Numer Algor (2006) 43: 1–24
DOI 10.1007/s11075-006-9032-8

On the fast solution of Toeplitz-block linear systems arising in multivariate approximation theory

Stefan Becuwe · Annie Cuyt

Received: 8 December 2005 / Accepted: 2 May 2006 /
Published online: 18 August 2006

Abstract When constructing multivariate Padé approximants, highly structured linear systems arise in almost all existing definitions [10]. Until now little or no attention has been paid to fast algorithms for the computation of multivariate Padé approximants, with the exception of [17]. In this paper we show that a suitable arrangement of the unknowns and equations, for the multivariate definitions of Padé approximant under consideration, leads to a Toeplitz-block linear system with coefficient matrix of low displacement rank. Moreover, the matrix is very sparse, especially in higher dimensions. In Section 2 we discuss this for the so-called equation lattice definition and in Section 3 for the homogeneous definition of the multivariate Padé approximant. We do not discuss definitions based on multivariate generalizations of continued fractions [12, 25], or approaches that require some symbolic computations [6, 18]. In Section 4 we present an explicit formula for the factorization of the matrix that results from applying the displacement operator to the Toeplitz-block coefficient matrix. We then generalize the well-known fast Gaussian elimination procedure with partial pivoting developed in [14, 19], to deal with a rectangular block structure where the number and size of the blocks vary. We do not aim for a superfast solver because of the higher risk for instability. Instead we show how the developed technique can be combined with an easy interval arithmetic verification step. Numerical results illustrate the technique in Section 5.

Keywords Gaussian elimination · partial pivoting · displacement structure · Toeplitz-block · Padé approximant

AMS subject classification 15A23 · 41A21 · 65F05 · 65G20

Research partly funded by FWO-Vlaanderen.

S. Becuwe (✉) · A. Cuyt
Departement wiskunde en informatica, Universiteit Antwerpen (Campus Middelheim),
Middelheimlaan 1, B-2020 Antwerpen, Belgium
e-mail: stefan.becuwe@ua.ac.be

A. Cuyt
e-mail: annie.cuyt@ua.ac.be

1. Introduction

Given a univariate function $f(z)$ through its Taylor series expansion at a certain point in the complex plane, the Padé approximant $[n/m]^f$ of degree n in the numerator and m in the denominator for f is defined by (for simplicity we use the Taylor series at the origin)

$$\begin{aligned} f(z) &= \sum_{i=0}^{\infty} c_i z^i, \\ p(z) &= \sum_{i=0}^{n} a_i z^i, \\ q(z) &= \sum_{i=0}^{m} b_i z^i, \\ (fq-p)(z) &= \sum_{i\geq n+m+1} d_i z^i, \end{aligned} \tag{1}$$

with $[n/m]^f$ equal to the irreducible form of p/q. The conditions

$$d_i = 0, \qquad i = n+1, \dots, n+m,$$

give rise to a Toeplitz system of linear equations:

$$\begin{cases} c_{n+1}b_0 + c_n b_1 + \cdots + c_{n+1-m}b_m = 0, \\ \quad\vdots \\ c_{n+m}b_0 + c_{n+m-1}b_1 + \cdots + c_n b_m = 0. \end{cases} \tag{2}$$

The concept of displacement rank is first introduced in [22]. We use the definition as given in [14] where the $\{L, R\}$-displacement rank, or displacement rank for short, of an $u \times v$ matrix T is defined as the rank of the matrix $\nabla T = LT - TR$ with L and R being so-called left and right displacement operators. If T is a Toeplitz matrix as in (2) and $L = Z_u^{(1)}$, $R = Z_v^{(-1)}$ with

$$Z_k^{(w)} = \begin{pmatrix} 0 & \dots & & 0 & w \\ 1 & 0 & \dots & & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix}_{k\times k}$$

then the resulting matrix $LT - TR$ is almost entirely filled with zeroes, except for its first row and last column. Hence T has displacement rank 2, irrespective of its size.

Linear systems with a low displacement rank coefficient matrix, can be solved much faster than by means of the traditional methods. In general, a linear system with an $m \times m$ coefficient matrix of displacement rank α, can be solved in $\mathcal{O}(\alpha m^2)$ operations instead of the traditional $\mathcal{O}(m^3)$. Here $\mathcal{O}(\cdot)$ indicates that the floating-point operation count of the algorithm is bounded above by a constant multiple of $(\cdot)$. It has been indicated that such fast solvers are usually less stable [1, 5]. Precisely for that reason we do not consider any so-called superfast algorithms.

In [19] a pivoting strategy for the fast solution of linear systems with a Cauchy-like coefficient matrix is developed and it is shown how Toeplitz-like systems can be converted into Cauchy-like systems. Other authors have also indicated how one class of structured matrices can be transformed into another one [15]. The technique developed for Cauchy-like matrices, which is also applicable to square block matrices of identical size, is based on the existence of a factorization of $LT - TR$ of the form

$$\nabla T = LT - TR = GB, \qquad G \in \mathbb{C}^{u \times \alpha}, \qquad B \in \mathbb{C}^{\alpha \times v}.$$

Here the factors G and B are not full-size, but of a 'thin' size α. For a Toeplitz matrix

$$T = \begin{pmatrix} c_n & c_{n-1} & \dots & c_0 \\ c_{n+1} & c_n & \dots & c_1 \\ \vdots & & & \vdots \\ c_{2n} & \dots & & c_n \end{pmatrix},$$

such a factorization is easy to find:

$$\begin{aligned} Z_{n+1}^{(1)} T - T Z_{n+1}^{(-1)} &= \begin{pmatrix} c_{2n} - c_{n-1} & c_{2n-1} - c_{n-2} & \dots & c_{n+1} - c_0 & 2c_n \\ 0 & & \dots & 0 & c_{n+1} + c_0 \\ \vdots & & & \vdots & \vdots \\ 0 & & & 0 & c_{2n} + c_{n-1} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 2c_n \\ 0 & c_{n+1} + c_0 \\ \vdots & \vdots \\ 0 & c_{2n} + c_{n-1} \end{pmatrix} \cdot \begin{pmatrix} c_{2n} - c_{n-1} & \dots & c_{n+1} - c_0 & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}. \end{aligned} \tag{3}$$

Let us now turn to the multivariate generalization of all this.

