February 25 – May 25, 1987