Non-structural zeros and singular uncertainty in index reduction of dae — examples and discussions

Henrik Tidefelt
Division of Automatic Control
E-mail: tidefelt@isy.liu.se

25th January 2007

Report no.: LiTH-ISY-R-2768

Address:
Department of Electrical Engineering
Linköpings universitet
SE-581 83 Linköping, Sweden

WWW: http://www.control.isy.liu.se
Abstract
When massaging differential-algebraic equations during an index reduction process, one usually keeps track of structural zeros in order to be able to detect when non-differential equations appear. However, in a numerical setting structural zeros may turn out as small numbers, impossible to distinguish from zero when numerical precision is taken into account. Such coefficients are referred to as non-structural zeros. This note aims to provide a better understanding of the ad hoc procedure of replacing those and other small numbers by structural zeros, at critical stages in the index reduction process.

Keywords: Differential-algebraic, Numeric, Index reduction
Non-structural zeros and singular uncertainty in index reduction of DAE — examples and discussions

Henrik Tidefelt
2007-01-25

Abstract

When massaging differential-algebraic equations during an index reduction process, one usually keeps track of structural zeros in order to be able to detect when non-differential equations appear. However, in a numerical setting structural zeros may turn out as small numbers, impossible to distinguish from zero when numerical precision is taken into account. Such coefficients are referred to as non-structural zeros. This note aims to provide a better understanding of the ad hoc procedure of replacing those and other small numbers by structural zeros, at critical stages in the index reduction process.

Keywords: Differential-algebraic, Numeric, Index reduction

1 Introduction

Nonlinear differential-algebraic equations (DAE, hereafter) is the natural outcome of component-based modeling of complex dynamic systems. Often, there is some known structure to the equations, and the long term goal of this development is to be able to deal with those that emerge in quasilinear form,

\[ E(x(t), t) \dot{x}(t) + A(x(t), t) = 0 \]  

In this note, the matrix-values function \( E \) shall be called the leading matrix, while the tuple-valued function \( A \) shall be called the algebraic term.

In theory, index reduction of equations in this form is simple: Manipulate the equations using invertible row operations so that the leading matrix becomes separated into one block which is completely zeroed, and one block with independent rows. Differentiate the discovered algebraic equations, and repeat until the leading matrix gets full rank. As examples of the in-theory ramifications of this description, consider the following list:

- It may be difficult to perform the row reduction in a numerically well-conditioned way.
- The produced equations could involve very big expressions.
- Testing whether an expression is zero is highly non-trivial.
This note focuses on the last of these ramifications. Typical examples in the literature have leading matrices whose rank is determined solely by a zero-pattern. For instance, if some variable does not appear differentiated in any equation, the corresponding column of the leading matrix will be zero. It is then easy to see that this column will remain zero after arbitrarily complex row operations, so if the operations are chosen to create structural zeros in the other columns at some row, it will follow that the whole row is structurally zero. Thus an algebraic equation is revealed, and when differentiating this equation, the presence of variables in the equation determines the zero-pattern of the newly created row in the leading matrix, and so the index reduction may be continued.

Now, consider what happens if a “change of variables and equations” is performed on the original DAE. It is sufficient to consider static linear transforms, represented by full non-singular matrices. Then any zero pattern is generally lost. Another situation when linear dependence between rows in the leading matrix are not visible in a zero pattern, is when a user happens to write down equations that are dependent within numerical precision. It must be emphasized here that numerical precision is often not a mere question of floating point number representation in numerical hardware, but a consequence of uncertainties in estimated model parameters.

It is proposed that a semi-numerical approach is taken to zero-testing, where an expression is taken for being (re-writable to) zero if it evaluates to zero at some trial point. Clearly, a tolerance will have to be used in this test, and showing that a meaningful threshold even exists is the main topic of this note. The analysis below will be restricted to linear time-invariant (LTI, hereafter) DAE, and then the choice of trial point will be of no consequence.

Notation. Let \( \lambda(X) \) denote the eigenvalues of \( X \), and let for instance \( \text{Re}\lambda(X) < 0 \) mean that all eigenvalues have negative real parts. The symbol \( \doteq \) is used to indicate an equality that shall be thought of as an equation.

2 Background

Elimination-differentiation approaches have been in use for a long time. As is often the case in the area of dynamic systems, the essence of the idea is best introduced by looking at linear time-invariant (LTI) systems, while the extension to nonlinearities brings many subtleties to the surface. The linear case was considered in Luenberger [1978], and the algorithm is commonly known as the shuffle algorithm.

While the compact representation of LTI systems make the translation of theory to computer programs rather straight-forward, the implementation of nonlinear theory is not at all as straight-forward. This seems, at least to some part, to be explained by the fact that there are no widespread computer tools for working with the mathematical concepts from differential algebra. A theoretical counterpart of the shuffle algorithm, but applying to general nonlinear DAE, was used in Rouchon et al. [1995]. However, its implementation is nontrivial since it requires a computable representation of the function whose existence is granted by the theorem. For quasilinear DAE, on the other hand, an implicit function can be computed explicitly, and our current interest in these methods owes to this fact. For references to implementation-oriented index reduction of quasilinear DAE along these lines, see for example Visconti [1999] or Steinbrecher [2006].
The very numerically oriented Brenan et al. [1996] contains overviews of work regarding the relation between DAE and stiff ODE, both in relation to backward difference formulas and implicit Runge-Kutta methods.