Given a Taylor series expansion (for simplicity we describe only the bivariate case but the higher dimensional case is only notationally more difficult)

$$f(x, y) = \sum_{(i,j) \in \mathbb{N}^2} c_{ij} x^i y^j, \tag{4}$$

one can group the different definitions for multivariate Padé approximants into four main categories, depending on how one deals with the information c_{ij}. Rewriting $f(x, y)$ as

$$f(x, y) = \sum_{k=0}^{\infty} c_{i_k j_k} x^{i_k} y^{j_k}$$

is done in what we call the 'equation lattice' group of definitions. Another way to deal with the information is to rewrite $f(x, y)$ as

$$f(x, y) = \sum_{k=0}^{\infty} \left(\sum_{i+j=k} c_{ij} x^i y^j \right)$$

and to process the ‘homogeneous’ subexpressions of degree k in the same way as the univariate terms of degree k in (1). A third group of definitions looks at the Taylor series development as

$$f(x, y) = \sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} c_{ij} y^j \right) x^i = \sum_{i=0}^{\infty} c_i(y) x^i$$

and treats the problem, at least partly, in a ‘symbolic’ way. It is therefore out of the scope of this paper. Since the ‘continued fraction’ approach [12, 25] does not compute its multivariate approximant from a defining system of equations for the numerator and denominator coefficients, we do not discuss this generalization either.

2. The equation lattice approach

2.1. Definition

For $f(x, y)$ given by (4), we can define a multivariate Padé approximant p/q to f by determining $p(x, y)$ and $q(x, y)$ from accuracy-through-order conditions as follows. Let the polynomials $p(x, y)$ and $q(x, y)$ be of the general form

$$\begin{aligned} p(x, y) &= \sum_{(i,j)\in N} a_{ij} x^i y^j, \\ q(x, y) &= \sum_{(i,j)\in D} b_{ij} x^i y^j, \end{aligned}$$

where N (for the numerator) and D (for the denominator) are nonempty finite subsets of $\mathbb{N}^2$. The sets N and D indicate in a way the degree of the polynomials $p(x, y)$ and $q(x, y)$. Let us denote

$$\#N = n + 1, \qquad \#D = m + 1.$$

In analogy with the univariate case, we choose a set of indices E (for the equations) such that

$$N \subseteq E, \tag{5a}$$

$$\#(E\backslash N) = m = \#D - 1, \tag{5b}$$

$$E \text{ satisfies the inclusion property.} \tag{5c}$$

Here (5c) means that when a point belongs to the index set E, then the rectangular subset of points emanating from the origin with the given point as its furthermost corner, also lies in E. In other words,

$$(i, j) \in E \Rightarrow \{(k, \ell) \mid k \le i, \ell \le j\} \subseteq E.$$

We then impose the following accuracy-through-order conditions on the polynomials $p(x, y)$ and $q(x, y)$, namely

$$(fq - p)(x, y) = \sum_{(i,j)\in\mathbb{N}^2\backslash E} d_{ij} x^i y^j. \tag{6}$$

Condition (5a) enables us to split the system of equations

$$d_{ij} = 0, \qquad (i, j) \in E,$$

in a nonhomogeneous part defining the numerator coefficients

$$\sum_{\mu=0}^{i} \sum_{\nu=0}^{j} c_{\mu\nu} b_{i-\mu, j-\nu} = a_{ij}, \qquad (i, j) \in N,$$

and a homogeneous part defining the denominator coefficients

$$\sum_{\mu=0}^{i} \sum_{\nu=0}^{j} c_{\mu\nu} b_{i-\mu, j-\nu} = 0, \qquad (i, j) \in E \backslash N. \tag{7}$$

By convention $b_{k\ell} = 0$ if $(k, \ell) \notin D$. Condition (5b) guarantees the existence of a nontrivial denominator $q(x, y)$ because the homogeneous system has one equation less than the number of unknowns and so one unknown coefficient can be chosen freely. We denote the set of rational functions p/q satisfying (6) by $[N/D]_E^f$ and call it the general multivariate Padé approximant for f.

Because of the freedom in choosing the sets N, D and E, the equation lattice definition covers a variety of approximation schemes, sometimes with minor variations on the general definition above. In [2, 3, 27–31] rectangular schemes are studied, in [3, 4, 8, 16, 24] triangular schemes, and in [7, 20, 21, 23] a combination of both. For more information we also refer to [11, 26].

In general, uniqueness of the general multivariate Padé approximant, in the sense that all rational functions in $[N/D]_E^f$ reduce to the same irreducible form, is not guaranteed, unless the index set $E \setminus N$ supplies a homogeneous system of linearly independent equations. As already mentioned, at least one nontrivial solution of (6) exists because the number of unknown coefficients b_{ij} is one more than the number of conditions in (7). But it is not so (unlike in the univariate case) that different solutions p_1, q_1 and p_2, q_2 of (6) are necessarily equivalent, meaning that $(p_1 q_2)(x, y) = (p_2 q_1)(x, y)$. Hence p_1/q_1 and p_2/q_2 may be different functions. In general one can only say that

$$(p_1 q_2 - q_1 p_2)(x, y) = \sum_{(i,j) \in N \times D \backslash E} e_{ij} x^i y^j,$$

where

$$N \times D = \{(i + k, j + \ell) \mid (i, j) \in N, (k, \ell) \in D\}.$$

One way to enforce a unicity property is to choose the index set E as large as possible, by adding conditions as soon as there are linearly dependent equations in (7), but this is not possible in all approximation schemes.

