**BIT manuscript No.**
(will be inserted by the editor)

# Conversion Methods, Block Triangularization, and Structural Analysis of Differential-Algebraic Equation Systems

**Guangning Tan** · **Nedialko S. Nedialkov** · **John D. Pryce**

Received: date / Accepted: date

**Abstract** In a previous article, the authors developed two conversion methods to improve the $\Sigma$-method for structural analysis (SA) of differential-algebraic equations (DAEs). These methods reformulate a DAE on which the $\Sigma$-method fails into an equivalent problem on which this SA is more likely to succeed with a generically nonsingular Jacobian. The basic version of these methods processes the DAE as a whole. This article presents the block version that exploits block triangularization of a DAE. Using a block triangular form of a Jacobian sparsity pattern, we identify which diagonal blocks of the Jacobian are identically singular and then perform a conversion on each such block. This approach improves the efficiency of finding a suitable conversion for fixing SA's failures. All of our conversion methods can be implemented in a computer algebra system so that every conversion can be automated.

**Keywords** differential-algebraic equations · structural analysis · block triangular form · modeling · symbolic computation

**Mathematics Subject Classification (2000)** 34A09 · 65L80 · 41A58 · 68W30

## 1 Introduction.

This article is a continuation of [13], in which we presented two conversion methods for improving the $\Sigma$-method [9] for structural analysis (SA) of DAEs. When this SA fails on a DAE with an identically singular (but structurally nonsingular) System

G. Tan
School of Computational Science and Engineering, McMaster University,
1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada, E-mail: tang4@mcmaster.ca

N. S. Nedialkov
Department of Computing and Software, McMaster University,
1280 Main Street West, , Hamilton, Ontario L8S 4K1, Canada, E-mail: nedialk@mcmaster.ca

J. D. Pryce
School of Mathematics, Cardiff University,
Senghennydd Road, Cardiff CF24 4AG, Wales, UK., E-mail: prycejd1@cardiff.ac.uk

Jacobian, our conversion methods reformulate the DAE into an equivalent problem on which the SA is more likely to succeed with a generically nonsingular System Jacobian [12, 13].

These two conversion methods are the linear combination (LC) method and the expression substitution (ES) method. The former is based on replacing an existing equation by a linear combination of some equations and derivatives of them. The latter is based on replacing some existing derivatives[1] by expressions that contain newly introduced variables and derivatives of them. In the ES method, the equations that prescribe such replacements are also appended to the original DAE, so the resulting system is an enlarged one. The main result of a conversion using either method is a strict decrease in the value of the signature matrix [13]. Based on our experience, we conjecture that such a decrease tends to give a better problem formulation of a DAE from SA perspective.

Our works [6, 10, 11] show how to construct block triangular forms (BTFs) of a DAE using the structural data obtained from the $\Sigma$-method. A BTF indicates how each part of the DAE influences [resp. is influenced by] other parts. The interdependences between all pairs of blocks may be depicted by a fine-block graph [11]. Exploiting the underlying structure of a DAE, we can compute the derivatives of its solution in a blockwise fashion [7], or perform a dummy derivative index reduction algorithm [4, 5].

We refer the reader to the previous article [13] for a summary of the $\Sigma$-method, details of its failures, the basic conversion methods, and explanations of the equivalence of DAEs. By "basic" we mean that these methods do not exploit BTFs of a DAE. We shall follow the notation in [13].

In this article, we combine our conversion methods with a block triangularization of a DAE and derive our block conversion methods. When the System Jacobian is identically singular, and the DAE has a nontrivial BTF—that is, having at least two diagonal blocks—we can identify which blocks are identically singular and perform a conversion on each such block. Now that we only deal with equations and variables within a block, which is usually of a smaller size compared to the whole DAE, these block methods require fewer symbolic computations and hence are expected to be more efficient in finding a useful conversion for fixing SA's failures.

Section 2 reviews BTFs of a sparsity pattern and BTFs of a DAE. Section 3 presents our block conversion methods and demonstrates their application on a DAE from [1]. Section 4 gives more examples, in which the two DAEs are obtained from electrical circuit analysis [3]. Section 5 gives concluding remarks.

## 2 Block triangularization of DAEs.

In §2.1, we introduce notation for a BTF of a sparsity pattern. In §2.2, we review how to derive a BTF of a DAE; more details are in [10, 11].

We do not repeat the definitions and formulas for the notation in the $\Sigma$-method theory, such as a *signature matrix* $\boldsymbol{\Sigma} = (\sigma_{ij})$ and its *value* $\mathrm{Val}(\boldsymbol{\Sigma})$, a *highest-value*

[1] Throughout this article, "derivatives of a variable" include the variable itself as its 0th derivative.

*transversal (HVT) T* of $\mathbf{\Sigma}$, a *valid offset pair* $(\mathbf{c};\mathbf{d})$, a *System Jacobian* $\mathbf{J}(\mathbf{c};\mathbf{d}) = (J_{ij})$, and so forth. We refer the reader to [13] for details.

Terms are in *slanted font* at their defining occurrence. We use bold font for matrices that may split into blocks, and for the sub-matrices. Individual entries of a matrix are in lowercase. For example, matrix $\mathbf{A}$ has sub-matrices $\mathbf{A}_{lm}$ and entries $a_{ij}$.

## 2.1 Block triangular forms of a sparsity pattern.

Let[2] $R = 1:n$ be the set of indices of $n$ rows (equations), and let $C = 1:n$ be the set of indices of $n$ columns (variables). A *sparsity pattern* $\mathbf{A}$ is a subset of the Cartesian product $R \times C$ that contains row-column index pairs $(i, j)$. We can view $\mathbf{A}$ as its incidence matrix $(a_{ij})$, where $a_{ij}$ equals 1 if $(i, j) \in \mathbf{A}$ and 0 otherwise. A *transversal* of $\mathbf{A}$ is $n$ positions in $\mathbf{A}$ with exactly one position in each row and each column. If $\mathbf{A}$ has some transversal, then it is *structurally nonsingular*. The union of all transversals of $\mathbf{A}$ comprise its *essential sparsity pattern* $\mathbf{A}_{\text{ess}}$ [11]. Obviously, $\mathbf{A}$ is structurally nonsingular if and only if $\mathbf{A}_{\text{ess}}$ is nonempty.

Assume henceforth that $\mathbf{A}$ is structurally nonsingular. Let $P$ and $Q$ be two suitable permutation matrices for $\mathbf{A}$, such that the permuted incidence matrix $\mathbf{A}' = P\mathbf{A}Q$ can be written in a $p \times p$ block form

$$\mathbf{A}' = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1p} \\ & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2p} \\ & & \ddots & \vdots \\ & & & \mathbf{A}_{pp} \end{bmatrix} , \tag{2.1}$$

where each diagonal block $\mathbf{A}_{qq}$, $q = 1:p$, is square of positive size $N_q$. We say the block form (2.1) is a *BTF* of $\mathbf{A}$. Blanks in (2.1) mean that a sub-matrix $\mathbf{A}_{kl}$ below the block diagonal with $k > l$ is empty.

A sparsity pattern is *irreducible*, if it cannot be permuted to the form (2.1) with $p > 1$ [2]; otherwise it is *reducible*. A BTF is *irreducible* if each diagonal block is *irreducible*; otherwise it is *reducible* [11]. Hence, if (2.1) is irreducible, then $p$ is the largest number of diagonal blocks among all possible BTFs of $\mathbf{A}'$.

When we say block $q$ of a matrix in a BTF, we shall refer to the $q$th diagonal block submatrix. For $q = 1:p$, we define for block $q$ the index set

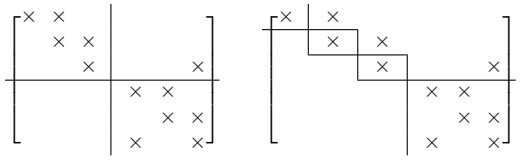
$$B_q = \text{the set of indices } i \text{ that belong to block } q \ .$$

Throughout this article, they are the indices of the permuted $\mathbf{A}'$, not those of the original $\mathbf{A}$.

Another useful notation is blockOf$(i)$ that denotes the block number $q$ such that index $i \in B_q$. Since each diagonal block is square, both $B_q$ and blockOf$(i)$ notation apply to rows and columns equally. To summarize, for $i \in 1:n$ and $q \in 1:p$,

$$\text{blockOf}(i) = q \iff i \in B_q \iff \sum_{m=1}^{q-1} N_m + 1 \le i \le \sum_{m=1}^{q} N_m \ .$$

[2] The colon notation $p:q$ for integers $p,q$ denotes either the unordered set or the enumerated list of integers $i$ with $p \le i \le q$, depending on context.



(a)

| $q$ | $N_q$ | $B_q$ |
|---|---|---|
| 1 | 1 | $\{1\}$ |
| 2 | 1 | $\{2\}$ |
| 3 | 1 | $\{3\}$ |
| 4 | 3 | $\{4,5,6\}$ |

(b)

Fig. 2.1: (a) Two nontrivial BTFs of the same sparsity pattern. The left one is reducible with number of blocks $p = 2$. The right one is irreducible with $p = 4$. (b) Block information for the irreducible BTF.

*Example 2.1* We illustrate in Figure 2.1 the above block notation with a sparsity pattern of two nontrivial BTFs.

The following lemma connects the transversals of a sparsity pattern $\mathbf{A}$ and the transversals of its diagonal blocks in some BTF.

**Lemma 2.1** *[11, Lemma 2.4] Any transversal $T$ of a sparsity pattern $\mathbf{A}$ is contained in the union of the diagonal blocks of any BTF of $\mathbf{A}$, that is, $T \subseteq \mathbf{A}_{11} \cup \cdots \cup \mathbf{A}_{pp}$.*

Equivalently, the intersection of $T$ with block $q$ of $\mathbf{A}$ is a transversal $T_q$ of $\mathbf{A}_{qq}$.

## 2.2 Block triangular forms of a DAE.

The natural sparsity pattern of a DAE indicates if a variable $x_j$ occurs in an equation $f_i$. Each such occurrence corresponds to a finite entry $\sigma_{ij}$ in $\boldsymbol{\Sigma}$, and hence we have

$$\mathbf{S} = \{\, (i,j) \mid \sigma_{ij} > -\infty \,\} \qquad \text{(the sparsity pattern of } \boldsymbol{\Sigma}\text{)} .$$

If $\mathbf{S}$ has some transversal, then $\boldsymbol{\Sigma}$ has a transversal with finite $\sigma_{ij}$'s and a finite $\mathrm{Val}(\boldsymbol{\Sigma})$ [9], so the DAE is structurally well posed (SWP) [10]; otherwise it is structurally ill posed. Here, we shall deal with the SWP case only.

A more informative BTF derives from the sparsity pattern $\mathbf{S}_0 = \mathbf{S}_0(\mathbf{c};\mathbf{d})$ of a System Jacobian $\mathbf{J} = \mathbf{J}(\mathbf{c};\mathbf{d})$ as defined in [13, (2.6)]:

$$\mathbf{S}_0 = \{\, (i,j) \mid d_j - c_i = \sigma_{ij} \,\} \qquad \text{(the sparsity pattern of } \mathbf{J}\text{)} . \tag{2.2}$$

By [13, (2.2)], $d_j - c_i = \sigma_{ij}$ holds on a HVT $T$ of $\boldsymbol{\Sigma}$, so $T$ is also a transversal of $\mathbf{S}_0$.

A less obvious set contains the positions that contribute to $\det(\mathbf{J})$:

$$\mathbf{S}_{\text{ess}} = \text{the union of all HVTs of } \boldsymbol{\Sigma} \qquad \text{(the essential sparsity pattern of } \boldsymbol{\Sigma}\text{)} ,$$

which is also the essential sparsity pattern of $\mathbf{S}_0$ for any valid offset pair $(\mathbf{c};\mathbf{d})$ [11, Lemma 3.1].

Since $d_j - c_i = \sigma_{ij}$ holds on each HVT and hence implies $\sigma_{ij} > -\infty$, we have

$$\mathbf{S}_{\text{ess}} \subseteq \mathbf{S}_0 \subseteq \mathbf{S} \qquad \text{for any offset pair } (\mathbf{c};\mathbf{d}) .$$

Our experience suggests that the (irreducible) BTF based on $\mathbf{S}_0$ can be significantly finer than that based on $\mathbf{S}$. We refer to the former BTF as *fine BTF*, and to the latter as *coarse BTF*. We refer to the diagonal blocks in the fine BTF as *fine blocks*, and refer to those in the coarse BTF as *coarse blocks*.

Assume that $\mathbf{S}_0$ is permuted into a $p \times p$ BTF. Following this BTF, we apply the same permutations on $\mathbf{J}$ and $\boldsymbol{\Sigma}$, and write them in $p \times p$ block forms:

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} & \cdots & \mathbf{J}_{1p} \\ \mathbf{0} & \mathbf{J}_{22} & \cdots & \mathbf{J}_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{J}_{pp} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1p} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{p1} & \cdots & \cdots & \boldsymbol{\Sigma}_{pp} \end{bmatrix}. \tag{2.3}$$

We call this procedure a *block triangularization* of the DAE, and note that the sparsity pattern $\mathbf{S}$ of $\boldsymbol{\Sigma}$ may not be in the same BTF as $\mathbf{S}_0$ of $\mathbf{J}$. That is, every $\sigma_{ij}$ below the block diagonal of $\boldsymbol{\Sigma}$ is not necessarily $-\infty$, but must satisfy $\sigma_{ij} < d_j - c_i$ as $J_{ij} \equiv 0$. Hence,

$$d_j - c_i \begin{cases} > \sigma_{ij} & \text{if blockOf}(j) < \text{blockOf}(i) \\ \geq \sigma_{ij} & \text{if blockOf}(j) \geq \text{blockOf}(i)\ . \end{cases} \tag{2.4}$$

We refer the reader to [6, 11] for more details on BTFs.