The method used to understand the problem in this note reminds of the analysis in Kunkel and Mehrmann [2006, chapter 3], where equation and variable transforms are applied to a time-varying linear DAE to obtain a form where solvability and other questions are easily answered. However, it is a method which neither assumes nor makes approximations. The method of linearizing along a solution trajectory in Campbell [1995] comes with an analysis of the error caused by small errors in the initial conditions. This method, with its classic solvability conditions, is pushed further in Campbell and Griepentrog [1995] to make it amenable to numerical implementation.

The problem of non-structural zeros is closely related to model reduction and singular perturbation theory. Model reduction can be done in many ways. The method called residualization removes some of the system dynamics by substituting zero for the derivatives of some variables. This way, differential equations are replaced by algebraic ones. Skogestad and Postlethwaite, 1996 These methods have been extended to LTI DAE. Stykel, 2004 Unfortunately, model reduction techniques cannot readily be applied in the framework of this note, since we are interested in defending a given model reduction rather than finding one with good properties. In singular perturbation theory, a similar reduction can be seen as the limiting system as some dynamics become arbitrarily fast. Kokotović et al., 1986 However, some of the assumptions made in the singular perturbation framework are not always satisfied in the presence of non-structural zeros, and this is a major concern in this note. The connection to model reduction and singular perturbation theory is interesting also for another reason, namely that the classical motivation in those areas is that the underlying system being modeled is singularly perturbed in itself, and one is interested in studying how this can be handled in modeling and model-based techniques. Although that framework is built around ordinary differential equations, the situation is just as likely when DAE are used to model the same systems. It is a goal of this note to highlight the relation between treatment of small numbers in the leading matrix that are due to stiffness in the system being modeled, and the treatment of small numbers that are artifacts of numerical errors and the like.

The following singular perturbation theorem found in Kokotović et al. [1986, theorem 5.1] will be useful. Consider the singularly perturbed LTI ordinary differential equation

\[
\begin{pmatrix}
  \dot{x}(t)
  \\
  \epsilon \dot{z}(t)
\end{pmatrix} =
\begin{pmatrix}
  M_{11} & M_{12} \\
  M_{21} & M_{22}
\end{pmatrix}
\begin{pmatrix}
  x(t) \\
  z(t)
\end{pmatrix}
\begin{pmatrix}
  x(t_0) \\
  z(t_0)
\end{pmatrix} =
\begin{pmatrix}
  x^0 \\
  z^0
\end{pmatrix}
\]  

(2)

where we are interested in small \( \epsilon > 0 \). Define \( M_0 := M_{11} - M_{12} M_{22}^{-1} M_{21} \), denote

\[
\dot{x}_s(t) \quad x_s(t_0) \quad \dot{z}_f(\tau) \quad z_f(\tau_0)
\]  

(3)

the slow model (obtained by setting \( \epsilon := 0 \) and eliminating \( z \) using the thereby obtained non-differential equations), and denote

\[
\dot{z}_f(\tau) = M_{22} \dot{z}_f(\tau) \\
\dot{z}_f(\tau_0) = z^0 + M_{22}^{-1} M_{21} x^0
\]  

(4)

the fast model (which is expressed in the timescale given by \( \epsilon \tau = t - t_0 \)).

**Theorem 1.** If \( \text{Re} \lambda( M_{22} ) < 0 \), there exists an \( \epsilon^* > 0 \) such that, for all \( \epsilon \in (0, \epsilon^*] \), the states of the original system (2) starting from any bounded initial conditions \( x^0 \).
and \( z^0, \| z^0 \| < c_1, \| z^0 \| < c_2 \), where \( c_1 \) and \( c_2 \) are constants independent of \( \epsilon \), are approximated for all finite \( t \geq t_0 \) by

\[
\begin{align*}
x(t) &= x_s(t) + O(\epsilon) \\
z(t) &= -M^{-1} - \frac{1}{22} M x_s(t) + z_1(\tau) + O(\epsilon)
\end{align*}
\]  

(5)

where \( x_s(t) \) and \( z_1(\tau) \) are the respective states of the slow model (3) and the fast model (4).

If also \( \text{Re} \lambda(M_0) < 0 \) then (5) holds for all \( t \in [t_0, \infty) \).

Moreover, the boundary layer correction \( z_1(\tau) \) is significant only during the initial short interval \( [t_0, t_1] \), \( t_1 - t_0 = O(\epsilon \log \epsilon) \), after which

\[
z(t) = -M_{22}^{-1} M_{21} x_s(t) + O(\epsilon)
\]

3 The reduction algorithm

This section will detail the index reduction algorithm mentioned in the introduction. It will be defined for LTI DAE, and is a shuffle algorithm. The restriction to LTI DAE is for the sake of the analysis below. Extending the index reduction algorithm as such to quasilinear systems is almost immediate; it is the analysis of it that becomes more difficult.

Consider the square DAE (square meaning that the number of equations and variables match)

\[
E \dot{x}(t) + A x(t) = 0
\]  

(6)

where \( E \) and \( A \) are no longer considered functions, but constant matrices. To ease the notation, it is assumed that the variables are ordered to suit the presentation.