2.2. Toeplitz-block structure

Let us denote the index set $E \setminus N$ by S and introduce the notations

$$I^{(S)} = \max\{ i \mid (i, j) \in S \},$$
$$J^{(S)} = \max\{ j \mid (i, j) \in S \}.$$

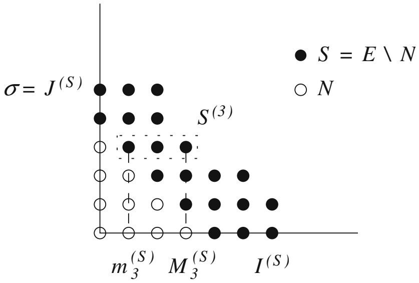

Figure 1 Breaking down $E \setminus N$.

For the sake of simplicity we assume for now that the set N also satisfies the inclusion property, so that the set S does not look like a Swiss cheese with holes. If $I^{(S)} > J^{(S)}$ then we decompose S as (figure 1)

$$\begin{aligned} m_j^{(S)} &= \min\{\, i \mid (i,j) \in S \,\}, \\ M_j^{(S)} &= \max\{\, i \mid (i,j) \in S \,\}, \\ S^{(j)} &= \{\, (i,j) \mid m_j^{(S)} \le i \le M_j^{(S)} \,\}, \\ S &= S^{(0)} \cup \cdots \cup S^{(J^{(S)})}. \end{aligned}$$

In the other case the role of i and j is interchanged. Note that some of the sets $S^{(j)}$ may be empty, including the set $S^{(0)}$. In the sequel we assume that $I^{(S)} > J^{(S)}$ and we introduce the shorthand σ for $J^{(S)}$. Similarly we need δ for $\min\left(I^{(D)}, J^{(D)}\right)$. Using these notations, we rewrite (7) as

$$\sum_{\mu=0}^{i} \sum_{\nu=0}^{j} c_{\mu\nu} b_{i-\mu,j-\nu} = 0, \qquad (i,j) \in S^{(j)}, \qquad j = 0,\ldots,\sigma, \tag{8}$$

and we arrange the unknown coefficients b_{ij} as

$$\left(b_{m_0^{(D)},0}, \ldots, b_{M_0^{(D)},0} \mid b_{m_1^{(D)},1}, \ldots, b_{M_1^{(D)},1} \mid \ldots \mid b_{m_\delta^{(D)},\delta}, \ldots, b_{M_\delta^{(D)},\delta} \right).$$

Introducing the $(M_k^{(S)} - m_k^{(S)} + 1) \times (M_\ell^{(D)} - m_\ell^{(D)} + 1)$ Toeplitz matrices

$$C_{k\ell}^{(S)} = \begin{pmatrix} c_{m_k^{(S)},k-\ell} & \cdots & c_{m_k^{(S)}-(M_\ell^{(D)}-m_\ell^{(D)}),k-\ell} \\ \vdots & & \vdots \\ c_{M_k^{(S)},k-\ell} & \cdots & c_{M_k^{(S)}-(M_\ell^{(D)}-m_\ell^{(D)}),k-\ell} \end{pmatrix}$$

the coefficient matrix of the system of linear equations Equation (7) takes the Toeplitz-block structure

$$\begin{pmatrix} C_{00}^{(S)} & 0 & & \cdots & 0 \\ C_{10}^{(S)} & C_{11}^{(S)} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ & & & & 0 \\ & & & & C_{\delta,\delta}^{(S)} \\ & & & & \vdots \\ C_{\sigma,0}^{(S)} & C_{\sigma,1}^{(S)} & \cdots & & C_{\sigma,\delta}^{(S)} \end{pmatrix}. \tag{9}$$

If $k < \ell$ then the matrix $C_{k\ell}^{(S)}$ is a zero matrix with the same dimensions. If $\sigma - \delta$ is not too large, then (9) can be called block lower triangular.

2.3. Displacement rank and sparsity

For a Toeplitz-block matrix T with $\sigma + 1$ block rows and $\delta + 1$ block columns and rectangular Toeplitz blocks of size $u_i \times v_j = (M_i^{(S)} - m_i^{(S)} + 1) \times (M_j^{(D)} - m_j^{(D)} + 1)$ the displacement operators

$$L = \bigoplus_{t=0}^{\sigma} Z_{u_t}^{(1)}, \qquad R = \bigoplus_{t=0}^{\delta} Z_{v_t}^{(-1)},$$

are used, where $\bigoplus W_t$ denotes the block diagonal matrix of which the t^{th} block is given by W_t. When applied to the coefficient matrix T of (9) the resulting matrix $LT - TR$ has the same block structure as (9). Moreover, each block of ∇T consists of zeroes except for its first row and last column. Hence the displacement rank of (9) is at most

$$(\sigma + 1) + (\delta + 1) = \min\left(I^{(S)}, J^{(S)}\right) + \min\left(I^{(D)}, J^{(D)}\right) + 2, \qquad S = E \setminus N.$$

The fact that the matrices $C_{k\ell}^{(S)}$ are filled with zeroes when $k < \ell$ implies that a not insignificant number of matrix entries can be zero, namely at least a fraction of

$$\frac{\sum_{\ell=1}^{\delta}\left[\left(M_\ell^{(D)} - m_\ell^{(D)} + 1\right)\sum_{k=0}^{\sigma-\delta+\ell-1}\left(M_k^{(S)} - m_k^{(S)} + 1\right)\right]}{m(m+1)},$$

where $m = \#S = \#D - 1$ with $S = E \setminus N$. This approaches 50%. Additional zeroes occur when some of the indices i, ranging between $m_k^{(S)} - (M_\ell^{(D)} - m_\ell^{(D)}) \leq i \leq M_k^{(S)}$,