*Example 2.2* We illustrate the coarse and fine BTFs with the (artificially) modified double pendula DAE in [7]. The state variables are $x, y, \lambda, u, v, \mu$; $G$ is gravity, $L > 0$ is the length of both pendula, and $\alpha$ is a constant.

$$\begin{aligned} 0 &= f_1 = x'' + x\lambda & 0 &= f_4 = u'' + u\mu \\ 0 &= f_2 = y'' + y\lambda + (x')^3 - g & 0 &= f_5 = (v''')^3 + v\mu - G \\ 0 &= f_3 = x^2 + y^2 - L^2 & 0 &= f_6 = u^2 + v^2 - (L + \alpha\lambda)^2 + \lambda'' \end{aligned}$$

$$\boldsymbol{\Sigma} = \begin{array}{c|cccccc|c} & v & \mu & u & x & y & \lambda & c_i \\ \hline f_5 & 3^\bullet & 0 & & & & & 0 \\ f_4 & & 0^\bullet & 2 & & & & 0 \\ f_6 & 0 & & 0^\bullet & & & 2 & 2 \\ \hline f_3 & & & & 0^\bullet & 0 & & 6 \\ f_2 & & & & 1 & 2^\bullet & 0 & 4 \\ f_1 & & & & 2 & & 0^\bullet & 4 \\ \hline d_j & 3 & 0 & 2 & 6 & 6 & 4 & \end{array} \qquad \mathbf{J} = \begin{array}{c|cccccc} & v''' & \mu & u'' & x^{(6)} & y^{(6)} & \lambda^{(4)} \\ \hline f_5 & 2v''' & v & & & & \\ f_4 & & u & 1 & & & \\ f_6'' & & & 2u & & & 1 \\ \hline f_3^{(6)} & & & & 2x & 2y & \\ f_2^{(4)} & & & & & 1 & y \\ f_1^{(4)} & & & & 1 & & x \end{array}$$

The row and column labels in $\mathbf{J}$, showing equations and variables differentiated to order $c_i$ and $d_j$, aim to remind the reader of the formula for $\mathbf{J}$ in [13, (2.6)].

There are two $3 \times 3$ coarse blocks. The first one, comprising equations $f_5, f_4, f_6$ and variables $v, \mu, u$, can further decompose into three $1 \times 1$ fine blocks, while the second coarse block, comprising equations $f_3, f_2, f_1$ and variables $x, y, \lambda$, is irreducible. Hence there are four blocks in the fine BTF.

The sparsity pattern $\mathbf{S}_0$ of $\mathbf{J}$ is exactly the one in Figure 2.1(a), so the fine BTF information is in Figure 2.1(b).

If we state Lemma 2.1 in the context of a Jacobian sparsity pattern, then we have the following lemma.

**Lemma 2.2** *[11, Lemma 3.3] Assume that a Jacobian sparsity pattern $\mathbf{S}_0$ is in some BTF. Let $(\boldsymbol{\Sigma}_{qm})_{q,m=1:p}$ be the corresponding sub-matrices of $\boldsymbol{\Sigma}$. Then a HVT $T$ of $\boldsymbol{\Sigma}$ is the union of HVTs $T_q$ of the diagonal blocks $\boldsymbol{\Sigma}_{qq}$: $T = T_1 \cup \cdots \cup T_p$.*

This lemma is not difficult to prove, given that a transversal $T$ of $\mathbf{S}_0$ is the union of transversals $T_q$ of the diagonal blocks of $\mathbf{S}_0$.

The following lemma is useful for proving the main Theorems 3.1 and 3.2 of the block conversion methods in §3.

**Lemma 2.3** *Assume that $\boldsymbol{\Sigma}$ has a finite* Val$(\boldsymbol{\Sigma})$ *and is in a $p \times p$ block form as in (2.3). Let* $\mathbf{c}$ *and* $\mathbf{d}$ *be two nonnegative integer n-vectors. Assume also that*

*(a)* $d_j - c_i > \sigma_{ij}$ *holds for all entries below the diagonal blocks of* $\boldsymbol{\Sigma}$,
*(b)* $d_j - c_i \geq \sigma_{ij}$ *holds elsewhere, and*
*(c)* Val$(\boldsymbol{\Sigma}) = \sum_j d_j - \sum_i c_i$.

*Then*

*(i)* $(\mathbf{c};\mathbf{d})$ *is a valid offset pair of* $\boldsymbol{\Sigma}$,
*(ii) the block form of* $\boldsymbol{\Sigma}$ *is a BTF of the Jacobian sparsity pattern* $\mathbf{S}_0$, *and*
*(iii) a HVT of* $\boldsymbol{\Sigma}$ *is the union of HVTs* $T_q$ *of the diagonal blocks* $\boldsymbol{\Sigma}_{qq}$, *for all* $q = 1:p$.

*Proof (i)* We let $T$ denote a HVT of $\boldsymbol{\Sigma}$. Since Val$(\boldsymbol{\Sigma})$ is finite, $\sigma_{ij} \geq 0$ for all $(i,j) \in T$. For $(\mathbf{c};\mathbf{d})$ to be a valid offset of $\boldsymbol{\Sigma}$, $d_j - c_i \geq \sigma_{ij}$ must hold for all $i, j = 1:n$, with equalities for all $(i,j) \in T$ [9].

By (a) and (b), $d_j - c_i \geq \sigma_{ij}$ holds everywhere. Summing these inequalities over $T$ gives

$$\sum_{(i,j)\in T} (d_j - c_i) \geq \sum_{(i,j)\in T} \sigma_{ij} \ .$$

The left-hand side equals $\sum_j d_j - \sum_i c_i$, and the right-hand side equals Val$(\boldsymbol{\Sigma})$ by definition. By (c), these two values are equal, so $d_j - c_i = \sigma_{ij}$ holds for all $(i,j) \in T$, and $(\mathbf{c};\mathbf{d})$ is valid for $\boldsymbol{\Sigma}$.

*(ii)* By (a), the blocks below the block diagonal in $\mathbf{S}_0$, derived from $\boldsymbol{\Sigma}$ and $(\mathbf{c};\mathbf{d})$ using (2.2), are empty. By the definition of a BTF of a Jacobian sparsity pattern, $\mathbf{S}_0$ is in a BTF as described by the $p \times p$ block form.

*(iii)* This follows immediately from *(ii)* and Lemma 2.2. □

Following a $p \times p$ BTF based on $\mathbf{S}_0$, we can write any valid offset pair $(\mathbf{c};\mathbf{d})$ of $\boldsymbol{\Sigma}$ in a block form as

$$(\mathbf{c}_1;\mathbf{d}_1), (\mathbf{c}_2;\mathbf{d}_2), \ldots, (\mathbf{c}_p;\mathbf{d}_p) \ , \tag{2.5}$$

where each of the sub-vectors $\mathbf{c}_q$ and $\mathbf{d}_q$ is of length $N_q$, where $q = 1:p$.

**Lemma 2.4** *Assume that a Jacobian pattern* $\mathbf{S}_0$*, derived by* $\boldsymbol{\Sigma}$ *and a valid offset pair* $(\mathbf{c};\mathbf{d})$*, is in some BTF. If we write* $(\mathbf{c};\mathbf{d})$ *into block form as in (2.5), then* $(\mathbf{c}_q;\mathbf{d}_q)$ *is a valid offset pair of* $\boldsymbol{\Sigma}_{qq}$*.*

*Proof* Let $T$ be a HVT of $\boldsymbol{\Sigma}$. By Lemma 2.2, the intersection of $T$ with block $q$ is a HVT $T_q$ of $\boldsymbol{\Sigma}_{qq}$. Then $d_j - c_i = \sigma_{ij}$ holds for all $(i,j) \in T_q \subseteq T$. Since $(\mathbf{c};\mathbf{d})$ is valid for $\boldsymbol{\Sigma}$, $d_j - c_i \geq \sigma_{ij}$ and $c_i \geq 0$ hold on $\boldsymbol{\Sigma}_{qq}$, where $i,j \in B_q$. Thus the offset pair $(\mathbf{c}_q;\mathbf{d}_q)$ matched to block $q$ satisfies the conditions [13, (2.2)] for being valid for $\boldsymbol{\Sigma}_{qq}$. □

From the view of Lemma 2.4, we can regard each diagonal block $\boldsymbol{\Sigma}_{qq}$ as a signature matrix in its own right. Equivalently, each block $q$, having $N_q$ equations in $N_q$ variables, can be viewed as a sub-DAE, with a signature matrix $\boldsymbol{\Sigma}_{qq}$, a finite $\mathrm{Val}(\boldsymbol{\Sigma}_{qq})$, a *local offset pair* $(\mathbf{c}_q;\mathbf{d}_q)$, and a *sub-Jacobian* $\mathbf{J}_{qq}$. Expressions that contribute to entries in an off-diagonal block $\boldsymbol{\Sigma}_{qm}$, where $q \neq m$, can be considered as driving terms, or equivalently, the influence of variables in block $m$ on those in block $q$. We refer to $(\mathbf{c};\mathbf{d})$ of $\boldsymbol{\Sigma}$ as a *global offset pair*. The reader is referred to [11] for more theoretical results about block triangularization and global/local offset pairs.

## 3 Block conversion methods.

They are suitable for improving the efficiency of finding a useful conversion for fixing SA's failures. If $\mathbf{J}$ is identically singular, then by (2.3), $\det(\mathbf{J}) = \prod_{q=1}^{p} \det(\mathbf{J}_{qq}) \equiv 0$, so at least one $\mathbf{J}_{qq}$ for some $q \in 1:p$ is identically singular. As discussed before, we can regard block $q$ as a sub-DAE with a signature matrix $\boldsymbol{\Sigma}_{qq}$. Then we wish to apply the basic conversion methods on this sub-DAE to achieve a strict decrease in $\mathrm{Val}(\boldsymbol{\Sigma}_{qq})$, provided the conditions for applying these methods are satisfied for those variables and equations within block $q$.

However, what we should ensure is a strict decrease in the value of the *whole* signature matrix, namely $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) < \mathrm{Val}(\boldsymbol{\Sigma})$, where $\overline{\boldsymbol{\Sigma}}$ is the signature matrix of the resulting DAE. Proving this inequality from a decrease in $\mathrm{Val}(\boldsymbol{\Sigma}_{qq})$ is nontrivial, because a conversion on block $q$ may affect blocks $\boldsymbol{\Sigma}_{qm}$ for $m = 1,\ldots,q-1,q+1,\ldots,p$. Especially in the ES method, $\boldsymbol{\Sigma}_{qq}$ and these blocks are enlarged. Hence, the conditions and the conversion process need to be carefully modified, so that the conversion methods can adapt to a BTF based on $\mathbf{S}_0$.

We give an introductory example in §3.1, present the block LC method in §3.2, and present the block ES method in §3.3.

Hereafter we use the fine BTF in the examples for demonstration, since each fine block contains an irreducible sub-Jacobian sparsity pattern. Our experience suggests that a useful conversion can usually be derived from the fine BTF of a DAE. However, we emphasize that the block conversion methods can be applied not only to the irreducible BTF of a Jacobian sparsity pattern $\mathbf{S}_0$ with some valid $(\mathbf{c};\mathbf{d})$, but also to any BTF of $\mathbf{S}_0$. For example, the basic conversion methods consider a DAE in a (trivial) BTF of one $n \times n$ block.

3.1 An introductory example.

We illustrate these block methods with the following DAE:

$$\begin{aligned} 0 &= f_1 = x_1 + x_2 + h_1(t) \\ 0 &= f_2 = x_1 + (x_1' + x_2')x_3' + h_2(t) \\ 0 &= f_3 = x_3' + h_3(t)\ . \end{aligned} \tag{3.1}$$

$$\boldsymbol{\Sigma} = \begin{array}{c} \\ f_1 \\ f_2 \\ f_3 \\ d_j \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \left[\begin{array}{ccc|c} 0^{\bullet} & 0 & & 1 \\ 1 & 1^{\bullet} & 1 & 0 \\ \hline & & 1^{\bullet} & 0 \end{array}\right] \\ \begin{array}{ccc} 1 & 1 & 1 \end{array} \end{array} \qquad \mathbf{J} = \begin{array}{c} \\ f_1' \\ f_2 \\ f_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1' & x_2' & x_3' \end{array} \\ \left[\begin{array}{cc|c} 1 & 1 & \\ x_3' & x_3' & x_1' + x_2' \\ \hline & & 1 \end{array}\right] \end{array}$$

(Here $h_1, h_2, h_3$ are driving functions.) The coarse BTF and the fine BTF are identical, both having two diagonal blocks.

In the basic LC method, we can choose $\mathbf{u} = [-x_3', 1, -x_1' - x_2']^T \in \operatorname{coker}(\mathbf{J})$. Using [13, (4.3)], we have

$$I = \{\, i \mid u_i \not\equiv 0 \,\} = \{\, 1,2,3 \,\}, \quad \underline{c} = \min_{i \in I} c_i = 0, \quad L = \{\, l \in I \mid c_l = \underline{c} \,\} = \{\, 2,3 \,\}\ .$$

We let $\sigma(x_j, \mathbf{u})$ denote the order of the highest derivative to which $x_j$ occurs in $\mathbf{u}$, or $-\infty$ if $x_j$ does not occur in $\mathbf{u}$ [13, (4.1)]. The LC condition [13, (4.4)] is violated since

$$\sigma(x_j, \mathbf{u}) = 1 \not< 1 = d_j - \underline{c} \quad \text{for all} \quad j = 1{:}3\ .$$

Not surprisingly, replacing either $f_2$ or $f_3$ by

$$\overline{f} = \sum_{i \in I} u_i f_i^{(c_i - \underline{c})} = -x_3' h_1'(t) + \big(x_1 + h_2(t)\big) - (x_1' + x_2')\big(x_3' + h_3(t)\big)$$

does not result in a decrease in $\operatorname{Val}(\boldsymbol{\Sigma})$; verifying this is not difficult.