The algorithm (described below) applies row operations represented by the matrix \( K_0 \) to the equations, maintaining a form,

\[
\begin{align*}
K_0 E &= \begin{pmatrix} E_{11} & E_{12} \\ 0 & \epsilon E_{22} \end{pmatrix} \\
K_0 A &= \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}
\end{align*}
\]  

(7)

where the scalar \( \epsilon \geq 0 \) is chosen such that \( E_{22} \) has element-wise max-norm 1. The matrix \( E_{11} \) is upper triangular and non-singular.

As long as \( E_{22} \) contains elements that are distinct non-zero (taking numerical precision into account), the algorithm proceeds with the row reduction until either \( E_{11} \) spans all of the leading matrix, or the small size of \( \epsilon \) is suspected to indicate a stiffness that needs to be taken care of by model reduction. In the first case, index reduction is complete since the underlying ODE has been revealed. The case \( \epsilon = 0 \) is standard in theoretically oriented contexts; when this happens the purely algebraic equations are differentiated (shuffled), which replaces the zeroed block in the leading matrix by the corresponding rows from the matrix in the algebraic term. The differential index has then been lowered by 1, and the index reduction process can start over from the beginning. The last case, and when no elements are distinct non-zero, will be discussed shortly, with the goal of approximating \( \epsilon \) by zero. Values for \( \epsilon \) admitting such approximation will be referred to as tiny from here on.

Besides the assumptions that are the goal of this note, determining whether \( \epsilon \) is tiny cannot be done looking at \( \epsilon \) alone. Clearly, the threshold must be a function of
the involved matrices (with the exception of $E_{22}$). In particular, note that $\epsilon$ can be made arbitrarily small by scaling of equations or variables. Although answering the tinyness-question is outside the scope of this note, equation and variable scalings need to be specified to some extent for the forthcoming analysis.

In order to define equation and variable scaling without too much inconvenience, one of the assumptions needed in the analysis below is made now:

A1 It is assumed that $A_{22}$ is well-conditioned whenever tinyness of $\epsilon$ is investigated.

This assumption can be relaxed, but that requires a more sophisticated algorithm that will not be detailed here. In particular, this assumption implies that the rows in \((A_{21} A_{22})\)
are non-zero, and it is possible to augment the row reduction scheme with row scalings such that the following property holds:

P1 The norm of the smallest row in \((A_{21} A_{22})\) equals the norm of the biggest row in \((A_{11} A_{12})\), unless zero.

Note that the act of ensuring this property may effectively change the value of $\epsilon$.

If the row reduction stopped because $E_{22}$ contained only elements that are not distinctively non-zero, the problem is ill-posed unless $\epsilon$ may be considered tiny. This relates the problem of lack of numerical precision to the problem of stiff systems. In view of this and the close resemblance to the singular perturbation framework, it is natural to say that row reduction stooped due to singular uncertainty.

Having to decide whether $\epsilon > 0$ may be considered tiny, the algorithm is faced with several alternatives:

- Ignorantly replace the small numbers by structural zeros (thereby making $\epsilon = 0$), and proceed as above.
- Reject the problem with a message of the kind not implemented.
- Make an informed decision whether it is right to replace the small numbers by structural zeros. If so, proceed as above, and if not, reject the problem as hopelessly ill-conditioned.

The note pursues the last of these.

Although not the issue here, one more termination condition should be mentioned for completeness. Besides the error situations and the case of successful completion when the leading matrix has obtained full rank, the algorithm terminates when the differentiation step does not add rank to the leading matrix. (In this case the solution is not well defined.)

Before ending this section, the choice of row reduction algorithm will be discussed briefly. There are several standard algorithms that maintain the upper triangular form of $E_{11}$ with zeros below, for instance Gaussian elimination and QR factorization. Any such algorithm can be combined with the equation and variable scalings prescribed here, although that may break good numerical properties. If extension to quasilinear DAE would not be a concern, using QR factorization methods would be a good choice for their good numerical properties (trying to ensure that the solution to the row reduced equations is close to the solution of the original equations), but with the quasilinear form in mind, Gaussian elimination is proposed.
4 Motivating example

The example of this section is meant to show the need for an algorithm like the one proposed; it will not show how the proposed algorithm is used to solve the problem.

Starting from an index 0 DAE in two variables,
\[
\begin{pmatrix}
1 & 7 \\
1 & 3
\end{pmatrix} \dot{x}(t) + \begin{pmatrix}
3 & 2 \\
2 & 1
\end{pmatrix} x(t) = 0
\]
an index 1 DAE in three variables is formed by making a copy of the second variable. In the leading matrix, the second variable is replaced to 70\% by the new variable.

\[
\begin{pmatrix}
1 & 2.1 & 4.9 \\
1 & 0.9 & 2.1 \\
0 & 0 & 0
\end{pmatrix} \dot{x}(t) + \begin{pmatrix}
3 & 2 & 0 \\
2 & 1 & 0 \\
0 & 1 & -1
\end{pmatrix} x(t) = 0
\]