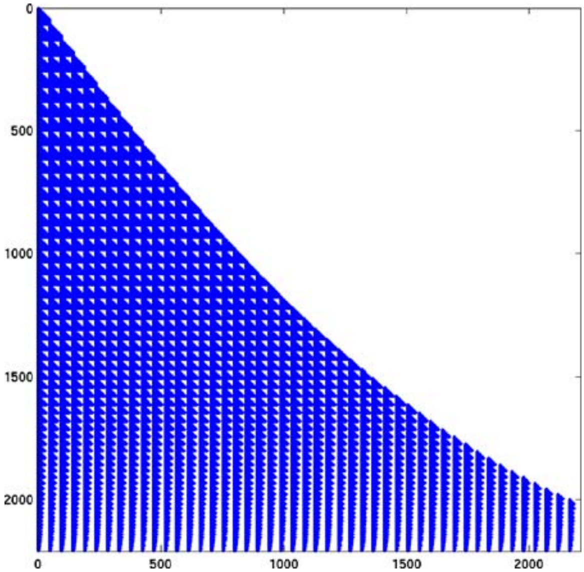

Figure 2 Sparsity of (9) for $[N_2/D_{2208}]_{E_{2210}}$.

become negative. To illustrate the sparsity of (9) we show in figure 2 the resulting matrix (zero entries are blank) for $s = 2$ with N, D and E given by

$$\begin{array}{ll} N = \{(0,0),(1,0),(0,1)\}, & \#N = 3, \\ D = \{(i,j) \mid 0 \le i, j \le 46\}, & \#D = 2209, \\ E = \{(i,j) \mid 0 \le i + j \le 65\}, & \#E = 2211. \end{array}$$

3. The homogeneous approach

3.1. Definition

The main difference between the equation lattice approach on one hand and the homogeneous approach on the other hand, is that in the former the number N_e of equations imposed on the coefficients of the multivariate Padé approximant is one less than the number N_u of unknown coefficients that have to be determined, just like in the univariate case, while in the latter the system of equations is seriously overdetermined as soon as one is dealing with more than two variables. Despite this overdetermination, the system inherently only consists of at most $N_u - 1$ linearly independent equations, making it soluble without having to resort to least squares techniques. This remarkable fact is already pointed out in [9].

For the definition of the homogeneous multivariate Padé approximant $[n/m]_H^f$ we introduce the notations

$$\begin{aligned} A_k(x,y) &= \sum_{i+j=nm+k} a_{ij}x^i y^j, & k &= 0,\ldots,n, \\ B_k(x,y) &= \sum_{i+j=nm+k} b_{ij}x^i y^j, & k &= 0,\ldots,m, \\ C_k(x,y) &= \sum_{i+j=k} c_{ij}x^i y^j, & k &= 0,1,2\ldots. \end{aligned}$$

For chosen n and m the polynomials

$$\begin{aligned} p(x,y) &= \sum_{k=0}^{n} A_k(x,y), \\ q(x,y) &= \sum_{k=0}^{m} B_k(x,y), \end{aligned}$$

are then computed from the conditions

$$(fq - p)(x,y) = \sum_{i+j \ge nm+n+m+1} d_{ij}x^i y^j, \tag{10}$$

where the conditions of homogeneous degree $nm + n + 1$ up to $nm + n + m$ can be rewritten as

$$\begin{cases} C_{n+1}(x,y)B_0(x,y) + \cdots + C_{n+1-m}(x,y)B_m(x,y) \equiv 0, \\ \quad \vdots \\ C_{n+m}(x,y)B_0(x,y) + \cdots + C_n(x,y)B_m(x,y) \equiv 0, \end{cases} \tag{11}$$

with $C_k(x,y) \equiv 0$ if $k < 0$. This is exactly the system of defining equation (2) for univariate Padé approximants where the term $c_k z^k$ in the univariate definition is substituted by

$$C_k(x,y) = \sum_{i+j=k} c_{ij} x^i y^j, \qquad k = 0, 1, 2, \ldots .$$

A simple count of unknowns and conditions in (10) shows that in the bivariate case the number of equations is one less than the number of unknowns, just like in the univariate case. But in the general multivariate case the system of defining equation (10) is overdetermined. Nevertheless it has been proven that a nontrivial solution also exists in the multivariate case [9, pp. 60–62]. It is therefore unnecessary to consider the linear conditions (10) in a least squares sense. This inherent dependence among the homogeneous Padé approximation conditions is still not fully understood and may lead to new developments.

For the homogeneous Padé approximants we can also prove that if p_1 and q_1 as well as p_2 and q_2 satisfy condition (10), then

$$(p_1 q_2)(x,y) = (p_2 q_1)(x,y).$$

The homogeneous multivariate Padé approximant $[n/m]_H^f$ for $f(x,y)$ can then be defined as the unique irreducible form of a solution $p(x,y)/q(x,y)$ of (10). Several suitable normalizations are possible. This unicity of the irreducible form is a distinctive characteristic of the homogeneous approach.