Notice that only the sub-Jacobian of block 1, $\mathbf{J}_{11} = \partial(f_1', f_2)/\partial(x_1', x_2')$, is singular. Suppose we consider block 1, with $B_1 = \{\, 1,2 \,\}$, as a sub-DAE, and choose $\mathbf{u} = [-x_3', 1]^T \in \operatorname{coker}(\mathbf{J}_{11})$. Within block 1, the LC method derives

$$I = \{\, i \in B_1 \mid u_i \not\equiv 0 \,\} = \{\, 1,2 \,\}, \quad \underline{c} = \min_{i \in I} c_i = 0, \quad L = \{\, l \in I \mid c_l = \underline{c} \,\} = \{\, 2 \,\}\ .$$

Now the LC condition [13, (4.4)] is satisfied for the column indices in block 1:

$$\sigma(x_j, \mathbf{u}) = -\infty < d_j - \underline{c} \quad \text{for } j = 1,2 \in B_1\ .$$

Replacing $f_2$ by $\overline{f}_2 = u_1 f_1' + u_2 f_2 = x_1 + h_2(t) - x_3' h_1'(t)$ results in the DAE with the following SA result.

$$\overline{\boldsymbol{\Sigma}} = \begin{array}{c} \\ \overline{f}_1 \\ \overline{f}_2 \\ f_3 \\ d_j \end{array} \begin{array}{c} \begin{array}{cccc} x_2 & x_1 & x_3 & c_i \end{array} \\ \left[\begin{array}{ccc|c} 0^{\bullet} & 0 & & 0 \\ \hline & 0^{\bullet} & 1 & 0 \\ \hline & & 1^{\bullet} & 0 \end{array}\right] \\ \begin{array}{cccc} 0 & 0 & 1 & \end{array} \end{array} \qquad \overline{\mathbf{J}} = \begin{array}{c} \\ \overline{f}_1 \\ \overline{f}_2 \\ f_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_2 & x_1 & x_3' \end{array} \\ \left[\begin{array}{cc|c} 1 & 1 & \\ \hline & 1 & g_1'(t) \\ \hline & & 1 \end{array}\right] \end{array}$$

The SA succeeds as $\overline{\mathbf{J}}$ is nonsingular. The conversion results in a decrease in the value of the signature matrix: $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) = 1 < 2 = \mathrm{Val}(\boldsymbol{\Sigma})$.

The basic ES method can work on (3.1) by choosing $\mathbf{v} = [1,-1,0]^T \in \ker(\mathbf{J})$. It is simpler—though trivial for this example—to work on block 1 only. We find $\mathbf{v} = [1,-1]^T \in \ker(\mathbf{J}_{11})$, and use [13, (4.10)] to derive

$$J = \{\, l \in B_1 \mid v_l \not\equiv 0 \,\} = \{\, 1,2 \,\}, \quad s = |J| = 2, \quad M = \{\, 1,2 \,\}, \quad \bar{c} = \max_{i \in M} c_i = 1\ .$$

Since $\mathbf{v}$ is constant, it is not difficult to verify that the ES conditions [13, (4.11)] hold.

We choose $l = 2 \in J$ and introduce for $x_1$ a new variable

$$y_1 = x_1^{(d_1 - \bar{c})} - \frac{v_1}{v_2} \cdot x_2^{(d_2 - \bar{c})} = x_1 + x_2\ .$$

The ES method hence says: replace $x_1$ by $y_1 - x_2$ in $f_1$, and replace $x_1'$ by $y_1' - x_2'$ in $f_2$. Finally we append the equation $g_1$ that prescribes such replacements, and obtain

$$\begin{aligned} 0 &= \overline{f}_1 = y_1 + h_1(t) & \qquad 0 &= \overline{f}_3 = x_3' + h_3(t) \\ 0 &= \overline{f}_2 = x_1 + y_1' x_3' + h_2(t) & \qquad 0 &= g_1 = -y_1 + x_1 + x_2\ . \end{aligned}$$

$$\overline{\boldsymbol{\Sigma}} = \begin{array}{c} \\ g_1 \\ \overline{f}_2 \\ \overline{f}_1 \\ f_3 \\ d_j \end{array} \begin{array}{c} \begin{array}{ccccc} x_2 & x_1 & y_1 & x_3 & c_i \end{array} \\ \left[\begin{array}{ccc|c|c} 0^{\bullet} & 0 & 0 & & 0 \\ \hline & 0^{\bullet} & 1 & 1 & 0 \\ \hline & & 0^{\bullet} & & 1 \\ \hline & & & 1^{\bullet} & 0 \end{array}\right] \\ \begin{array}{cccc} 0 & 0 & 1 & 1 \end{array} \end{array} \qquad \overline{\mathbf{J}} = \begin{array}{c} \\ g_1 \\ \overline{f}_2 \\ \overline{f}_1 \\ f_3 \end{array} \begin{array}{c} \begin{array}{cccc} x_2 & x_1 & y_1' & x_3' \end{array} \\ \left[\begin{array}{cc|c|c} 1 & 1 & -1 & \\ \hline & 1 & x_3' & y_1' \\ \hline & & 1 & \\ \hline & & & 1 \end{array}\right] \end{array}$$

Again $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) = 1 < 2 = \mathrm{Val}(\boldsymbol{\Sigma})$, and the SA succeeds as $\det(\overline{\mathbf{J}}) = 1$.

### 3.2 Block linear combination method.

We first introduce some convenient notation for the block LC method. Let $\mathbf{0}_r$ denote the zero column vector of size $r$. Assume that a $\mathbf{J}_{qq}$ is identically singular. Let $\widehat{\mathbf{u}} \in \mathrm{coker}(\mathbf{J}_{qq})$, where $\widehat{\mathbf{u}} \not\equiv \mathbf{0}_{N_q}$. Let also

$$\mathbf{u} = \begin{bmatrix} \mathbf{0}_{N_1 + \cdots + N_{q-1}} \\ \widehat{\mathbf{u}} \\ \mathbf{0}_{N_{q+1} + \cdots + N_p} \end{bmatrix}\ .$$

Denote

$$\begin{aligned} &I = \{\, i \mid u_i \not\equiv 0 \,\} \subseteq B_q, \quad \underline{c} = \min_{i \in I} c_i, \\ &L = \{\, l \in I \mid c_l = \underline{c} \,\}, \quad \text{and} \quad \overline{L} = \{\, l \in L \mid u_l \text{ is (nonzero) constant} \,\}\ . \end{aligned} \tag{3.2}$$

The set $\overline{L}$ is used to seek a conversion that guarantees equivalence between the original DAE and the converted one. The block LC method is based on the following theorem.

**Theorem 3.1** *If*

$$\sigma(x_j, \mathbf{u}) < d_j - \underline{c} \qquad \textit{for all } j \in B_q \tag{3.3}$$

*and we replace an equation $f_l$, $l \in L$, by*

$$\overline{f} = \sum_{i \in I} u_i f_i^{(c_i - \underline{c})}\ ,$$

*then* $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) < \mathrm{Val}(\boldsymbol{\Sigma})$, *where* $\overline{\boldsymbol{\Sigma}} = (\overline{\sigma}_{ij})$ *is the signature matrix of the resulting DAE.*

Before proving this theorem, we show how to apply the block LC method and prove a related lemma.

*Example 3.1* We illustrate the block LC method with the Campbell-Griepentrog two-link robot arm DAE [1]. We slightly simplify the problem formulation to (3.4), allowing the first-order derivatives $x_1'$, $x_2'$, and $x_3'$ to occur implicitly in the equations. The two state variables $u_1$ and $u_2$ in the original formulation are renamed $x_4$ and $x_5$, respectively (and not to be confused with entries in a vector $\mathbf{u}$ in our notation).

The equations of this problem are

$$\begin{aligned} 0 = A = {} & x_1'' - \Big[2c(x_3)(x_1' + x_3')^2 + x_1'^2 d(x_3) + (2x_3 - x_2)\big(a(x_3) + 2b(x_3)\big) \\ & + a(x_3)(x_4 - x_5)\Big] \\ 0 = B = {} & x_2'' - \Big[-2c(x_3)(x_1' + x_3')^2 - x_1'^2 d(x_3) + (2x_3 - x_2)\big(1 - 3a(x_3) - 2b(x_3)\big) \\ & - a(x_3)x_4 + \big(a(x_3) + 1\big)x_5\Big] \\ 0 = C = {} & x_3'' - \Big[-2c(x_3)(x_1' + x_3')^2 - x_1'^2 d(x_3) + (2x_3 - x_2)\big(a(x_3) - 9b(x_3)\big) \\ & - 2x_1'^2 c(x_3) - d(x_3)\big(x_1' + x_3'\big)^2 - \big(a(x_3) + b(x_3)\big)(x_4 - x_5)\Big] \\ 0 = D = {} & \cos x_1 + \cos(x_1 + x_3) - p_1(t) \\ 0 = E = {} & \sin x_1 + \sin(x_1 + x_3) - p_2(t)\ , \end{aligned} \tag{3.4}$$

where

$$\begin{aligned} a(\theta) &= 2/(2 - \cos^2\theta) & b(\theta) &= \cos\theta/(2 - \cos^2\theta) \\ c(\theta) &= \sin\theta/(2 - \cos^2\theta) & d(\theta) &= \sin\theta\cos\theta/(2 - \cos^2\theta) \\ p_1(t) &= \cos(1 - e^t) + \cos(1 - t) & p_2(t) &= \sin(1 - e^t) + \sin(1 - t)\ . \end{aligned}$$

$$\boldsymbol{\Sigma} = \begin{array}{c|ccccc|c} & x_2 & x_4 & x_5 & x_1 & x_3 & c_i \\ f_1\ B & 2^\bullet & 0 & 0 & 1 & 1 & 0 \\ f_2\ C & 0 & 0^\bullet & 0 & 1 & 2 & 0 \\ f_3\ A & 0 & 0 & 0^\bullet & 2 & 1 & 0 \\ f_4\ D & & & & 0^\bullet & 0 & 2 \\ f_5\ E & & & & 0 & 0^\bullet & 2 \\ \hline d_j & 2 & 0 & 0 & 2 & 2 & \end{array} \qquad \mathbf{J} = \begin{array}{c|ccccc} & x_2'' & x_4 & x_5 & x_1'' & x_3'' \\ B & 1 & a_3 & -a_3-1 & & \\ C & & a_3+b_3 & -a_3-b_3 & & 1 \\ A & & -a_3 & a_3 & 1 & \\ D'' & & & & \frac{\partial D}{\partial x_1} & \frac{\partial D}{\partial x_3} \\ E'' & & & & \frac{\partial E}{\partial x_1} & \frac{\partial E}{\partial x_3} \end{array}$$

Here in $\mathbf{J}$,

$$\begin{aligned} a_3 = a(x_3) &= 2/(2-\cos^2 x_3) & b_3 = b(x_3) &= \cos x_3/(2-\cos^2 x_3) \\ \partial D/\partial x_1 &= -\sin x_1 - \sin(x_1+x_3) & \partial D/\partial x_3 &= -\sin(x_1+x_3) \\ \partial E/\partial x_1 &= \cos x_1 + \cos(x_1+x_3) & \partial E/\partial x_3 &= \cos(x_1+x_3)\,. \end{aligned}$$

The DAE (3.4) is of differentiation index 5, while the SA reports structural index $\nu_S = 3$. Hence this must be a failure case, because $\nu_S$ is an upper bound for the differentiation index when the SA succeeds [9]. We can see that the sub-Jacobian $\mathbf{J}_{22}$ of block 2 is identically singular.

Our method first computes $\widehat{\mathbf{u}} = [2, 2+\cos x_3]^T \in \operatorname{coker}(\mathbf{J}_{22})$. Then $\mathbf{u} = [0, 2, 2+\cos x_3, 0, 0]^T$. Using (3.2), we have

$$I = \{\, i \mid u_i \not\equiv 0 \,\} = \{2,3\}, \quad \underline{c} = \min_{i\in I} c_i = 0, \quad L = \{2,3\}, \quad \text{and} \quad \overline{L} = \{2\}\,.$$

The variables $x_4$ and $x_5$ in block 2 do not occur in $\mathbf{u}$, so the condition (3.3) is satisfied.

Considering equivalence, we pick $l = 2 \in \overline{L}$ over $l = 3 \in L \setminus \overline{L}$, and replace $f_l = C$ by $\overline{C} = u_1 C + u_2 A = 2C + (2+\cos x_3)A$. The SA results of the resulting DAE are as follows.

$$\overline{\boldsymbol{\Sigma}} = \begin{array}{c|ccccc|c} & x_4 & x_5 & x_2 & x_1 & x_3 & c_i \\ A & 0^\bullet & 0 & 0 & 2 & 1 & 0 \\ B & 0 & 0^\bullet & 2 & 1 & 1 & 0 \\ \overline{C} & & & 0^\bullet & 2 & 2 & 2 \\ D & & & & 0^\bullet & 0 & 4 \\ E & & & & 0 & 0^\bullet & 4 \\ \hline d_j & 0 & 0 & 2 & 4 & 4 & \end{array} \qquad \overline{\mathbf{J}} = \begin{array}{c|ccccc} & x_4 & x_5 & x_2'' & x_1^{(4)} & x_3^{(4)} \\ A & -a_3 & a_3 & & & \\ B & a_3 & -a_3-1 & 1 & & \\ \overline{C}'' & & & \frac{\partial \overline{C}}{\partial x_2} & 2+\cos x_3 & 2 \\ D^{(4)} & & & & \frac{\partial D}{\partial x_1} & \frac{\partial D}{\partial x_3} \\ E^{(4)} & & & & \frac{\partial E}{\partial x_1} & \frac{\partial E}{\partial x_3} \end{array}$$

Here $\partial\overline{C}/\partial x_2 = 2(a_3^2 - 3a_3b_3 + b_3^2)(2-\cos^2 x_3)$. The SA reports $\nu_S = 5$ and succeeds at any point where

$$\det(\overline{\mathbf{J}}) = 4(a_3^2 - 3a_3b_3 + b_3^2)\sin x_3 \neq 0\,.$$

Now $\operatorname{Val}(\overline{\boldsymbol{\Sigma}}) = 0 < 2 = \operatorname{Val}(\boldsymbol{\Sigma})$.