To perform index reduction of this DAE, it suffices to note that the first two rows of the leading matrix are independent, and as the last row is completely zeroed, the last equation is differentiated. This leads to

\[
\begin{pmatrix}
1 & 0.21 & 7 \cdot 0.49 \\
1 & 0.09 & 0.21 \\
0 & 1 & -1
\end{pmatrix} \dot{x}(t) + \begin{pmatrix}
3 & 2 & 0 \\
2 & 1 & 0 \\
0 & 0 & 0
\end{pmatrix} x(t) = 0
\]

Now, perform a well-conditioned change of equations given by the matrix:

\[
T := 4 \cdot \begin{pmatrix}
2 & -9 & 0 \\
8 & -5 & 3 \\
1 & -5 & 7
\end{pmatrix}^{-1}
\]

It is natural to expect that this should not make a big difference to the difficulty in solving the DAE via reduction to index 0, but when the computation is performed on a computer, the picture is not quite as clear. The new DAE has the matrices \( T^{-1} E \) and \( T^{-1} A \). By computing a QR factorization (using standard computer software) of the leading matrix, a structurally upper triangular leading matrix was obtained together with an orthogonal matrix \( Q_1 \) associated with this form. The corresponding matrix of the algebraic term is computed by multiplication by \( Q_1 \) from the left. This leads to

\[
\begin{pmatrix}
-0.62 & -0.95 & -2.2 \\
0 & 0.62 & 1.4 \\
0 & 0 & 3.4 \cdot 10^{-16}
\end{pmatrix} \dot{x}(t)
\]

\[
+ \begin{pmatrix}
-1.6 & -0.53 & -0.41 \\
0.51 & 0.56 & -0.048 \\
-7.2 \cdot 10^{-17} & 0.46 & -0.46
\end{pmatrix} x(t) = 0
\]

Although looking like an implicit ODE, this view is unacceptable for two reasons. First, the system of equations is extremely stiff. (Even worse, the stiff mode happens to be unstable this time, not at all like the original system.) Second, considering numerical precision in hardware, it would not make sense to compute a solution that depends so critically on a coefficient that is not distinctly non-zero.

The ad hoc solution to this problem is to replace the small coefficient in the leading matrix by zero, and then proceed as usual, but suppose \textit{ad hoc} is not good enough. How can one then determine if 3.4 \cdot 10^{-16} is tiny, or just looks tiny due to equation and variable scalings? What is the theoretical excuse for the replacement of small numbers by zeros? What assumptions have to be made?
5 Analysis

Consider the form (7), when the algorithm is faced with the question whether $\epsilon = 0$ is a good approximation. The analysis of this section will produce additional assumptions (the first one being A1) that will make the approximation errors $O(\epsilon)$. Although not providing a way to compute a threshold, this shows that it is meaningful to speak of $\epsilon$ small enough. There will be assumptions both regarding robust features of the equations and those reflecting the intended meaning of the equations. Thus the assumptions will not be possible to verify in a purely mathematical setting, but only when the user has some idea of what object the equations are a model of, and the user also has some knowledge about how the underlying object could possibly behave. Let these assumptions (to be determined) be referred to as the assumptions about the underlying object.

Introduce an $\epsilon$-uncertainty in the non-structural elements of $E_{22}$, so that $E_{22} = 0$ is considered a possible model for the underlying object. Then the DAE is really a family of DAE, obtained by considering the possible perturbations while excluding any equations that do not conform to the assumptions about the underlying object.

It is clear that the proposed index reduction scheme is useless in a numeric setting if the solutions to the family of equations are not close. If it can be shown that the solutions approach a limit as the degree of uncertainty gets smaller, then taking any of the solutions as the solution to the original equations should be considered just as good. In particular, the solutions should approach that obtained by setting to zero any numbers that cannot be distinguished from zero under the given uncertainty.

Two assumptions about the underlying object are added from start.

A2 It is assumed that the given DAE is meant to be of differential index 1.

A3 It is assumed that for every very fast mode of the underlying ODE, the duration of the boundary layer is bounded by a common time $t_1$, which is no further into the future than that it is acceptable to have potentially large transient errors until then.

Admittedly, A3 is an odd bird, but the discussion of this is deferred until section 7.

For simplicity the study is restricted to finite times. This way, it will not be necessary to assume stability of the slow model.

5.1 Pointwise approximation

In this section, the analysis is concerned with perturbations that only differ in magnitude, not in direction. It will be shown that the deviations in the solution decay at the same rate as the magnitude of the perturbation.

The goal will be to rewrite the DAE in standard singular perturbation form, where the approximation in setting $\epsilon := 0$ is reasonably well understood, and then revert the rewriting process to see how the approximation relates to the original formulation. Rewriting of the equations will be performed in two ways, namely equation transforms (or equivalently, row operations on the matrices), and variable transforms (or equivalently, column operations on the matrices). Since setting $\epsilon := 0$ can be seen as a row operation on the leading matrix, this will interact with the other equation transforms when reverting the rewriting process. Hence, the equation transforms must be well conditioned in certain ways in order to end up not too far from the original equations. On the other hand, the variable transforms are unaffected by the row operation $\epsilon := 0$,
and need not be well-conditioned in order for the reversed rewriting process to yield useful results.