3.2. Block-Toeplitz-block structure

In order to better understand the structure of this system, we start by writing it as a linear system in the coefficients b_{ij}. In order to do so we arrange the unknown denominator coefficients in the order:

$$\left(b_{nm,0} \ldots b_{0,nm} \mid b_{nm+1,0} \ldots b_{0,nm+1} \mid \ldots \mid b_{nm+m,0} \ldots b_{0,nm+m}\right).$$

When we arrange the conditions (11) in a similar (upward sloping diagonal) way and introduce the Toeplitz blocks

$$C_n^{(nm)} = \begin{pmatrix} c_{n,0} & 0 & \ldots & 0 \\ \vdots & \ddots & & \vdots \\ c_{0,n} & & & 0 \\ 0 & & & c_{n,0} \\ \vdots & & \ddots & \vdots \\ 0 & \ldots & 0 & c_{0,n} \end{pmatrix}_{(n+nm+1)\times(nm+1)} \tag{12}$$

then the coefficient matrix of the system of equation (11) looks like

$$\begin{pmatrix} C_{n+1}^{(nm)} & C_n^{(nm+1)} & \ldots & C_{n-m+1}^{(nm+m)} \\ C_{n+2}^{(nm)} & C_{n+1}^{(nm+1)} & & \\ \vdots & \vdots & \ddots & \\ C_{n+m}^{(nm)} & \ldots & & C_n^{(nm+m)} \end{pmatrix} \tag{13}$$

which is very similar to (2): Since the superscript in the notation of a block $C_i^{(j)}$ only influences the block's dimensions and not it contents, (13) clearly has contentwise identical blocks along downward sloping diagonals. Moreover, behind each entry $C_i^{(j)}$ in this Toeplitz-structured matrix unfolds a simpler Toeplitz matrix (12). Actually (13) is block Toeplitz with individual Toeplitz blocks.

This is only the bivariate case. Now let us discuss the higher dimensional case. We shall see that this principle of unfolding can be applied recursively. The structure of the block-Toeplitz-block coefficient matrix resembles that of a set of Russian nested Matrioshka dolls. When going from one to two variables, the coefficient matrix of (2) is transformed into (13) which looks identical until we 'open' each entry $C_i^{(j)}$ and find that there is another Toeplitz matrix inside. We now describe the transition from two to more variables.

Let us denote the number of variables by s and let us denote by 0_t a sequence of t zeroes in a multi-index. The generalization of (10) and (11) to s variables is straightforward and so for reasons of conciseness is not repeated. First we arrange the unknown coefficients $b_{i_1 \ldots i_s}$ and afterwards the entries of the coefficient matrix of (11). We start by arranging a subset of the coefficients and then describe an unfolding mechanism to include all the other coefficients. The first $b_{i_1 \ldots i_s}$ to be selected and ordered are

$$\left(b_{nm,0,0_{s-2}}, b_{nm-1,1,0_{s-2}} \ldots b_{0,nm,0_{s-2}} \mid \ldots \mid b_{nm+m,0,0_{s-2}} \ldots b_{0,nm+m,0_{s-2}}\right).$$

We have clearly focused on the first and second index. Then we let each $b_{ij0_{s-2}}$ unfold to

$$\left(b_{i,j,0,0_{s-3}}, b_{i,j-1,1,0_{s-3}}, \ldots, b_{i,0,j,0_{s-3}}\right).$$

Here we have focused on the second and third index. Let us now repeat the procedure for the third and fourth index and so on. We let each $b_{i,j,k,0_{s-3}}$ unfold to

$$\left(b_{i,j,k,0,0_{s-4}}, b_{i,j,k-1,1,0_{s-4}}, \ldots, b_{i,j,0,k,0_{s-4}}\right).$$

If this unfolding is performed $s-2$ times then all the unknown denominator coefficients are ordered. Before constructing the coefficient matrix of (11) according to the same principle, let us count the number of equations and the number of unknowns.

Each homogeneous expression $B_k(x_1, \ldots, x_s)$ contains $\binom{s+nm+k-1}{nm+k}$ coefficients $b_{i_1 \ldots i_s}$. So the total of unknown denominator coefficients $b_{i_1 \ldots i_s}$ equals

$$N_u = \sum_{k=0}^{m} \binom{s+nm+n+k-1}{nm+k}.$$

The k^{th} equation in (11) equates an $(nm+n+k)$-linear operator in s variables to zero. So it equates $\binom{s+nm+n+k-1}{nm+n+k}$ coefficients $d_{i_1 \ldots i_s}$ to zero. Hence the number of homogeneous equations is in total

$$N_e = \sum_{k=1}^{m} \binom{s+nm+n+k-1}{nm+n+k}. \tag{14}$$

If $nm > 0$ then

$$N_u = \binom{s+nm+m}{nm+m} - \binom{s+nm-1}{nm-1},$$
$$N_e = \binom{s+nm+n+m}{nm+n+m} - \binom{s+nm+n}{nm+n}.$$

If $nm = 0$ then $N_u = \binom{s+m}{m}$. For $s = 2$ the above values lead to $N_u - N_e = 1$ while for $s > 2$ the system is clearly overdetermined.

The construction of the coefficient matrix of the overdetermined homogeneous system of equations becomes straightforward if the principle of unfolding is again applied. Take the first column of each $C_i^{(j)}$ in (13) and expand

$$\left(c_{i,0,0_{s-2}} \dots c_{0,i,0_{s-2}}\right)$$

in the same way as the subvector $(b_{i,0,0_{s-2}} \dots b_{0,i,0_{s-2}})$. During the unfolding process, the size of every Toeplitz block at each step in the process can be determined from the following: The k^{th} term in the expression (14) for N_e is linked to the block entries in the k^{th} row of (13) and can be decomposed as

$$\binom{s+nm+n+k-1}{nm+n+k} = \sum_{\ell=0}^{nm+n+k} \binom{(s-1)+\ell-1}{\ell},$$

indicating that each unfolded block of row size $\binom{s+nm+n+k-1}{nm+n+k}$ consists of a block-Toeplitz-block structure with subblocks of row size $\binom{(s-1)+\ell-1}{\ell}$. Take for instance $s = 3$, $n = 1$ and $m = 2$ and construct the three-dimensional analogue of the upper left block $C_{n+1}^{(nm)}$ of (13). The 5×3 matrix $C_2^{(2)}$ for $s = 2$ is given by

$$C_2^{(2)} = \begin{pmatrix} c_{20} & 0 & 0 \\ c_{11} & c_{20} & 0 \\ c_{02} & c_{11} & c_{20} \\ 0 & c_{02} & c_{11} \\ 0 & 0 & c_{02} \end{pmatrix}.$$