**Lemma 3.1** *Consider a BTF of a Jacobian pattern* $\mathbf{S}_0$ *derived from* $\boldsymbol{\Sigma}$ *and* $(\mathbf{c};\mathbf{d})$. *If we perform the LC conversion as described in Theorem 3.1, then in the resulting* $\overline{\boldsymbol{\Sigma}}$,

$$d_j - c_i \begin{cases} > \overline{\sigma}_{ij} & \text{if } blockOf(j) < blockOf(i) \\ \geq \overline{\sigma}_{ij} & \text{if } blockOf(j) \geq blockOf(i)\,. \end{cases} \tag{3.5}$$

*Proof* We only replace $f_l$ by $\overline{f}_l = \overline{f}$ in a conversion, so $\overline{\sigma}_{ij} = \sigma_{ij}$ for all $i \neq l$ and all $j$. By (2.4), (3.5) holds for all $i \neq l$.

When $i = l$, we consider two cases: (a) blockOf$(j) <$ blockOf$(l)$ and (b) blockOf$(j) \geq$ blockOf$(l)$.

(a) blockOf$(j) <$ blockOf$(l)$. By (2.4), $\sigma_{lj} < d_j - c_l$. Then $\overline{\sigma}_{lj}$ is

$$\sigma\left(x_j, \overline{f}_l\right) = \sigma\left(x_j, \sum_{i\in I} u_i f_i^{(c_i-\underline{c})}\right) \leq \max\left\{\sigma\left(x_j, \mathbf{u}\right), \max_{i\in I} \sigma\left(x_j, f_i^{(c_i-\underline{c})}\right)\right\}. \tag{3.6}$$

We use some simple derivations to obtain

$$\begin{aligned} \sigma\left(x_j, \mathbf{u}\right) &\leq \sigma\left(x_j, \mathbf{J}_{qq}\right) \leq \max_{i\in I} \sigma\left(x_j, f_i\right) = \max_{i\in I} \sigma_{ij} \\ &< \max_{i\in I}(d_j - c_i) = d_j - \min_{i\in I} c_i = d_j - c_l \quad \text{and} \end{aligned} \tag{3.7a}$$

$$\max_{i\in I} \sigma\left(x_j, f_i^{(c_i-\underline{c})}\right) = \max_{i\in I}\left(\sigma_{ij} + c_i - \underline{c}\right) < d_j - \underline{c} = d_j - c_l\,. \tag{3.7b}$$

Substituting (3.7a) and (3.7b) in (3.6), we obtain $\overline{\sigma}_{lj} = \sigma\left(x_j, \overline{f}_l\right) < d_j - c_l$.

(b) blockOf$(j) \geq$ blockOf$(l)$. By (2.4), $\sigma_{lj} \leq d_j - c_l$. We can replace the two "$<$" in (3.7) by "$\leq$", and using these inequalities in (3.6), we have $\overline{\sigma}_{lj} \leq d_j - c_l$. □

Using Lemma 3.1, we can now prove Theorem 3.1.

*Proof* By Lemma 2.4, we can regard block $q$ as a sub-DAE with $\boldsymbol{\Sigma}_{qq}$ and $(\mathbf{c}_q; \mathbf{d}_q)$. The conversion described in Theorem 3.1 can be considered as an application of the basic LC method to this sub-DAE. Since the block LC condition (3.3) holds, that is, $\sigma\left(x_j, \mathbf{u}\right) < d_j - \underline{c}$ for all $j \in B_q$ that belong to this sub-DAE, the basic LC condition [13, (4.4)] also holds for the sub-DAE. Hence $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}_{qq}) < \mathrm{Val}(\boldsymbol{\Sigma}_{qq})$.

Let $\overline{T}$ be a HVT of $\overline{\boldsymbol{\Sigma}}$. Using (3.5) in Lemma 3.1, we have $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) = \sum_{(i,j)\in\overline{T}} \overline{\sigma}_{ij} \leq d_j - c_i = \mathrm{Val}(\boldsymbol{\Sigma})$. Now we prove $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) < \mathrm{Val}(\boldsymbol{\Sigma})$ by contradiction. First assume that $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) = \sum_j d_j - \sum_i c_i = \mathrm{Val}(\boldsymbol{\Sigma}) \geq 0$. With (3.5) , the three conditions in Lemma 2.3 are satisfied. From this lemma, it follows that the Jacobian patterns $\overline{\mathbf{S}}_0$, derived from $\overline{\boldsymbol{\Sigma}}$ and $(\mathbf{c}; \mathbf{d})$, and $\mathbf{S}_0$, derived from $\boldsymbol{\Sigma}$ and $(\mathbf{c}; \mathbf{d})$, are in the same $p \times p$ BTF.

By Lemma 2.2, $\overline{T}$ is the union of HVTs $\overline{T}_m$ of all diagonal blocks $\overline{\boldsymbol{\Sigma}}_{mm}$, $m = 1 : p$. By the construction of $\overline{\boldsymbol{\Sigma}}$, $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}_{mm}) = \mathrm{Val}(\boldsymbol{\Sigma}_{mm})$ for all $m \neq q$. Then a contradiction follows from

$$\begin{aligned} \mathrm{Val}(\overline{\boldsymbol{\Sigma}}) &= \sum_{(i,j)\in\overline{T}} \overline{\sigma}_{ij} = \sum_{m=1}^{p} \sum_{(i,j)\in\overline{T}_m} \overline{\sigma}_{ij} = \sum_{m=1}^{p} \mathrm{Val}(\overline{\boldsymbol{\Sigma}}_{mm}) \\ &= \sum_{m\neq q} \mathrm{Val}(\overline{\boldsymbol{\Sigma}}_{mm}) + \mathrm{Val}(\overline{\boldsymbol{\Sigma}}_{qq}) < \sum_{m\neq q} \mathrm{Val}(\boldsymbol{\Sigma}_{mm}) + \mathrm{Val}(\boldsymbol{\Sigma}_{qq}) \\ &= \sum_{m=1}^{p} \mathrm{Val}(\boldsymbol{\Sigma}_{mm}) = \sum_{m=1}^{p} \sum_{(i,j)\in T_m} \sigma_{ij} = \sum_{(i,j)\in T} \sigma_{ij} = \mathrm{Val}(\boldsymbol{\Sigma})\,, \end{aligned} \tag{3.8}$$

where $T$ is a HVT of $\boldsymbol{\Sigma}$ and $T_m$ are HVTs of its diagonal blocks $\boldsymbol{\Sigma}_{mm}$. The assumption $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) = \mathrm{Val}(\boldsymbol{\Sigma})$ is hence false, so $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) < \mathrm{Val}(\boldsymbol{\Sigma})$ holds.

3.3 Block expression substitution method.

Assume again that a $\mathbf{J}_{qq}$ is identically singular. Let $\widehat{\mathbf{v}} \in \ker(\mathbf{J}_{qq})$, where $\widehat{\mathbf{v}} \not\equiv \mathbf{0}_{N_q}$. Similarly, we construct the column $n$-vector

$$\mathbf{v} = \begin{bmatrix} \mathbf{0}_{N_1+\cdots+N_{q-1}} \\ \widehat{\mathbf{v}} \\ \mathbf{0}_{N_{q+1}+\cdots+N_p} \end{bmatrix} .$$

We use notation similar to that used in the basic ES method (see [13, §4.2]):

$$\begin{aligned} &J = \left\{ j \mid v_j \not\equiv 0 \right\} \subseteq B_q, \quad M = \left\{ i \in B_q \mid d_j - c_i = \sigma_{ij} \text{ for some } j \in J \right\} , \\ &s = |J|, \quad \overline{c} = \max_{i \in M} c_i \quad \text{and} \quad \overline{J} = \left\{ l \mid v_l \text{ is (nonzero) constant} \right\} . \end{aligned} \tag{3.9}$$

The set $\overline{J}$ is used to seek a conversion that guarantees equivalence between the original DAE and the converted one. The conditions for applying the block ES method are

$$\begin{aligned} &\sigma(x_j, \mathbf{v}) \begin{cases} < d_j - \overline{c} & \text{if } j \in J \text{ or } \operatorname{blockOf}(j) < q \\ \le d_j - \overline{c} & \text{if } j \in B_q \setminus J \text{ or } \operatorname{blockOf}(j) > q, \end{cases} \quad \text{and} \\ &d_j - \overline{c} \ge 0 \quad \text{for all } j \in J . \end{aligned} \tag{3.10}$$

We choose an $l \in J$, and introduce $s-1$ new variables

$$y_j = x_j^{(d_j - \overline{c})} - \frac{v_j}{v_l} \cdot x_l^{(d_l - \overline{c})} \quad \text{for all } j \in J \setminus \left\{ l \right\} . \tag{3.11}$$

In each $f_i$ with $i \in B_q$, we

$$\begin{aligned} \text{replace each} \quad & x_j^{(\sigma_{ij})} \quad \text{with } d_j - c_i = \sigma_{ij} \text{ and } j \in J \setminus \left\{ l \right\} \\ \text{by} \quad & \left( y_j + \frac{v_j}{v_l} \cdot x_l^{(d_l - \overline{c})} \right)^{(\overline{c} - c_i)} . \end{aligned} \tag{3.12}$$

Note that because of $M$ in (3.9), we actually perform replacements (equivalently referred to as "expression substitutions") in only $f_i$'s with $i \in M \subseteq B_q$. Denote each new $f_i$ by $\overline{f}_i$, and let also $\overline{f}_i = f_i$ for the unchanged equations with $i \notin M$.

By (3.11), we append $s-1$ equations that prescribe the substitutions in (3.12):

$$0 = g_j = -y_j + x_j^{(d_j - \overline{c})} - \frac{v_j}{v_l} \cdot x_l^{(d_l - \overline{c})} \quad \text{for all } j \in J \setminus \left\{ l \right\} . \tag{3.13}$$

The block ES method is based on the following theorem.

**Theorem 3.2** *Let $J$, $s$, $M$, and $\overline{c}$ be as defined in (3.9). Assume that the conditions (3.10) hold. For an $l \in J$, if we*

*1) introduce $s-1$ new variables $x_j$, $j \in J \setminus \left\{ l \right\}$, as defined in (3.11),*
*2) perform replacements in $f_i$, for all $i \in B_q$, as described in (3.12), and*
*3) append $s-1$ equations $g_j$, $j \in J \setminus \left\{ l \right\}$, as defined in (3.13),*

*then* $\operatorname{Val}(\overline{\boldsymbol{\Sigma}}) < \operatorname{Val}(\boldsymbol{\Sigma})$, *where $\overline{\boldsymbol{\Sigma}}$ is the signature matrix of the resulting DAE.*

Before proving this theorem, we illustrate the block ES method with the robot arm DAE (3.4) and prove two related lemmas.

*Example 3.2* The method finds first $\widehat{\mathbf{v}} = [1,1]^T \in \ker(\mathbf{J}_{22})$. Then $\mathbf{v} = [0,0,1,1,0]^T$. Using (3.9), we have

$$J = \overline{J} = \{\, j \mid v_j \not\equiv 0 \,\} = \{2,3\}, \quad s = |J| = 2, \quad M = \{2,3\}, \quad \overline{c} = \max_{i\in M} c_i = 0\,.$$

Since $\mathbf{v}$ is constant, $J = \overline{J}$ and the first condition in (3.10) holds. The second condition in (3.10) holds also, as $d_4 - \overline{c} = d_5 - \overline{c} = 0$. We choose $x_4$, whose column index in the permuted $\boldsymbol{\Sigma}$ is $l = 2 \in \overline{J}$. Then we introduce for $x_5$, the other variable in block 2 with column index $j = 3$, a new variable

$$y_5 = x_5^{(d_5-\overline{c})} - \frac{v_3}{v_2} \cdot x_4^{(d_4-\overline{c})} = x_5 - x_4\,.$$

Correspondingly, we append $0 = g_5 = -y_5 + x_5 - x_4$ and replace $x_5$ by $y_5 + x_4$ in $C$ and $A$, the equations in block 2.