The form (7) requires that $E_{22}$ have norm 1, but it may be singular. In particular, it may be singular due to structural properties (a variable not appearing in any equation, or some equation being non-differential), but it can also be singular due to numerical coincidence.

Since we are now concerned with a particular instance of the family, it makes sense to compute the QR factorization $Q_1 E_{22} = \begin{pmatrix} R_0 & 0 \end{pmatrix}$. Extend the orthogonal matrix such that it can be applied to all equations:

$$Q_1 K_0 E = \begin{pmatrix} E_{11} & E_{12} \\ 0 & \epsilon & R_0 \\ 0 & 0 \end{pmatrix}$$

(8)

The corresponding matrix in the algebraic term,

$$Q_1 K_0 A = \begin{pmatrix} A_{11} & A_{12} \\ A'_{21} & A'_{22} \\ A_{31} & A_{32} \end{pmatrix}$$

(9)

must have full rank in the last block row, or the equations would not have a well-defined solution. However, the analysis benefits from the stronger $A_1$ which ensures that (the square)

$$\begin{pmatrix} A'_{22} \\ A'_{32} \end{pmatrix} = Q_1 A_{22}$$

is well-conditioned in itself. Let $A'_{22} Q_2^T = \begin{pmatrix} 0 & R_1 \end{pmatrix}$ be another QR factorization. Then, the change of variables $x = K^T_1 y$ with

$$K^T_1 = \begin{pmatrix} I & 0 \\ -A_{22}^{-1} A_{21} & Q_2^T \end{pmatrix}$$

is well defined (although the condition number may be large) and brings the equations in the form

$$Q_1 K_0 E K^T_1 = \begin{pmatrix} E''_{11} & E''_{12} & E''_{13} \\ \epsilon E''_{21} & \epsilon E''_{22} & \epsilon E''_{23} \\ 0 & 0 & 0 \end{pmatrix}$$

$$Q_1 K_0 A K^T_1 = \begin{pmatrix} A''_{11} & A''_{12} & A''_{13} \\ 0 & A''_{22} & A''_{23} \\ 0 & 0 & R_1 \end{pmatrix}$$

(10)

where

$$E''_{11} = E_{11} - E_{12} A_{22}^{-1} A_{21}$$

(11)

Since this is a DAE of index 1 (by A2), differentiating the last equation must yield a full-rank leading matrix, and by the block triangular form of that matrix, it may be concluded that

$$\begin{pmatrix} E''_{11} \\ \epsilon E''_{21} \end{pmatrix}$$
is non-singular. At the same time, $R_1 y_3 = 0$ implies $y_3 = 0$, and hence $y_3$ together with the last group of equations can be removed without changing the solution for $y_1$ and $y_2$. Partition $\bar{Q}_1$ and $Q_2$ to reflect this. The equations given by the following matrices then determine $y_1$ and $y_2$.

$$
\begin{align*}
\bar{Q}_{1,1} K_0 E \bar{Q}_{2,1}^T & = \begin{pmatrix} E''_{11} & E''_{12} \\ \epsilon E_{21} & \epsilon E_{22} \end{pmatrix} \\
\bar{Q}_{1,1} K_0 A \bar{Q}_{2,1}^T & = \begin{pmatrix} A''_{11} & A''_{12} \\ 0 & A''_{22} \end{pmatrix}
\end{align*}
$$

where the leading matrix is now known to be non-singular. Even better, since $E''_{11}$ shows that $E''_{22}$ does not depend on the perturbed $E''_{22}$, it makes sense to make an assumption regarding it.

**A4** It is assumed that $E''_{11} - E''_{12} A''_{22}$ is well-conditioned with eigenvalues much bigger than $\epsilon$.

This way, the equations can soon be turned into the standard singular perturbation form (2). By well-conditioned row operations represented by

$$
K_2 = \begin{pmatrix} I & 0 \\ -\epsilon E''_{21} & I \end{pmatrix}
$$

and a well defined change of variables, $y = K_3^T z$, with

$$
K_3^T = \begin{pmatrix} I & -E''_{21} & E''_{12} \\ 0 & 0 & I \end{pmatrix}
$$

we reach

$$
\begin{align*}
K_2 \bar{Q}_{1,1} K_0 E \bar{Q}_{2,1}^T K_3^T & = \begin{pmatrix} E''_{11} & 0 \\ 0 & \epsilon E''_{22} \end{pmatrix} \\
K_2 \bar{Q}_{1,1} K_0 A \bar{Q}_{2,1}^T K_3^T & = \begin{pmatrix} A''_{11} & A''_{12} \\ A''_{21} & A''_{22} \end{pmatrix}
\end{align*}
$$

where the leading matrix is still non-singular. Hence, the standard singular perturbation form is

$$
\begin{pmatrix}
\dot{z}_1(t) \\
\epsilon \dot{z}_2(t)
\end{pmatrix} = \begin{pmatrix} E''_{11} & A''_{12} \\ E''_{21} A''_{11} & E''_{22} A''_{12} \end{pmatrix} \begin{pmatrix} z_1(t) \\
z_2(t) \end{pmatrix}
$$

(12)