In the transition from two to three variables the vector $(c_{200} \mid c_{110} \mid c_{020})$ unfolds to

$$(c_{200} \mid c_{110}, c_{101} \mid c_{020}, c_{011}, c_{002})$$

and the vector $(b_{200} \mid b_{110} \mid b_{020})$ unfolds to the vector

$$(b_{200} \mid b_{110}, b_{101} \mid b_{020}, b_{011}, b_{002})$$

in which the unknown coefficients of $B_0(x_1, x_2, x_3)$ are arranged for $nm = 2$. The size of each compartment in this last vector determines the column size of the rectangular Toeplitz blocks in the three-dimensional analogue of (13) while the row size of each block can be determined from the unfolding of the first column in $C_2^{(2)}$. For instance the Toeplitz block that takes the place of the entry on row 3 and column 2 of $C_2^{(2)}$ has three rows because the entry on row 3 in the first column unfolded to $(c_{020}, c_{011}, c_{002})$. It has two columns because in the vector of unknowns, which is multiplied with the coefficient matrix, the second compartment contains two elements. Also, the total number of rows of the three-dimensional analogue of $C_2^{(2)}$ is given by the first ($k = 1$)

term of (14) which equals 15. As explained, these 15 rows split up in five smaller constructions according to

$$\binom{s+nm+n+k-1}{nm+n+k} = \binom{6}{4} = \sum_{\ell=0}^{4} \binom{\ell+1}{\ell}, \qquad k=1.$$

Besides determining the correct dimensions of the blocks, we should also point out that the contents of the blocks and the entries in the blocks are copied along downward sloping diagonals, in Toeplitz fashion. Hence we finally obtain for the three-dimensional analogue of $C_{n+1}^{(nm)}$ with $n=1$ and $m=2$:

$$\begin{pmatrix}
c_{2,0,0} & 0 & 0 & 0 & 0 & 0 \\
c_{1,1,0} & c_{2,0,0} & 0 & 0 & 0 & 0 \\
c_{1,0,1} & 0 & c_{2,0,0} & 0 & 0 & 0 \\
c_{0,2,0} & c_{1,1,0} & 0 & c_{2,0,0} & 0 & 0 \\
c_{0,1,1} & c_{1,0,1} & c_{1,1,0} & 0 & c_{2,0,0} & 0 \\
c_{0,0,2} & 0 & c_{1,0,1} & 0 & 0 & c_{2,0,0} \\
0 & c_{0,2,0} & 0 & c_{1,1,0} & 0 & 0 \\
0 & c_{0,1,1} & c_{0,2,0} & c_{1,0,1} & c_{1,1,0} & 0 \\
0 & c_{0,0,2} & c_{0,1,1} & 0 & c_{1,0,1} & c_{1,1,0} \\
0 & 0 & c_{0,0,2} & 0 & 0 & c_{1,0,1} \\
0 & 0 & 0 & c_{0,2,0} & 0 & 0 \\
0 & 0 & 0 & c_{0,1,1} & c_{0,2,0} & 0 \\
0 & 0 & 0 & c_{0,0,2} & c_{0,1,1} & c_{0,2,0} \\
0 & 0 & 0 & 0 & c_{0,0,2} & c_{0,1,1} \\
0 & 0 & 0 & 0 & 0 & c_{0,0,2}
\end{pmatrix}. \tag{15}$$

3.3. Displacement rank and sparsity

For a Toeplitz-block matrix with m block rows and $m+1$ block columns and rectangular Toeplitz blocks of size $u_i = nm+n+i+1$ by $v_j = nm+j+1$ the displacement operators

$$L = \bigoplus_{t=1}^{m} Z_{u_t}^{(1)}, \qquad R = \bigoplus_{t=0}^{m} Z_{v_t}^{(-1)}, \tag{16}$$

are used. When applied to the coefficient matrix of (11) the resulting matrix consists of zeroes except for the last column of each block at the lowest level of the recursive unfolding process. So in order to know the displacement rank, the number of block columns must be counted. From the construction of our block-Toeplitz-block matrix it should be clear that, for $s \neq 2$, this number is given by

$$\sum_{k=nm}^{nm+m} \sum_{\ell=0}^{k} \binom{(s-2)+\ell-1}{\ell}$$

and, for $s=2$, the displacement rank of (13) equals at most $m+1$.

When computing the actual size $N_e \times N_u$ of the coefficient matrix of (11), it is apparent that as the number of variables grows when $s>2$, the system is soon very much overdetermined. For instance for $s=4, n=3$ and $m=4$ we have $N_e = 4979$ and $N_u = 3480$. However, it is proved in [9, pp. 60–62] that a nontrivial solution of the system always exists and that the superfluous equations in the overdetermined system are mathematically linearly dependent.