The resulting DAE has the following new equations

$$\begin{aligned}
0 = \overline{A} \;&= x_1'' - \Big[2c(x_3)(x_1' + x_3')^2 + x_1'^2 d(x_3) + (2x_3 - x_2)\big(a(x_3) + 2b(x_3)\big) + a(x_3)y_5\Big] \\
0 = \overline{C} \;&= x_3'' - \Big[-2c(x_3)(x_1' + x_3')^2 - x_1'^2 d(x_3) + (2x_3 - x_2)\big(a(x_3) - 9b(x_3)\big) \\
&\qquad - 2x_1'^2 c(x_3) - d(x_3)\big(x_1' + x_3'\big)^2 - \big(a(x_3) + b(x_3)\big)y_5\Big] \\
0 = g_5 \;&= -y_5 + x_4 - x_5\,.
\end{aligned}$$

$$\overline{\boldsymbol{\Sigma}} = \begin{array}{c} \\ g_5 \\ B \\ \overline{C} \\ \overline{A} \\ D \\ E \\ d_j \end{array}
\begin{array}{c}
\begin{array}{cccccccc} x_4 & x_5 & x_2 & y_5 & x_1 & x_3 & c_i \end{array} \\
\left[\begin{array}{cccccc}
0^{\bullet} & 0 & & 0 & & & 0 \\
0 & 0^{\bullet} & 2 & & 1 & 1 & 0 \\
 & & 0^{\bullet} & 0 & 1 & 2 & 2 \\
 & & 0 & 0^{\bullet} & 2 & 1 & 2 \\
 & & & & 0^{\bullet} & 0 & 4 \\
 & & & & 0 & 0^{\bullet} & 4
\end{array}\right] \\
\begin{array}{cccccc} 0 & 0 & 2 & 2 & 4 & 4 \end{array}
\end{array}
\qquad
\overline{\mathbf{J}} = \begin{array}{c} \\ g_5 \\ B \\ \overline{C}'' \\ \overline{A}'' \\ D^{(4)} \\ E^{(4)} \end{array}
\begin{array}{c}
\begin{array}{cccccc} x_4 & x_5 & x_2'' & y_5'' & x_1^{(4)} & x_3^{(4)} \end{array} \\
\left[\begin{array}{cccccc}
-1 & 1 & & & & \\
a_3 & -a_3 - 1 & 1 & & & \\
 & & a_3 - 9b_3 & -a_3 - b_3 & & 1 \\
 & & a_3 + 2b_3 & a_3 & 1 & \\
 & & & & \frac{\partial D}{\partial x_1} & \frac{\partial D}{\partial x_3} \\
 & & & & \frac{\partial E}{\partial x_1} & \frac{\partial E}{\partial x_3}
\end{array}\right]
\end{array}$$

Now the System Jacobian $\overline{\mathbf{J}}$ is generically nonsingular. The SA reports the correct index 5, and succeeds at any point where $\det(\overline{\mathbf{J}}) = 2(a_3^2 - 3a_3b_3 + b_3^2)\sin x_3 \neq 0$. Again $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) = 0 < 2 = \mathrm{Val}(\boldsymbol{\Sigma})$.

In [8], Pryce fixed the SA's failure on (3.4), and pointed out that *only* the introduction of $x_4 - x_5$ as a separate variable is essential to his fix. Example 3.2 verifies Pryce's argument and shows that the block ES method finds his reformulation in a systematic way.

To prove Theorem 3.2, we shall use the following two assumptions.

(a) Without loss of generality, we assume the entries $\widehat{v}_j \not\equiv 0$ are in the first $s$ positions of $\widehat{\mathbf{v}}$, that is, $\widehat{\mathbf{v}} = [\widehat{v}_1, \ldots, \widehat{v}_s, 0, \ldots, 0]^T$. By (3.9), $J = \sum_{m=1}^{q-1} N_m + 1 : \sum_{m=1}^{q-1} N_m + s$.

(b) We introduce one more variable $y_l = x_l^{(d_l-\overline{c})}$ for the chosen $l \in J$, and append correspondingly one more equation $0 = g_l = -y_l + x_l^{(d_l-\overline{c})}$.

We show first that the signature matrix $\overline{\boldsymbol{\Sigma}}$ of the resulting DAE can be put in the block structure as shown in Figure 3.1. Then we construct two $(n+s)$-vectors $\widetilde{\mathbf{c}}$ and $\widetilde{\mathbf{d}}$ in (3.14), and prove in Lemma 3.2 that $\widetilde{d}_j - \widetilde{c}_i > \overline{\sigma}_{ij}$ holds below the block diagonal, while $\widetilde{d}_j - \widetilde{c}_i \geq \overline{\sigma}_{ij}$ holds elsewhere. Lastly, we prove Theorem 3.2.

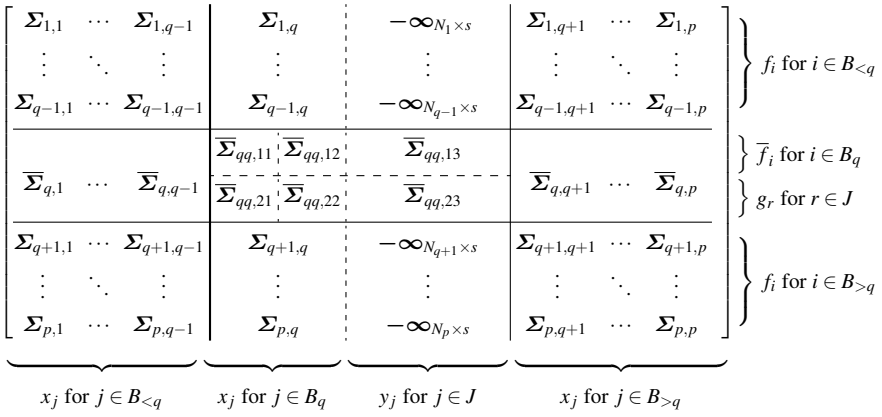



Fig. 3.1: Block structure of $\overline{\boldsymbol{\Sigma}}$ of the resulting DAE by the block ES method. The notation $B_{<q}$ is short for $\cup_{m=1}^{q-1} B_w$, and $B_{>q}$ is short for $\cup_{m=q+1}^{p} B_m$.

From the description of the conversion in Theorem 3.2, the substitutions (3.12) only occur in equations $f_i$ with $i \in B_q$. Hence, in the resulting DAE, variables $y_j$ for $j \in J$ only appear in $\overline{f}_i$ for $i \in B_q$ and $g_r$ for $r \in J$.

Considering the block structure of $\overline{\boldsymbol{\Sigma}}$ in Figure 3.1, we distinguish between the four cases for a block submatrix $\overline{\boldsymbol{\Sigma}}_{m_1 m_2}$: (a) $m_1 \neq q$ and $m_2 \neq q$, (b) $m_1 \neq q$ and $m_2 = q$, (c) $m_1 = q$ and $m_2 \neq q$, and (d) $m_1 = m_2 = q$.

(a) $m_1 \neq q$ and $m_2 \neq q$. In $\overline{\boldsymbol{\Sigma}}_{m_1 m_2}$, equations $f_i$ are of indices $i \in B_{<q} \cup B_{>q}$. As noted in (3.9), the expression substitutions described in (3.12) only take place in $f_{i'}$ with $i' \in M \subseteq B_q$, so do not happen in such blocks $\overline{\boldsymbol{\Sigma}}_{m_1 m_2}$. Hence, each $\boldsymbol{\Sigma}_{m_1 m_2}$ remains unchanged in $\overline{\boldsymbol{\Sigma}}$: $\overline{\boldsymbol{\Sigma}}_{m_1 m_2} = \boldsymbol{\Sigma}_{m_1 m_2}$ for $m_1 \neq q$ and $m_2 \neq q$.

(b) $m_1 \neq q$ and $m_2 = q$. In $\overline{\boldsymbol{\Sigma}}_{m_1 q}$, we include variables $y_j$ for $j \in J$ as defined in (3.11). By the same arguments as in (a), the expression substitutions do not happen in these blocks. That is, $y_j$ for $j \in J$ do not appear in equations $f_i$ for $i \in B_{<q} \cup B_{>q}$. Hence, we can obtain $\overline{\boldsymbol{\Sigma}}_{m_1 q}$ by concatenating horizontally $\boldsymbol{\Sigma}_{m_1 q}$ with an $N_{m_1} \times s$ matrix of $-\infty$'s: $\overline{\boldsymbol{\Sigma}}_{m_1 q} = \left[\boldsymbol{\Sigma}_{m_1 q},\ -\boldsymbol{\infty}_{N_{m_1} \times s}\right]$ for $m_1 = 1, \ldots, q-1, q+1, \ldots, p$.

(c) $m_1 = q$ and $m_2 \neq q$. In $\overline{\boldsymbol{\Sigma}}_{q m_2}$, we include equations $g_r$ for $r \in J$ as defined in (3.13). Also, due to the expression substitutions (3.12) occurring in $f_i$ with $i \in M \subseteq B_q$, $\boldsymbol{\sigma}(x_j, f_i)$ and $\boldsymbol{\sigma}\left(x_j, \overline{f}_i\right)$ may not be the same for all $i \in B_q$ and all $j = 1:n$.

Hence, in contrast to cases (a) and (b), there are no obvious connections between $\boldsymbol{\Sigma}_{m_1 m_2}$ and $\overline{\boldsymbol{\Sigma}}_{m_1 m_2}$ for $m_1 = q$ and $m_2 \neq q$.

(d) $m_1 = m_2 = q$. $\overline{\boldsymbol{\Sigma}}_{qq}$ contains signature entries for equations $\overline{f}_i$ and $g_r$, where $i \in B_q$ and $r \in J$, in variables $x_j$ and $y_r$, where $j \in B_q$ and $r \in J$. Similar to the resulting signature matrix $\overline{\boldsymbol{\Sigma}}$ by the basic ES method [13, §4.2], $\overline{\boldsymbol{\Sigma}}_{qq}$ by the block ES method also has a (sub)block structure; c.f (A.11). We shall use it in the proof of Lemma 3.2 in Appendix A.

Let $Q = \sum_{m=1}^{q} N_m$ denote the total number of equations (or variables) in the first $q$ blocks. Using a valid $(\mathbf{c};\mathbf{d})$ of $\boldsymbol{\Sigma}$, we construct an offset pair $(\widetilde{\mathbf{c}};\widetilde{\mathbf{d}})$ defined as

$$\widetilde{c}_i = \begin{cases} c_i & \text{if } i = 1:Q \\ \overline{c} & \text{if } i = Q+1:Q+s \\ c_{i-s} & \text{if } i = Q+s+1:n+s, \end{cases} \qquad \widetilde{d}_j = \begin{cases} d_j & \text{if } j = 1:Q \\ \overline{c} & \text{if } j = Q+1:Q+s \\ d_{j-s} & \text{if } j = Q+s+1:n+s\,. \end{cases} \tag{3.14}$$

Then we have the following lemma. Since its proof is rather technical, we present it in Appendix A.

**Lemma 3.2** *Let $(\widetilde{\mathbf{c}};\widetilde{\mathbf{d}})$ be as constructed in (3.14). In the block structure of $\overline{\boldsymbol{\Sigma}}$ in Figure 3.1, if a position $(i,j)$ in $\overline{\boldsymbol{\Sigma}}$ is below the block diagonal, then $\widetilde{d}_j - \widetilde{c}_i > \overline{\sigma}_{ij}$; otherwise, $\widetilde{d}_j - \widetilde{c}_i \geq \overline{\sigma}_{ij}$.*

Using this lemma, we can now prove Theorem 3.2.

*Proof* Let $\overline{T}$ be a transversal of $\overline{\boldsymbol{\Sigma}}$. Using Lemma 3.2 and (3.14), we derive

$$\begin{aligned} \mathrm{Val}(\overline{\boldsymbol{\Sigma}}) &= \sum_{(i,j)\in\overline{T}} \overline{\sigma}_{ij} \leq \sum_{(i,j)\in\overline{T}} (\widetilde{d}_j - \widetilde{c}_i) = \sum_{j=1}^{n+s} \widetilde{d}_j - \sum_{i=1}^{n+s} \widetilde{c}_i \\ &= \left( \sum_{j=1}^{Q} d_j + s\overline{c} + \sum_{j=Q+s+1}^{n+s} d_{j-s} \right) - \left( \sum_{j=1}^{Q} c_i + s\overline{c} + \sum_{i=Q+s+1}^{n+s} c_{i-s} \right) \\ &= \sum_{j=1}^{n} d_j - \sum_{i=1}^{n} c_i = \mathrm{Val}(\boldsymbol{\Sigma})\,. \end{aligned}$$

As in the proof of Theorem 3.1, we prove $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) < \mathrm{Val}(\boldsymbol{\Sigma})$ by contradiction. Assume $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) = \sum_j^{n+s} \widetilde{d}_j - \sum_i^{n+s} \widetilde{c}_i = \mathrm{Val}(\boldsymbol{\Sigma})$, which is $\geq 0$. With Lemma 3.2, the three conditions in Lemma 2.3 are satisfied. Then it follows from Lemma 2.3 that

(a) $(\widetilde{\mathbf{c}};\widetilde{\mathbf{d}})$ is a valid offset pair of $\overline{\boldsymbol{\Sigma}}$;

(b) the Jacobian pattern $\overline{\mathbf{S}}_0$, derived from $\overline{\boldsymbol{\Sigma}}$ and $(\widetilde{\mathbf{c}};\widetilde{\mathbf{d}})$, is in the $p \times p$ BTF shown in Figure 3.1;

(c) $\overline{T}$ is the union of HVTs $\overline{T}_m$ of all diagonal blocks $\overline{\boldsymbol{\Sigma}}_{11}, \ldots, \overline{\boldsymbol{\Sigma}}_{pp}$ of $\overline{\boldsymbol{\Sigma}}$.

We can consider block $q$ of the original DAE as a sub-DAE, with signature matrix $\boldsymbol{\Sigma}_{qq}$ and offset pair $(\mathbf{c}_q;\mathbf{d}_q)$—this follows from Lemma 2.4. The conversion described in Theorem 3.2 can be regarded as an application of the basic ES method to this sub-DAE, given that the ES conditions [13, (4.11)] hold because of (3.10). By [13, Theorem 4.2] for the basic ES method, a conversion results in $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}_{qq}) < \mathrm{Val}(\boldsymbol{\Sigma}_{qq})$. Also, since $\overline{\boldsymbol{\Sigma}}_{mm} = \boldsymbol{\Sigma}_{mm}$ for $m \neq q$, $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}_{mm}) = \mathrm{Val}(\boldsymbol{\Sigma}_{mm})$. Then a contradiction $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) < \mathrm{Val}(\boldsymbol{\Sigma})$ follows by the same derivations as in (3.8). The assumption $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) = \mathrm{Val}(\boldsymbol{\Sigma})$ is hence false, so $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) < \mathrm{Val}(\boldsymbol{\Sigma})$ holds. □

## 4 More examples.

We demonstrate how to apply the block conversion methods on two DAE problems originated from electrical circuit analysis [3]. They are the transistor amplifier and the ring modulator. We describe them in §4.1 and §4.2, respectively.