Here, A3 grants that $E''_{21} A''_{22}$ is Hurwitz so that theorem 1 applies. Hence, after the short boundary layer, the solution is close to the solution of

$$
\begin{pmatrix}
\dot{z}_1(t) \\
0
\end{pmatrix} = \begin{pmatrix} E''_{11} & A''_{12} \\ E''_{22} A''_{11} & E''_{22} A''_{12} \end{pmatrix} \begin{pmatrix} z_1(t) \\
z_2(t) \end{pmatrix}
$$

or equivalently, the DAE in $z$ with matrices

$$
\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} K_2 \bar{Q}_{1,1} K_0 E \bar{Q}_{2,1}^T K_3^T \\
K_2 \bar{Q}_{1,1} K_0 A \bar{Q}_{2,1}^T K_3^T
$$

Now, undoing the change of variables from $y$ to $z$, reintroducing $y_3$, and undoing the change of variables from $x$ to $y$, we find that $x$ is approximately given by the DAE with
matrices
\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & I
\end{pmatrix}
\begin{pmatrix}
1 & 0 & 0 \\
0 & Q_{1,11} & Q_{1,12} \\
0 & Q_{1,21} & Q_{1,22}
\end{pmatrix}
K_0 E
\]
\[
\begin{pmatrix}
\mathcal{I} & -\epsilon E_{11}'' & 0 & 0 \\
0 & \mathcal{I} & 0 & 0 \\
0 & 0 & \mathcal{I}
\end{pmatrix}
\begin{pmatrix}
\mathcal{I} & 0 & 0 \\
0 & Q_{1,11} & Q_{1,12} \\
0 & Q_{1,21} & Q_{1,22}
\end{pmatrix}
K_0 A
\]

Using (8) and (9), these matrices evaluate to
\[
\begin{pmatrix}
E_{11} & E_{12} \\
0 & 0 \\
0 & 0
\end{pmatrix}
\begin{pmatrix}
A_{11} & A_{12} \\
A_{21}' - \epsilon E_{21}'' E_{11}'' - A_{11} & A_{22}' - \epsilon E_{21}'' E_{11}'' - A_{12}
\end{pmatrix}
\]

It is seen that the row operation \(K_2\) is not to be undone, only \(Q_1\) is. This does not change the leading matrix, while the lower part of the matrix in the algebraic term can be written
\[
(A_{21} & A_{22}) - \epsilon \left( \begin{pmatrix}
\mathcal{I} \\
0
\end{pmatrix} Q_1 E_{11}'' E_{11}'' - A_{11} & A_{12}
\end{pmatrix}
\]

By A4 and P1, the second of these terms tends is small and tends to zero at the rate of \(\epsilon\). The problem of understanding a tiny perturbation in the leading matrix has thus been shifted to understanding a small perturbation in the matrix of the algebraic term. Since \(A_{22}\) is well-conditioned by \(A_1\), also the slightly perturbed matrix in the final form will be well-conditioned.

Since the DAE now readily can be written as an LTI ODE after elimination of \(x_2\), and the associated matrix is like \(E''_{11}\) (see A4) but with a small perturbation, we are close to a conclusion. How the solution depends on the size of the perturbation is related to the perturbed matrix itself in a non-trivial manner, as is shown in Van Loan [1977]; the statements about perturbations in the matrix exponential in this note are all from this source. To begin with, at any fixed time of the solution, the relative error due to the perturbation vanishes at the same rate as the size of the perturbation, as the size of the perturbation tends to zero. It may be surprising then, that \(E''_{11}\) being Hurwitz is not enough to ensure that the absolute perturbation error tends to zero with time, even for very small perturbations. However, such conditions exist, and one is to be presented shortly. It makes use of the logarithmic norm of a matrix, see for instance Ström [1975] for a discussion. The reason this condition is not made an assumption here is that, by looking at the solution during a fixed time interval, the decay of the perturbation error over time is not really needed. In addition, the condition to be presented here is just one of many conditions, and it recommended that the reader refer to Van Loan [1977] rather than checking only this: If the logarithmic norm of \(E''_{11}\) is negative, in addition to \(E''_{11}\) being Hurwitz, then the absolute error due to sufficiently small perturbations will eventually tend to zero as time tends to infinity.

### 5.2 Uniform approximation

In this section, it will be established that the size of the deviations in the solution are \(O(\epsilon)\) maximizing both over time and the uncertainty-perturbations of \(E_{22}\). While the
time dimension was “handled” by only considering a finite time interval, it remains to ensure that the results are not affected by the exact content of $E_{22}$; knowing that the norm is $1$ must be enough, so that $\epsilon$ alone defines the size of the perturbation.

Clearly, the analysis takes different paths depending on the rank of $E_{22}$, but this causes no problems since there are only finitely many values of the rank to consider. Hence, it is sufficient to consider the rank as given.

Note that much of the dependency of (12) on $E_{21}^\prime$ is via the unitary matrix $Q_1$ that represents an equation transformation. However, it suffices without loss of generalization to consider the case $Q_1 = I$, since if closeness can be shown with “$Q_1$ incorporated in $K_0^\prime$” (leading to $Q_1 = I$), any general case is obtained by just undoing the equation operations $Q_1$ once the approximation properties has been established. From here on, consider the case $Q_1 = I$.