When inspecting the coefficient matrix it is also clear that it is very sparse and at the same time highly structured. In a first attempt to get a grip on the redundant equations in (11), when considering it symbolically, we tried to pinpoint the $N_e - N_u + 1$ linearly dependent equations. Although the structure is responsible for the redundancy, the linearly dependent equations did not show up at specified entries in the matrix. For instance, the following interesting experiment can be performed using a computer algebra system. Take $s = 3$, $n = 1$ and $m = 2$. The 36×31 symbolic three-dimensional homogeneous system (symbolic entries c_{ijk}) has rank 30 and every system of 30 equations out of the 36 imposed ones has rank 30. It is not so that particular rows in the matrix constitute the linearly dependent equations. In order to reduce the size of the overdetermined linear system, another strategy has to be followed. While removing equations, ideally, one prefers to disrupt the structure of the matrix as little as possible. At the same time one wants to obtain a linear system that is as well-conditioned as possible. Let us first focus on the former.

Because of the structure, preferably equations are eliminated at the end of Toeplitz blocks and not in the middle, a restriction which is apparently not in conflict with the location of the linearly dependent equations. At no point should a combination of equations be removed such that some of the given coefficients $c_{i_1 \ldots i_s}$ are totally deleted from the system. In order to be sure that this is always possible, we count the number N_d of data $c_{i_1 \ldots i_s}$ necessary for the construction of the denominator of $[n/m]_H^f$ and compare it to N_u. This is the number of coefficients $c_{i_1 \ldots i_s}$ appearing in the first column of (11). From (11) it is clear that for s variables, N_d is given by

$$N_d = \sum_{k=n+1}^{n+m} \binom{s+k-1}{k} = \binom{s+n+m}{n+m} - \binom{s+n}{n}.$$

Since

$$\binom{s+nm+k-1}{nm+k} \geq \binom{s+n+k-1}{n+k}, \qquad k \geq 1, \quad s \geq 2,$$

we obtain $N_u > N_d$. Hence it is always possible to cut away equations without cutting away data.

Now let us inspect the condition number of the square subsystem. Let us again inspect an example, namely

$$f(x, y, z) = \frac{e^{2x} \log(3-y)}{4-z} = \sum_{i,j,k=0}^{\infty} c_{ijk} x^i y^j z^k$$

and reconsider the overdetermined system of 36×31 homogeneous equations obtained for the choice $n = 1$ and $m = 2$. In total 62,329,344 possible 30×30 square inhomogeneous systems of linear equations are possible. According to MATLAB, condition numbers vary from 746 to overflow. When removing only equations at the end of Toeplitz blocks, condition numbers start at 798, which is absolutely comparable. The square linear system which results by cutting away the last two equations and the entire block consisting of the rows 7 to 10 in (15), thus optimizing the displacement rank, still leads to a condition number of 6138. This relationship between structure and condition number should clearly be the subject of future research.

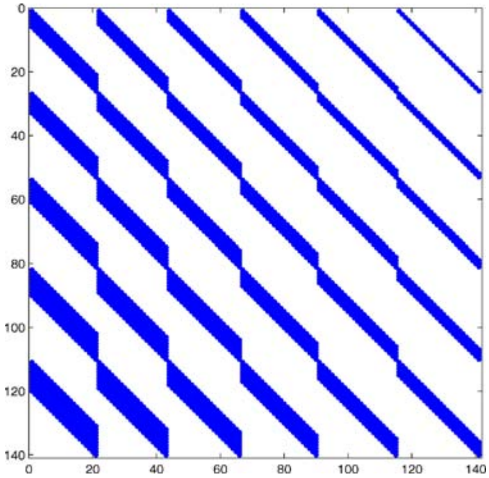

Figure 3 Sparsity of (13) for $s=2, n=4, m=5$.

In the case $s=4$, $n=3$ and $m=4$ one has to remove 1,500 equations from the overdetermined system before it can be passed to a solver. When inspecting the coefficient matrix, one counts 1,420 trailing zero entries in the first column. This part can be cut away, but another 80 equations have to be eliminated higher up, with minimal influence on the structure. Here minimal effect on the structure means with an increase of the displacement rank by one per block row in which equations are cut away.

From the structure of (13) it is also possible to compute the sparsity of the block-Toeplitz-block matrix (figure 3). The full matrix is of size $N_e \times N_u$. The number of matrix entries filled with coefficients $c_{i_1 \dots i_s}$ equals

$$N_c = \sum_{\ell=0}^{m} \sum_{k=n+1-\ell}^{n+m-\ell} \binom{s+k-1}{k} \binom{s+nm+\ell-1}{nm+\ell}.$$

Hence only a fraction of at most $N_c/(N_e \times N_u)$ coefficients in the matrix are nonzero. After removing the redundant equations, a fraction of at most $N_c/\,(N_u(N_u-1))$ in the matrix is nonzero. This is clearly an upper bound because some of the rows have been removed and hence the true numerator is less than N_c. Nevertheless for several values of s, n and m the sparsity lower bound, which we compute as

$$\zeta = 1 - \frac{N_c}{N_u(N_u-1)}$$

is soon close to 100% as can be seen from table 1.

Fortunately, the problem of having to remove redundant equations does not play in the bivariate case, only when $s>2$, and hence does not affect the numerical examples in the subsequent sections.