### 4.1 Transistor amplifier.

The transistor amplifier DAE is

$$
\begin{aligned}
0 = f_1 &= \quad C_1(x_1' - x_2') + R_0^{-1}(x_1 - U_e(t)) \\
0 = f_2 &= -C_1(x_1' - x_2') - R_2^{-1}U_b + x_2\left(R_1^{-1} + R_2^{-1}\right) - (\alpha - 1)g(x_2 - x_3) \\
0 = f_3 &= \quad C_2 x_3' - g(x_2 - x_3) + R_3^{-1}x_3 \\
0 = f_4 &= \quad C_3(x_4' - x_5') + R_4^{-1}(x_4 - U_b) + \alpha g(x_2 - x_3) \\
0 = f_5 &= -C_3(x_4' - x_5') - R_0^{-1}U_b + x_5\left(R_5^{-1} + R_6^{-1}\right) - (\alpha - 1)g(x_5 - x_6) \\
0 = f_6 &= \quad C_4 x_6' - g(x_5 - x_6) + R_7^{-1}x_6 \\
0 = f_7 &= \quad C_5(x_7' - x_8') + R_8^{-1}(x_7 - U_b) + \alpha g(x_5 - x_6) \\
0 = f_8 &= -C_5(x_7' - x_8') + R_9^{-1}x_8 \,.
\end{aligned}
$$

$\boldsymbol{\Sigma} =$

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $c_i$ |
|---|---|---|---|---|---|---|---|---|---|
| $f_1$ | $1^\bullet$ | 1 | | | | | | | 0 |
| $f_2$ | 1 | $1^\bullet$ | 0 | | | | | | 0 |
| $f_3$ | 0 | | $1^\bullet$ | | | | | | 0 |
| $f_4$ | 0 | 0 | | $1^\bullet$ | 1 | | | | 0 |
| $f_5$ | | | | 1 | $1^\bullet$ | 0 | | | 0 |
| $f_6$ | | | | 0 | | $1^\bullet$ | | | 0 |
| $f_7$ | | | | 0 | 0 | | $1^\bullet$ | 1 | 0 |
| $f_8$ | | | | | | | 1 | $1^\bullet$ | 0 |
| $d_j$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |

$\mathbf{J} =$

| | $x_1'$ | $x_2'$ | $x_3'$ | $x_4'$ | $x_5'$ | $x_6'$ | $x_7'$ | $x_8'$ | $c_i$ |
|---|---|---|---|---|---|---|---|---|---|
| $f_1$ | $C_1$ | $-C_1$ | | | | | | | 0 |
| $f_2$ | $-C_1$ | $C_1$ | | | | | | | 0 |
| $f_3$ | | | $C_2$ | | | | | | 0 |
| $f_4$ | | | | $C_3$ | $-C_3$ | | | | 0 |
| $f_5$ | | | | $-C_3$ | $C_3$ | | | | 0 |
| $f_6$ | | | | | | $C_4$ | | | 0 |
| $f_7$ | | | | | | | $C_5$ | $-C_5$ | 0 |
| $f_8$ | | | | | | | $-C_5$ | $C_5$ | 0 |

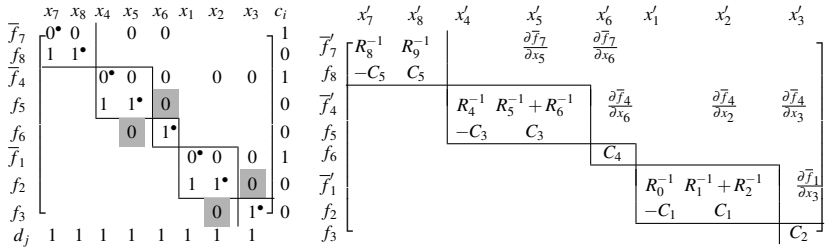
In the equations, $g(y) = \beta\left(e^{y/U_F} - 1\right)$ and $U_e(t) = 0.1\sin(200\pi t)$; $\alpha, \beta, U_b, U_F, R_0$, $R_k$ for $k = 1{:}9$, and $C_k$ for $k = 1{:}5$ are positive constants. The SA fails since $\mathbf{J}$ is identically singular. The fine BTF reveals that the three $2 \times 2$ sub-Jacobians $\mathbf{J}_{11}$, $\mathbf{J}_{33}$, $\mathbf{J}_{55}$ are identically singular and have a similar structure. Each block receives the same treatment when a conversion method is applied.

*LC method.* One can easily find $\widehat{\mathbf{u}} = [1,1]^T \in \operatorname{coker}(\mathbf{J}_{11}), \operatorname{coker}(\mathbf{J}_{33}), \operatorname{coker}(\mathbf{J}_{55})$. We perform on each singular block a conversion, and choose to replace the first equation in each such block.

| block | replace | by |
|---|---|---|
| 1 | $f_1$ | $\overline{f}_1 = f_1 + f_2$ |
| 3 | $f_4$ | $\overline{f}_4 = f_4 + f_5$ |
| 5 | $f_7$ | $\overline{f}_7 = f_7 + f_8$ |

The new equations in the resulting DAE are

$$
\begin{aligned}
0 &= \overline{f}_1 = R_0^{-1}(x_1 - U_e(t)) - R_2^{-1}U_b + x_2\left(R_1^{-1} + R_2^{-1}\right) - (\alpha - 1)g(x_2 - x_3) \\
0 &= \overline{f}_4 = R_4^{-1}(x_4 - U_b) + \alpha g(x_2 - x_3) - R_5^{-1}U_b + x_5\left(R_5^{-1} + R_6^{-1}\right) \\
&\qquad - (\alpha - 1)g(x_5 - x_6) \\
0 &= \overline{f}_7 = R_8^{-1}(x_7 - U_b) + \alpha g(x_5 - x_6) + R_9^{-1}x_8 \ .
\end{aligned}
$$




The SA still reports index 1, and succeeds with a nonzero constant $\det(\overline{\mathbf{J}})$:

$$
\det(\overline{\mathbf{J}}) = C_1C_2C_3C_4C_5\left(R_0^{-1} + R_1^{-1} + R_2^{-1}\right)\left(R_4^{-1} + R_5^{-1} + R_6^{-1}\right)\left(R_8^{-1} + R_9^{-1}\right) \neq 0 \ .
$$

Now $\mathrm{Val}(\overline{\mathbf{\Sigma}}) = 5 < 8 = \mathrm{Val}(\mathbf{\Sigma})$.

*ES method.* We can take $\widehat{\mathbf{v}} = [1,1]^T \in \ker(\mathbf{J}_{11}), \ker(\mathbf{J}_{33}), \ker(\mathbf{J}_{55})$. We show how to perform a conversion on block 1; block 3 and block 5 can be treated in the same way.

For block 1, we construct the corresponding $\mathbf{v} = [1,1,\mathbf{0}_8^T]^T$. Using (3.9), we have

$$
J = \overline{J} = \{\, j \mid v_j \not\equiv 0 \,\} = \{1,2\}, \quad s = |J| = 2, \quad M = \{1,2\}, \quad \text{and} \quad \overline{c} = 0 \ .
$$

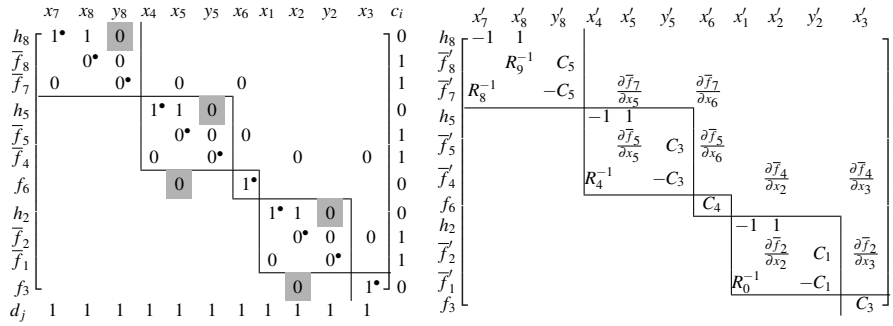
We choose $l = 1 \in \overline{J}$, introduce for $x_2$ a new variable

$$
y_2 = x_2^{(d_2 - \overline{c})} - \frac{v_2}{v_1} \cdot x_1^{(d_1 - \overline{c})} = x_2' - x_1' \ ,
$$

and append correspondingly the equation $0 = h_2 = -y_2 + x_2' - x_1'$. Then we replace $x_2'$ by $y_2 + x_1'$ in $f_1, f_2$.

After we complete similar conversions on block 3 and block 5, the resulting DAE has equations $f_3, f_6$ and the following equations:

$$
\begin{aligned}
0 &= \overline{f}_1 = -C_1y_2 + R_0^{-1}(x_1 - U_e(t)) \\
0 &= \overline{f}_2 = \ \ C_1y_2 - R_2^{-1}U_b + x_2\left(R_1^{-1} + R_2^{-1}\right) - (\alpha - 1)g(x_2 - x_3) \\
0 &= h_2 = -y_2 + x_2' - x_1' \\
0 &= \overline{f}_4 = -C_3y_5 + R_4^{-1}(x_4 - U_b) + \alpha g(x_2 - x_3) \\
0 &= \overline{f}_5 = \ \ C_3y_5 - R_5^{-1}U_b + x_5\left(R_5^{-1} + R_6^{-1}\right) - (\alpha - 1)g(x_5 - x_6) \\
0 &= h_5 = -y_5 + x_5' - x_4' \\
0 &= \overline{f}_7 = -C_5y_8 + R_8^{-1}(x_7 - U_b) + \alpha g(x_5 - x_6) \\
0 &= \overline{f}_8 = \ \ C_5y_8 + R_9^{-1}x_8 \\
0 &= h_8 = -y_8 + x_8' - x_7' \ .
\end{aligned}
$$




Here in $\overline{\mathbf{J}}$, $\partial\overline{f}_5/\partial x_5 = R_5^{-1} + R_6^{-1}$ and $\partial\overline{f}_2/\partial x_2 = R_1^{-1} + R_2^{-1}$. The SA succeeds with a nonzero constant $\det(\overline{\mathbf{J}})$ and $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) = 5 < 8 = \mathrm{Val}(\boldsymbol{\Sigma})$.

## 4.2 Ring modulator.

We study the ring modulator problem from [3]. When $C_s \neq 0$, it is a stiff ODE system of 15 nonlinear equations. Setting $C_s = 0$ gives a DAE of differentiation index 2 that consists of 11 differential and 4 algebraic equations:

$$
\begin{aligned}
0 = f_1 &= -x_1' + C^{-1}\left(x_8 - 0.5x_{10} + 0.5x_{11} + x_{14} - R^{-1}x_1\right)\\
0 = f_2 &= -x_2' + C^{-1}\left(x_9 - 0.5x_{11} + 0.5x_{12} + x_{15} - R^{-1}x_2\right)\\
0 = f_3 &= x_{10} - q(U_{D1}) + q(U_{D4})\\
0 = f_4 &= -x_{11} + q(U_{D2}) - q(U_{D3})\\
0 = f_5 &= x_{12} + q(U_{D1}) - q(U_{D3})\\
0 = f_6 &= -x_{13} - q(U_{D2}) + q(U_{D4})\\
0 = f_7 &= -x_7' + C_p^{-1}\left(-R_p^{-1}x_7 + q(U_{D1}) + q(U_{D2}) - q(U_{D3}) - q(U_{D4})\right)\\
0 = f_8 &= -x_8' + -L_h^{-1}x_1\\
0 = f_9 &= -x_9' + -L_h^{-1}x_2\\
0 = f_{10} &= -x_{10}' + L_{s2}^{-1}(0.5x_1 - x_3 - R_{g2}x_{10})\\
0 = f_{11} &= -x_{11}' + L_{s3}^{-1}(-0.5x_1 + x_4 - R_{g3}x_{11})\\
0 = f_{12} &= -x_{12}' + L_{s2}^{-1}(0.5x_2 - x_5 - R_{g2}x_{12})\\
0 = f_{13} &= -x_{13}' + L_{s3}^{-1}(-0.5x_2 + x_6 - R_{g3}x_{13})\\
0 = f_{14} &= -x_{14}' + L_{s1}^{-1}(-x_1 + U_{in1}(t) - (R_i + R_{g1})x_{14})\\
0 = f_{15} &= -x_{15}' + L_{s1}^{-1}(-x_2 - (R_c + R_{g1})x_{15})\,.
\end{aligned}
$$

The functions are

$$\begin{aligned} U_{D1} &= \phantom{-}x_3 - x_5 - x_7 - U_{in2}(t) \qquad & q(U) &= \gamma(e^{\delta U} - 1) \\ U_{D2} &= -x_4 + x_6 - x_7 - U_{in2}(t) & U_{in1}(t) &= 0.5\sin(2000\pi t) \\ U_{D3} &= \phantom{-}x_4 + x_5 + x_7 + U_{in2}(t) & U_{in2}(t) &= 2\sin(20000\pi t) \\ U_{D4} &= -x_3 - x_6 + x_7 + U_{in2}(t)\,. \end{aligned}$$

We refer the reader to [3] for the nonzero constants $C$, $C_p$, $R$, $R_p$, $R_c$, $\gamma$, $L_h$, $L_{s1}$, $L_{s2}$, $L_{s3}$, $R_{g1}$, $R_{g2}$, $R_{g3}$, $R_i$, and $\delta$.