The perturbation $E_{22}$ enters the analysis above in two places, first when setting $\epsilon := 0$, and then when estimating the error caused by the non-singular perturbation at the end of the analysis.

Beginning with the latter, note that the smallness of the second term in (13) requires that $E_{21}^\prime$ be bounded under the perturbations (this is the only part of the expression that depends on $E_{22}$). In (10), it can be seen that the dependence is linear, and the norm of this relation does not depend on $E_{22}$ since $Q_1$ is unitary, and $K_1^\prime$ does not depend on $E_{22}$ at all. From this, boundedness follows.

Turning to the singular perturbation part, it needs to be investigated how the form (12) depends on $E_{22}$, and then theorem 1 needs to be refined so that the constants of the “$O(\epsilon)$” can be related to the matrices.

It is readily seen that $Q_2$ and $K_1$ do not depend on $E_{22}$. Hence, $E_{12}^{\prime^\prime}$ (in addition to $E_{11}^{\prime^\prime}$) have no dependency, $E_{21}^{\prime^\prime}$ and $E_{22}^{\prime^\prime}$ depend only via the column operation $K_1$. Since $K_2$ does not change the first group of equations, it follows that the first group of equations in (12) have no dependency. To see how the second group of equations depend, note that the norm of

$$E_{22}^{\prime^\prime^\prime} = E_{22}^{\prime^\prime} - E_{21}^{\prime^\prime} E_{11}^{\prime^\prime-1} E_{12}^{\prime^\prime}$$

can be bounded independently of $E_{22}$. On the side of the algebraic term, the only dependency on $E_{22}$ is via the equation operation $K_2$. However, by P1, A4, and the recently used norm bound on $E_{21}^{\prime}$, it follows that $(A_{12}^{\prime^\prime}, A_{22}^{\prime^\prime})$ is close to $(A_{12}^{\prime}, A_{22}^{\prime})$.

Hence, the theorem 1 needs to be refined so that it is seen how the constant of the “$O(\epsilon)$” depends on $M_{21}$ and $M_{22}$. Inspection of the proof of this theorem reveals that the two matrices appear as $M_{22}^{-1} M_{21}$ in most places, which in the present case is just a small perturbation of $A_{22}^{\prime^\prime-1} A_{21}^{\prime^\prime}$, where the smallness can be bounded independently of $E_{22}$. This leaves us with the investigation of the remaining occurrences of $M_{21}$ and $M_{22}$, and inspection shows that this amounts to places where $M_{22}$ is the matrix being perturbed in a matrix exponential. At one place, this is related to the duration of the bounding layer, which needs no further inspection by A3. In the one place remaining, the matrix exponential is multiplied by $\epsilon$, but evolves in the fast time-scale. Hence, if the perturbed $M_{22}$ would have unstable poles, the “constant” of the associated “$O(\epsilon)$” would grow without bound as $\epsilon$ tends to zero. On the other hand, it is sufficient that $M_{22}$ be Hurwitz, which follows by A3.

This shows that the assumptions already listed before this section are enough to obtain uniform convergence with respect to $E_{22}$.
6 Examples

This section contains examples that illustrate some of the assumptions and the need to understand how to find a threshold for $\epsilon$. Let $?\Rightarrow$ be used to indicate a matrix which has no distinctively non-zero elements. The DAE are LTI in a variable denoted $x$.

6.1 A computed example

In this section, the results are illustrated using computer simulations with randomly generated perturbations for $E_{22}$. Two algebraic terms were used, given by selecting $\delta$ from $\{1, 10^{-2}\}$ in the pattern

$$A = \begin{pmatrix}
0.29 & 0.17 & 0.046 \\
0.34 & \delta 0.66 & \delta 0.66 \\
0.87 & \delta 0.83 & \delta 0.14
\end{pmatrix}$$

and then scaling the rows according to $P\Rightarrow$. Let the four numbers $(?_{11}, ?_{12}, ?_{21}, ?_{22})$ be generated by first taking four independent samples from a uniform distribution centered at 0, and then scaling to make the biggest number have the magnitude 1. Then $E$ was generated according to the pattern

$$E := \begin{pmatrix}
1 & 1 & 1 \\
0 & ?_{11} & ?_{12} \\
0 & ?_{21} & ?_{22}
\end{pmatrix}$$

for some chosen $\epsilon > 0$. The example is chosen such that $\epsilon = 0$ yields a stable slow system. Thus the perturbations of interest are those that make all modes of the stiff system stable. The initial conditions are chosen with $x_1(0) = 1$ and consistent with $\epsilon = 0$.

Simulation results are shown in figure 1. By choosing a threshold for $\epsilon$ based on visual appearance, the threshold can be related to $\delta$. Finding that $1 \cdot 10^{-3}$ and $1 \cdot 10^{-5}$ could be reasonable choices for $\delta$ being 1 and $10^{-2}$, respectively, it is tempting to conclude that it would be wise to base the scaling of the last two rows on $A_{22}$ alone.

6.2 Coping without $A1$

The following two DAE represent different trivial cases when $A1$ fails.

First, let $\epsilon$ be considered tiny and consider the 

$$\begin{pmatrix}
1 & 0 & 1 \\
0 & \epsilon ? & \epsilon ? \\
0 & \epsilon ? & \epsilon ?
\end{pmatrix}$$

The equation contains a completely useless equation and is clearly ill-posed. If input data is assumed to be well-posed, this situation should not happen.