$\boldsymbol{\Sigma} =$

| | $x_1$ | $x_2$ | $x_7$ | $x_{13}$ | $x_{11}$ | $x_{12}$ | $x_{10}$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_8$ | $x_9$ | $x_{14}$ | $x_{15}$ | $c_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $f_1$ | $1^\bullet$ | | | | 0 | | 0 | | | | | 0 | | 0 | | 0 |
| $f_2$ | | $1^\bullet$ | | 0 | | 0 | | | | | | | 0 | | 0 | 0 |
| $f_7$ | | | $1^\bullet$ | | | | | 0 | 0 | 0 | 0 | | | | | 0 |
| $f_{13}$ | | 0 | | $1^\bullet$ | | | | | | | 0 | | | | | 0 |
| $f_{11}$ | 0 | | | | $1^\bullet$ | | | | 0 | | | | | | | 0 |
| $f_{12}$ | | 0 | | | | $1^\bullet$ | | | | 0 | | | | | | 0 |
| $f_{10}$ | 0 | | | | | | $1^\bullet$ | 0 | | | | | | | | 0 |
| $f_3$ | | | 0 | | | | 0 | $0^\bullet$ | | 0 | 0 | | | | | 0 |
| $f_4$ | | | 0 | | 0 | | | | $0^\bullet$ | 0 | 0 | | | | | 0 |
| $f_5$ | | | 0 | | | 0 | | 0 | 0 | $0^\bullet$ | | | | | | 0 |
| $f_6$ | | | 0 | 0 | | | | 0 | 0 | | $0^\bullet$ | | | | | 0 |
| $f_8$ | 0 | | | | | | | | | | | $1^\bullet$ | | | | 0 |
| $f_9$ | | 0 | | | | | | | | | | | $1^\bullet$ | | | 0 |
| $f_{14}$ | 0 | | | | | | | | | | | | | $1^\bullet$ | | 0 |
| $f_{15}$ | | 0 | | | | | | | | | | | | | $1^\bullet$ | 0 |
| $d_j$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | |

Each block of size 1 has a nonsingular sub-Jacobian $-1$. Block 8 has an identically singular sub-Jacobian

$$\mathbf{J}_{88} = \begin{array}{c} \\ f_3 \\ f_4 \\ f_5 \\ f_6 \end{array} \begin{array}{c} \begin{array}{cccc} x_3 & x_4 & x_5 & x_6 \end{array} \\ \begin{bmatrix} -s_1 - s_4 & & s_1 & -s_4 \\ & -s_2 - s_3 & -s_3 & s_2 \\ s_1 & -s_3 & -s_1 - s_3 & \\ -s_4 & s_2 & & -s_2 - s_4 \end{bmatrix} \end{array}, \quad \text{where} \quad s_i = \gamma\delta e^{\delta U_{Di}}\,.$$

This is a nonlinear block, since variables $x_3, x_4, x_5, x_6$ do not occur jointly linearly in equations $f_3, f_4, f_5, f_6$. One can also see these variables appear in $\mathbf{J}_{88}$.

*LC method.* We find a constant vector $\widehat{\mathbf{u}} = [1, -1, 1, -1]^T \in \operatorname{coker}(\mathbf{J}_{88})$, which satisfies the block LC condition (3.3). Then $\mathbf{u} = [\mathbf{0}_7^T, 1, -1, 1, -1, \mathbf{0}_4^T]^T$. We use (3.2) to derive

$$I = \{\, i \mid u_i \not\equiv 0 \,\} = \{8, 9, 10, 11\}, \quad \underline{c} = 0, \quad \text{and} \quad L = \overline{L} = \{8, 9, 10, 11\}\,.$$

The row indices in $\overline{L}$ correspond to the equations $f_3, f_4, f_5, f_6$. We can pick any one of them and replace it by

$$\overline{f} = u_1 f_3 + u_2 f_4 + u_3 f_5 + u_4 f_6 = f_3 - f_4 + f_5 - f_6 = x_{10} + x_{11} + x_{12} + x_{13}\,.$$

We choose $f_3$ and replace it by $\overline{f}_3 = \overline{f}$. The resulting DAE has the following $\overline{\boldsymbol{\Sigma}}$ with $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) = 10 < 11 = \mathrm{Val}(\boldsymbol{\Sigma})$.

$\overline{\boldsymbol{\Sigma}} =$

| | $x_1$ | $x_2$ | $x_7$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_{10}$ | $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_8$ | $x_9$ | $x_{14}$ | $x_{15}$ | $c_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $f_1$ | $1^\bullet$ | | | | | | | 0 | 0 | | | 0 | | 0 | | 0 |
| $f_2$ | | $1^\bullet$ | | | | | | | | 0 | 0 | | 0 | | 0 | 0 |
| $f_7$ | | | $1^\bullet$ | 0 | 0 | 0 | 0 | | | | | | | | | 0 |
| $f_{10}$ | 0 | | | $0^\bullet$ | | | | 1 | | | | | | | | 0 |
| $f_5$ | | | 0 | 0 | $0^\bullet$ | 0 | | | | 0 | | | | | | 0 |
| $f_4$ | | | 0 | | 0 | $0^\bullet$ | 0 | | 0 | | | | | | | 0 |
| $f_6$ | | | 0 | 0 | 0 | | $0^\bullet$ | | | | 0 | | | | | 0 |
| $\overline{f}_3$ | | | | | | | | $0^\bullet$ | 0 | 0 | 0 | | | | | 1 |
| $f_{11}$ | 0 | | | | 0 | | | | $1^\bullet$ | | | | | | | 0 |
| $f_{12}$ | | 0 | | | | 0 | | | | $1^\bullet$ | | | | | | 0 |
| $f_{13}$ | | 0 | | | | | 0 | | | | $1^\bullet$ | | | | | 0 |
| $f_8$ | 0 | | | | | | | | | | | $1^\bullet$ | | | | 0 |
| $f_9$ | | 0 | | | | | | | | | | | $1^\bullet$ | | | 0 |
| $f_{14}$ | 0 | | | | | | | | | | | | | $1^\bullet$ | | 0 |
| $f_{15}$ | | 0 | | | | | | | | | | | | | $1^\bullet$ | 0 |
| $d_j$ | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |

Again, each $1\times 1$ block has a nonsingular Jacobian: $\partial f_i/\partial x_i' = -1$ for $i = 1, 2, 7, 8, 9, 14, 15$. The sub-Jacobian of block 4 in the resulting DAE is

$$\overline{\mathbf{J}}_{44} = \begin{array}{c} \\ f_{10} \\ f_5 \\ f_4 \\ f_6 \\ \overline{f}_3 \\ f_{11} \\ f_{12} \\ f_{13} \end{array} \begin{array}{c} \begin{array}{cccccccc} x_3 & x_4 & x_5 & x_6 & x_{10}' & x_{11}' & x_{12}' & x_{13}' \end{array} \\ \begin{bmatrix} -L_{s2}^{-1} & & & & -1 & & & \\ s_1 & -s_3 & -s_1-s_3 & & & & & \\ & -s_2-s_3 & -s_3 & s_2 & & & & \\ -s_4 & s_2 & & -s_2-s_4 & & & & \\ & & & & 1 & 1 & 1 & 1 \\ & L_{s3}^{-1} & & & & -1 & & \\ & & -L_{s2}^{-1} & & & & -1 & \\ & & & L_{s3}^{-1} & & & & -1 \end{bmatrix} \end{array},$$

whose determinant is $\det(\overline{\mathbf{J}}_{44}) = 2s_1s_2s_3s_4(s_1^{-1}+s_2^{-1}+s_3^{-1}+s_4^{-1})(L_{s2}^{-1}+L_{s3}^{-1})$. The SA succeeds at any point where $\det(\overline{\mathbf{J}}_{44}) \neq 0$, and the DAE is of index 2.

*ES method.* Find $\widehat{\mathbf{v}} = [-1,1,-1,1]^T \in \ker(\mathbf{J}_{88})$. Then $\mathbf{v} = [\mathbf{0}_7^T, -1, 1, -1, 1, \mathbf{0}_4^T]^T$. We use (3.9) to derive

$$J = \overline{J} = \{\, j \mid v_j \not\equiv 0 \,\} = \{8, 9, 10, 11\}, \quad s = |J| = 4, \quad M = J, \quad \text{and} \quad \overline{c} = 0\,.$$

We choose column index $l = 8 \in \overline{J}$ in the permuted $\boldsymbol{\Sigma}$. The variable of this column is $x_3$. The other variables in block 8 are $x_4$, $x_5$, $x_6$, so we introduce for them, respectively,

$$y_4 = x_4 - (v_9/v_8)\cdot x_3, \quad y_5 = x_5 - (v_{10}/v_8)\cdot x_3, \quad \text{and} \quad y_6 = x_6 - (v_{11}/v_8)\cdot x_3\ .$$

Then we append the equations corresponding to these variables

$$0 = g_4 = -y_4 + x_4 + x_3, \quad 0 = g_5 = -y_5 + x_5 - x_3, \quad \text{and} \quad 0 = g_6 = -y_6 + x_6 + x_3\ .$$

The equations in block 8 are $f_3$, $f_4$, $f_5$, $f_6$. In these equations, we perform the following substitutions.

| replace | by | in |
|---|---|---|
| $x_4$ | $y_4 - x_3$ | $f_4, f_5, f_6$ |
| $x_5$ | $y_5 + x_3$ | $f_3, f_4, f_5$ |
| $x_6$ | $y_6 - x_3$ | $f_3, f_4, f_6$ |

The resulting index-2 DAE is of size 18. (We do not display the results of SA here.) It has $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) = 10 < 11 = \mathrm{Val}(\boldsymbol{\Sigma})$ and $\det(\overline{\mathbf{J}}) = -2s_1s_2s_3s_4(s_1^{-1} + s_2^{-1} + s_3^{-1} + s_4^{-1})(L_{s2}^{-1} + L_{s3}^{-1})$. The largest fine block is of size 12, and the other six fine blocks are of size 1. The SA succeeds at any point where $\det(\overline{\mathbf{J}}) \neq 0$.

## 5 Conclusions.

We combined block triangularization with the LC and ES conversion methods for improving the $\Sigma$-method. When $\mathbf{J}$ is identically singular and the DAE has a nontrivial BTF, we can locate each diagonal block whose corresponding sub-Jacobian is identically singular, and perform a conversion on it. We base this strategy on the view that each diagonal block can be regarded as a sub-DAE, while formulas contributing to the off diagonal blocks are regarded as driving terms.

Compared with the basic conversion methods that work on the whole DAE, the block methods only work on singular blocks, which are usually smaller than the DAE itself. Hence the block methods require fewer symbolic computations, and can generally find a useful conversion for reducing $\mathrm{Val}(\boldsymbol{\Sigma})$ more efficiently. As in the basic case, a conversion applied on a singular block guarantees (a) a strict decrease in the value of the (whole) signature matrix, and (b) the equivalence between the original DAE and the resulting one. The rationale for choosing a desirable conversion method is in [13, Table 4.1].

We combine MATLAB's Symbolic Math Toolbox [14] with our structural analysis software DAESA [6,10], and have built a prototype code that automates the conversion process. We aim to incorporate them in a future version of DAESA.

With our prototype code, we have applied our methods on numerous DAEs on which the $\Sigma$-method fails. They are either arbitrarily constructed to be SA-failure cases for our investigations, or borrowed from the existing literature. Our conversion methods succeed in fixing all these solvable DAEs. We believe that our assumptions and conditions are reasonable for practical problems, and that these methods can help make the $\Sigma$-method more reliable.

We end these two articles with our main conjecture regarding SA's failure. In all our experiments, when we successfully fix the failure using our conversion methods, the value of a signature matrix always decreases. As Pryce points out in [8], the solvability of a DAE lies within its inherent nature, not the way it is formulated or analyzed. Hence we conjecture that a DAE formulation friendly to SA should have a reasonable but never overestimated $\mathrm{Val}(\boldsymbol{\Sigma})$ that can be interpreted as number of degrees of freedom (DOF) of the underlying problem. In other words, a DAE should not be formulated to exhibit more DOF than the underlying problem has. However, based on our current knowledge, it appears difficult to show why overestimating DOF can lead to an identically singular System Jacobian.

## A Proof of Lemma 3.2.

For $\overline{\boldsymbol{\Sigma}} = (\overline{\sigma}_{ij})$ in the block structure in Figure 3.1, we write the block sizes in the array

$$\overline{N} = (N_1, N_2, \ldots, N_{q-1}, N_q + s, N_{q+1}, \ldots, N_p) \,,$$

and also write the block sizes of $\boldsymbol{\Sigma}$ in the array $N = (N_1, N_2, \ldots, N_{q-1}, N_q, N_{q+1}, \ldots, N_p)$. Let $\overline{\mathrm{blockOf}}(i)$ denote the block number of a row or column index $i$ in $\overline{\boldsymbol{\Sigma}}$. From the construction of $\overline{N}$ and $N$, it is not difficult to show that

$$\overline{\mathrm{blockOf}}(j) < q \;\Leftrightarrow\; 1 \le j \le \sum_{w=1}^{q-1} N_w \;\Leftrightarrow\; \mathrm{blockOf}(j) < q \quad \text{and} \tag{A.1}$$

$$\overline{\mathrm{blockOf}}(j+s) > q \;\Leftrightarrow\; \sum_{w=1}^{q} N_w + s + 1 \le j + s \;\Leftrightarrow\; \mathrm{blockOf}(j) > q \,. \tag{A.2}$$

From the construction of $(\widetilde{\mathbf{c}}; \widetilde{\mathbf{d}})$ in (3.14), each variable $x_j$ for $j = 1{:}n$ has the same "variable offset" in $\overline{\boldsymbol{\Sigma}}$ as $x_j$ has in $\boldsymbol{\Sigma}$. Also, each equation $\overline{f}_i$ for $i = 1{:}n$ has the same "equation offset" in $\overline{\boldsymbol{\Sigma}}$ as $f_i$ has in $\boldsymbol{\Sigma}$. Quotation marks are used here because $(\widetilde{\mathbf{c}}; \widetilde{\mathbf{d}})$ is *not* a valid offset pair of $\overline{\boldsymbol{\Sigma}}$; this vector pair is merely used for proving $\mathrm{Val}(\overline{\boldsymbol{\Sigma}}) < \mathrm{Val}(\boldsymbol{\Sigma})$ in Theorem 3.2.

We aim to show that

$$\widetilde{d}_j - \widetilde{c}_i \begin{cases} > \overline{\sigma}_{ij} & \text{if } \overline{\mathrm{blockOf}}(j) < \overline{\mathrm{blockOf}}(i) \\ \ge \overline{\sigma}_{ij} & \text{if } \overline{\mathrm{blockOf}}(j) \ge \overline{\mathrm{blockOf}}(i) \,. \end{cases} \tag{A.3}$$

For the block structure of $\overline{\boldsymbol{\Sigma}}$ in Figure 3.1, we have shown on page 15 that $\overline{\boldsymbol{\Sigma}}_{m_1 m_2} = \boldsymbol{\Sigma}_{m_1 m_2}$ if $m_1 \ne q$ and $m_2 \ne q$, and that $\overline{\boldsymbol{\Sigma}}_{m_1 m_2} = \left[\boldsymbol{\Sigma}_{m_1 q},\ -\boldsymbol{\infty}_{N_{m_1} \times s}\right]$ if $m_1 \ne q$ and $m_2 = q$. Hence, provided $m_1 \ne q$, $\overline{\boldsymbol{\Sigma}}_{m_1 m_2}$ is below [resp. above] the block diagonal of $\overline{\boldsymbol{\Sigma}}$, if $\boldsymbol{\Sigma}_{m_1 m_2}$ is below [resp. above] the block diagonal of $\boldsymbol{\Sigma}$. By (2.4), the inequalities in (A.3) hold for $i$ with $\overline{\mathrm{blockOf}}(i) \ne q$.

What remains to show is the inequalities in (A.3) for $i$ with $\overline{\mathrm{blockOf}}(i) = q$. These inequalities are for the signature entries in $\overline{\boldsymbol{\Sigma}}_{q m_2}$, the blocks affected by the expression substitutions. We consider three cases for $\overline{\boldsymbol{\Sigma}}_{q m_2}$: it is (a) below the block diagonal, with $m_2 < q$, (b) above the block diagonal, with $m_2 > q$, or (c) the diagonal block $\overline{\boldsymbol{\Sigma}}_{qq}$, with $m_2 = q$.

(a) $\overline{\boldsymbol{\Sigma}}_{q m_2}$ with $m_2 < q$. An entry $(i, j)$ in this block satisfies $\overline{\mathrm{blockOf}}(j) < \overline{\mathrm{blockOf}}(i) = q$. By (A.1), $\mathrm{blockOf}(j) < q$ and hence $j \notin B_q$.

Recall from (3.12) that, in each $f_i$ with $i \in M \subseteq B_q$, we substitute $\left(y_r + \frac{v_r}{v_l} \cdot x_l^{(d_l - \overline{c})}\right)^{(\overline{c} - c_i)}$ for each $x_r^{(\sigma_{ir})}$ with $d_r - c_i = \sigma_{ir}$ and $r \in J \setminus \{\, l \,\} \subset B_q$. For a $j \notin B_q \supset J \setminus \{\, l \,\}$, the corresponding derivatives $x_j^{(d_j - c_i)}$ are not replaced in the ES conversion, and for $r \in J \setminus \{\, l \,\}$ (so $j, r, l$ are distinct),

$$\sigma\left(x_j, \left(y_r + \frac{v_r}{v_l} \cdot x_l^{(d_l - \overline{c})}\right)^{(\overline{c} - c_i)}\right) = \sigma\left(x_j, \left(\frac{v_r}{v_l}\right)^{(\overline{c} - c_i)}\right) \le \sigma\left(x_j, \mathbf{v}^{(\overline{c} - c_i)}\right) = \sigma\left(x_j, \mathbf{v}\right) + (\overline{c} - c_i) \,. \tag{A.4}$$

By (3.10), $\sigma\left(x_j,\mathbf{v}\right) < d_j - \overline{c}$. Using (2.4) and (A.4), we derive

$$\begin{aligned}\sigma\left(x_j,\overline{f}_i\right) &\le \max\left\{\sigma\left(x_j,f_i\right), \max_{r\in J\setminus\{l\}} \sigma\left(x_j,\left(y_r+\frac{v_r}{v_l}\cdot x_l^{(d_l-\overline{c})}\right)^{(\overline{c}-c_i)}\right)\right\}\\ &\le \max\left\{\sigma_{ij},\ \sigma\left(x_j,\mathbf{v}\right)+(\overline{c}-c_i)\right\}\\ &< \max\left\{d_j-c_i,\ (d_j-\overline{c})+(\overline{c}-c_i)\right\} = d_j-c_i \quad \text{for } i\in M\subseteq B_q\,. \end{aligned} \tag{A.5}$$

From the ES conversion described in Theorem 3.2, we have

$$\sigma\left(x_j,\overline{f}_i\right) = \sigma\left(x_j,f_i\right) < d_j-c_i \quad \text{for } i\in B_q\setminus M \text{ and} \tag{A.6}$$

$$\sigma\left(x_j,g_r\right) \le \sigma\left(x_j,v\right) < d_j-\overline{c} \quad \text{for } r\in J\,. \tag{A.7}$$

Since blocks $\overline{\mathbf{\Sigma}}_{qm_2}$ with $m_2 < q$ contain signature entries $\overline{\sigma}_{ij}$ for equations $\overline{f}_i$ and $g_r$, where $i\in B_q$ and $r\in J$, in variables $x_j$ with $\text{blockOf}(j) < q$, by taking together the inequalities in (A.5)-(A.7), we have

$$\overline{\sigma}_{ij} < \begin{cases} d_j-c_i & \text{if } \text{blockOf}(j) < q \text{ and } i\in B_q\\ d_j-\overline{c} & \text{if } \text{blockOf}(j) < q \text{ and } i\in Q+1:Q+s\,; \end{cases}$$

recall $Q=\sum_{w=1}^{q} N_w$. Using (A.1) and the construction of $(\widetilde{\mathbf{c}};\widetilde{\mathbf{d}})$ in (3.14), we have

$$\overline{\sigma}_{ij} < \widetilde{d}_j-\widetilde{c}_i \quad \text{for} \quad \overline{\text{blockOf}}(j) < \overline{\text{blockOf}}(i) = q\,.$$

($q$ is the block number of both the original and enlarged diagonal blocks.)

(b) $\overline{\mathbf{\Sigma}}_{qm_2}$ with $m_2 > q$. An entry $(i,j+s)$ in this block satisfies $\overline{\text{blockOf}}(j+s) > \overline{\text{blockOf}}(i) = q$. By (A.2), $\text{blockOf}(j) > q$ and hence $j\notin B_q \supset J\setminus\{l\}$. By the same arguments as in (a), the corresponding derivatives $x_j^{(d_j-c_i)}$ are not replaced in the ES conversion.

By (3.10), $\sigma\left(x_j,\mathbf{v}\right) \le d_j-\overline{c}$. Then by the same derivations as (A.5–A.7) in (a), we have

$$\sigma\left(x_j,\overline{f}_i\right) \le d_j-c_i \quad \text{for } i\in M\subseteq B_q\,, \tag{A.8}$$

$$\sigma\left(x_j,\overline{f}_i\right) = \sigma\left(x_j,f_i\right) \le d_j-c_i \quad \text{for } i\in B_q\setminus M, \text{ and} \tag{A.9}$$

$$\sigma\left(x_j,g_r\right) \le \sigma\left(x_j,v\right) \le d_j-\overline{c} \quad \text{for } r\in J\,. \tag{A.10}$$

Since blocks $\overline{\mathbf{\Sigma}}_{qm_2}$ with $m_2 > q$ contain signature entries $\overline{\sigma}_{i,j+s}$ for equations $\overline{f}_i$ and $g_r$, where $i\in B_q$ and $r\in J$, in variables $x_j$ with $\text{blockOf}(j) > q$, the inequalities (A.8–A.10) yield

$$\overline{\sigma}_{i,j+s} \le \begin{cases} d_j-c_i & \text{if } \text{blockOf}(j) > q \text{ and } i\in B_q\\ d_j-\overline{c} & \text{if } \text{blockOf}(j) > q \text{ and } i\in Q+1:Q+s\,. \end{cases}$$

Using again (A.2) and $(\widetilde{\mathbf{c}};\widetilde{\mathbf{d}})$ in (3.14), we have $\overline{\sigma}_{i,j+s} \le \widetilde{d}_{j+s}-\widetilde{c}_i$ for $\overline{\text{blockOf}}(j+s) > \overline{\text{blockOf}}(i) = q$, with $j = Q+1:n$. We can rewrite this inequality as

$$\overline{\sigma}_{ij} \le \widetilde{d}_j-\widetilde{c}_i \quad \text{for} \quad \overline{\text{blockOf}}(j) > \overline{\text{blockOf}}(i) = q\,,$$

with $j = Q+1+s:n+s$.

(c) $\overline{\mathbf{\Sigma}}_{qm_2}$ is $\overline{\mathbf{\Sigma}}_{qq}$, with $m_2 = q$. An entry $(i,j)$ in $\overline{\mathbf{\Sigma}}_{qq}$ satisfies $\overline{\text{blockOf}}(j) = \overline{\text{blockOf}}(i) = q$. We view block $q$ of the original DAE as a sub-DAE, with a signature matrix $\mathbf{\Sigma}_{qq}$ of size $N_q$ and an offset pair $(\mathbf{c}_q;\mathbf{d}_q)$. Given that the ES conditions are satisfied by (3.10), performing the ES conversion as described in Theorem 3.2 is equivalent to applying the basic ES method to this sub-DAE. After a conversion, the resulting enlarged signature matrix $\overline{\mathbf{\Sigma}}_{qq}$ of size $N_q+s$ has the form

$$\overline{\mathbf{\Sigma}}_{qq} = \left[\begin{array}{c|c|c} \overline{\mathbf{\Sigma}}_{qq,11} & \overline{\mathbf{\Sigma}}_{qq,12} & \overline{\mathbf{\Sigma}}_{qq,13}\\ \hline \overline{\mathbf{\Sigma}}_{qq,21} & \overline{\mathbf{\Sigma}}_{qq,22} & \overline{\mathbf{\Sigma}}_{qq,23} \end{array}\right]; \tag{A.11}$$

cf. Figure 3.1 and the details of the basic ES method in [13]. The two block rows of $\overline{\mathbf{\Sigma}}_{qq}$ correspond to $\overline{f}_i$ for $i \in B_q$ and $g_j$ for $j \in J$, respectively. The three block columns of $\overline{\mathbf{\Sigma}}_{qq}$ correspond to $x_j$ for $j \in J$, $x_j$ for $j \in B_q \setminus J$, and $y_j$ for $j \in J$, respectively. If we apply the same arguments in the proof of [13, Lemma 4.4] for the basic ES method, then we have $\widetilde{d}_j - \widetilde{c}_i \geq \overline{\sigma}_{ij}$ for all entries in $\overline{\mathbf{\Sigma}}_{qq}$. □

**Acknowledgements** The authors acknowledge with thanks the financial support for this research: GT is supported in part by the McMaster Centre for Software Certification through the Ontario Research Fund, Canada, NSN is supported in part by the Natural Sciences and Engineering Research Council of Canada, and JDP is supported in part by the Leverhulme Trust, the UK.

## References

1. Campbell, S.L., Griepentrog, E.: Solvability of general differential-algebraic equations. SIAM Journal on Scientific Computing **16**(2), 257–270 (1995)
2. Duff, I., Erisman, A., Reid, J.: Direct Methods for Sparse Matrices. Oxford Science Publications. Clarendon Press, Oxford (1986)
3. Mazzia, F., Iavernaro, F.: Test set for initial value problem solvers. Tech. Rep. 40, Department of Mathematics, University of Bari, Italy (2003). http://pitagora.dm.uniba.it/~testset/
4. McKenzie, R.: Structural analysis based dummy derivative selection for differential-algebraic equations. Tech. rep., Cardiff University, UK (2015). Submitted to BIT Numerical Mathematics, Oct 2015.
5. McKenzie, R.: Reducing the index of differential-algebraic equations by exploiting underlying structures. PhD Thesis, School of Mathematics, Cardiff University, Senghennydd Road, Cardiff CF24 4AG, UK (2016)
6. Nedialkov, N.S., Pryce, J.D., Tan, G.: Algorithm 948: DAESA—a Matlab tool for structural analysis of differential-algebraic equations: Software. ACM Trans. Math. Softw. **41**(2), 12:1–12:14 (2015)
7. Nedialkov, N.S., Tan, G., Pryce, J.D.: Exploiting fine block triangularization and quasilinearity in differential-algebraic equation systems (2016). McMaster University, Cardiff University. In preparation
8. Pryce, J.D.: Solving high-index DAEs by Taylor Series. Numerical Algorithms **19**, 195–211 (1998)
9. Pryce, J.D.: A simple structural analysis method for DAEs. BIT Numerical Mathematics **41**(2), 364–394 (2001)
10. Pryce, J.D., Nedialkov, N.S., Tan, G.: DAESA—a Matlab tool for structural analysis of differential-algebraic equations: Theory. ACM Trans. Math. Softw. **41**(2), 9:1–9:20 (2015)
11. Pryce, J.D., Nedialkov, N.S., Tan, G.: Fine block triangular structure of DAEs and its uses (2016). Cardiff University, McMaster University. In preparation
12. Tan, G., Nedialkov, N., Pryce, J.: Symbolic-numeric methods for improving structural analysis of differential-algebraic equation systems. Tech. rep., Department of Computing and Software, McMaster University, 1280 Main St. W., Hamilton, ON, L8S4L8, Canada (2015). CAS-15-07-NN, 84 pages
13. Tan, G., Nedialkov, N.S., Pryce, J.D.: Conversion methods for improving structural analysis of differential-algebraic equation systems. BIT Numerical Mathematics (2016). Submitted
14. The MathWorks, Inc.: Matlab Symbolic Math Toolbox (2016). http://www.mathworks.com/products/symbolic/