Second, keeping the same $A_{22}$ but adding structural zeros in the leading matrix and a non-zero element in $A_{21}$,

$$\begin{pmatrix}
1 & 0 & 1 \\
0 & \epsilon ? & \epsilon ? \\
0 & 0 & 0
\end{pmatrix}$$

it is found that $x_1 = 0$, and it follows that $\dot{x}_3 = 0$. Hence, $x_2$ is a stiff variable that quickly approaches zero.
Figure 1. Solutions for $x_1$ obtained by generating 50 random perturbations of given magnitudes. Details are given in the text. Left: $A$ defined by $\delta = 1$. Right: $A$ defined by $\delta = 10^{-2}$. Top: $\epsilon = 1 \cdot 10^{-1}$. Middle: $\epsilon = 1 \cdot 10^{-3}$. Bottom: $\epsilon = 1 \cdot 10^{-5}$.

6.3 Breaking A2

When A2 fails, it will generally not be possible to reach the standard singular perturbation form (2). Consider the DAE with matrices

\[
\begin{pmatrix}
1 & 0 & 1 \\
0 & \epsilon & ? \\
0 & \epsilon & ?
\end{pmatrix}
\begin{pmatrix}
1 \\
1 \\
0 \\
0
\end{pmatrix}
\]

Note that A1 implies that A4 must fail completely here.

Consider the case when the perturbed matrix has full rank. The algorithm then applies a variable transform to make the leading matrix block diagonal:

\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & \epsilon & ? \\
0 & \epsilon & ?
\end{pmatrix}
\begin{pmatrix}
1 & 1 & -1 \\
1 & 1 & 0 \\
0 & 0 & 0
\end{pmatrix}
\]

Inverting the perturbation is no good this time, since

\[
\begin{pmatrix}
? & ? \\
? & ? \\
? & ?
\end{pmatrix}^{-1}
\begin{pmatrix}
1 & 0 \\
0 & 0
\end{pmatrix}
\]

is not full-rank for any perturbation, and hence singular perturbation theory cannot be applied here immediately. However, since the two rows are dependent it is possible to proceed, although doing so is not included in the algorithm proposed here.

7 Discussion

Before concluding this paper, it is motivated to include a brief discussion of the analysis and its results.
7.1 Nature of the assumptions

$A_1$ and $A_4$ can easily be checked during index reduction. If they do not hold, the proposed index reduction scheme cannot be used to solve the DAE, and an implementation may choose to report this as an error. $A_3$ can often be validated using, for example, physical insight about the object that the DAE is supposed to describe. However, the judgment will become more difficult if $\epsilon$ is just small, but not tiny, since that is the distinction between unreasonable and reasonable stiffness of the object being modeled. $A_2$ typically requires a leap of faith, but it may also be the case that it is known that the equations have been obtained by equation and variable transforms applied to equations which are seen to be index 1 from their zero-patterns. Note that $A_2$ is very similar to $A_4$.

7.2 Applying the results

In Kokotović et al. [1986], it is argued that theorem 1 is often used to motivate the ad hoc method of just setting $\epsilon := 0$ because that seems like a reasonable thing to do. However, the cautious user will also need to define smallness of $\epsilon$ in terms of the allowable errors in the solution, and this will require a refined formulation of theorem 1 where any $\mathcal{O}(\epsilon)$ expressions are explicitly bounded. Without this, setting $\epsilon := 0$ is still an ad hoc procedure, and it must be emphasized that this remark applies just as much to the DAE index reduction scheme presented here.

In case the threshold for $\epsilon$ is well above the threshold obtained by tracking numerical precision, the user could choose to first use a lower threshold to do away with the really tiny numbers first, since $A_3$ can then be motivated using physical insight. Then, when tiny numbers are not so tiny any more, and physical insight can no longer support $A_3$, the assumption must instead be checked by computing $A_2''$ and verify that its poles are sufficiently far into the complex left half plane.

7.3 Formulating a theorem

As the analysis provides an $\mathcal{O}(\epsilon)$ result, formulating a theorem like theorem 1 seems within reach. This would require that $A_3$ be formulated more carefully in terms of eigenvalues of the fast modes of the underlying object. In return, $A_1$ and $A_4$ could be relaxed, not referring to $\epsilon$, and referring to non-singular instead of well-conditioned.

7.4 Extensions

Three directions for future research seem particularly important. One being extension to higher indices. Another being that these results be extended to more general classes of DAE, in particular time-varying, and then the quasilinear form. The third being that the analysis be refined so that the ad hoc tuning of the smallness threshold for $\epsilon$ can be replaced by a threshold that derive from error tolerances that a user can understand. Investigating non-square systems would also be of interest.

8 Conclusion

By analyzing how an ill-conditioned LTI DAE can be related to a standard singular perturbation formulation, a set of four assumptions has been established such that the “natural treatment” of such equations can be understood to a degree similar to that
which is typically established for standard singular perturbation problems. This means that the proposed index reduction procedure still remains an ad hoc procedure, albeit one which is better understood by taking the presented assumptions into account.

References