Table 1 Sparsity of (13).

s	n/m	N_e	N_d	N_c	N_u	ζ
2	1/2	11	7	56	12	0.576
3	1/2	36	16	246	31	0.735
2	4/5	140	40	3, 790	141	0.808
4	1/2	91	30	755	65	0.819
4	3/4	4979	295	415, 924	3, 480	0.966
6	5/5	2, 548, 596	7, 546	3, 283, 329, 337	1, 354, 017	0.998

4. Factorization and fast GEPP

4.1. Basic technique

The $t \times t$ matrix $Z_t^{(w)}$ is a companion matrix and can be factored as

$$Z_t^{(w)} = Q_{w,t}^H D_w Q_{w,t} = Q_{w,t}^H \mathrm{diag}(\lambda_1^{(w)}, \ldots, \lambda_t^{(w)}) Q_{w,t} \tag{17}$$

where the eigenvalues $\lambda_i^{(w)}$ are the t complex zeroes of

$$z^t - w = 0$$

and the columns of the unitary matrix $Q_{w,t}^H$ are the Schur vectors. Since $Z_t^{(w)}$ is normal, the Schur vectors are also the eigenvectors of $Z_t^{(w)}$. For $w = 1$,

$$\begin{aligned} \lambda_j^{(1)} &= \exp(\mathrm{i}2\pi(j-1)/t), \qquad j = 1, \ldots, t, \\ Q_{1,t} &= \frac{1}{\sqrt{t}} (\exp(\mathrm{i}2\pi(j-1)(k-1)/t))_{1 \le j,k \le t}\,, \end{aligned} \tag{18}$$

and for $w = -1$,

$$\begin{aligned} \lambda_j^{(-1)} &= \exp(\mathrm{i}\pi(2j-1)/t), \qquad j = 1, \ldots, t, \\ Q_{-1,t} &= \frac{1}{\sqrt{t}} (\exp(\mathrm{i}2\pi(j-1)(k-1)/t))_{1 \le j,k \le t} \\ &\qquad \times \mathrm{diag}\,(1, \exp(\mathrm{i}\pi/t), \ldots, \exp(\mathrm{i}\pi(t-1)/t))\,. \end{aligned} \tag{19}$$

Given the expressions (18) for $Q_{1,t}$, (19) for $Q_{-1,t}$ and (3) for a $t \times t$ Toeplitz matrix T, we can also write

$$\begin{aligned} Q_{1,t}\left(Z_t^{(1)} T - T Z_t^{(-1)}\right) Q_{-1,t}^H &= D_1 \hat{T} - \hat{T} D_{-1} \\ &= \left(Q_{1,t} G\right) \cdot \left(B Q_{-1,t}^H\right) = \hat{G}_{t\times 2} \hat{B}_{2\times t}, \end{aligned} \tag{20}$$

where the matrix $\hat{T} = Q_{1,t} T Q_{-1,t}^H$ is now a Cauchy-like matrix. For the solution of linear systems with Cauchy-like coefficient matrices a fast technique, incorporating partial pivoting, is proposed in [19]. It is based on the knowledge of the factorization $\hat{G}\hat{B}$ and heavily relies on the fact that the entries in the matrices D_1 and D_{-1} are distinct. The technique can easily be generalized to block matrices with square blocks of the same size, because in that case the block versions of D_1 and D_{-1} still consist of distinct entries. Problems arise when one is dealing with blocks of different size or, more general, rectangular blocks as in (9) and (13).

4.2. Rectangular Toeplitz blocks

Each $u_i \times v_j$ Toeplitz block $C_{ij}^{(S)}$ in (9) has dimensions

$$\begin{aligned} u_i &= M_i^{(S)} - m_i^{(S)} + 1, \\ v_j &= M_j^{(D)} - m_j^{(D)} + 1, \end{aligned}$$

and the displacement rank of the matrix given in (9) is at most

$$\alpha = 2 + \min\left(I^{(S)}, J^{(S)}\right) + \min\left(I^{(D)}, J^{(D)}\right), \qquad S = E \setminus N.$$

The block diagonal matrices L and R used as displacement operators are given by

$$L = \bigoplus_{t=0}^{\sigma} Z_{u_t}^{(1)} = \left(\bigoplus_{t=0}^{\sigma} Q_{1,u_t}^H\right) \oplus \left(\bigoplus_{t=0}^{\sigma} D_{1,u_t}\right) \oplus \left(\bigoplus_{t=0}^{\sigma} Q_{1,u_t}\right),$$

$$R = \bigoplus_{t=0}^{\delta} Z_{v_t}^{(-1)} = \left(\bigoplus_{t=0}^{\delta} Q_{-1,v_t}^H\right) \oplus \left(\bigoplus_{t=0}^{\delta} D_{-1,v_t}\right) \oplus \left(\bigoplus_{t=0}^{\delta} Q_{-1,v_t}\right),$$

and the factorization of $L\mathcal{T} - \mathcal{T}R$, where $\mathcal{T}$ now denotes the Toeplitz block matrix (9), into the product of a matrix G of size $(u_0 + \cdots + u_\sigma) \times \alpha$ and a matrix B of size $\alpha \times (v_0 + \cdots + v_\delta)$ can be written down explicitly. Through the left and right displacement operators each Toeplitz block $C_{ij}^{(S)}$ of (9) is transformed into a block of size $u_i \times v_j$ with mostly zeroes, except for the first row and last column. Let us denote the last column of this resulting block by $\gamma_{ij}^{(c)}$ and the first row (except for the element in the upper right corner) by $\gamma_{ij}^{(r)}$. Then $\gamma_{ij}^{(r)}$ is of size $1 \times (v_j - 1)$ and $\gamma_{ij}^{(c)}$ of size $u_i \times 1$. The matrix G consists of $\sigma + 1$ columns containing zeroes except for a 1 in one particular position and $\delta + 1$ columns composed of $\gamma_{ij}^{(c)}$:

$$G = \begin{pmatrix} & & & \gamma_{00}^{(c)} & \cdots & \gamma_{0\delta}^{(c)} \\ \delta_0 & \ldots & \delta_\sigma & \vdots & & \vdots \\ & & & \gamma_{\sigma 0}^{(c)} & \cdots & \gamma_{\sigma\delta}^{(c)} \end{pmatrix},$$

where

$$\begin{aligned} \delta_k &= \left(\delta_{i,1+\sum_{j=1}^{k} u_{j-1}}\right)_{m\times 1}, & k &= 0, \ldots, \sigma, \\ \delta_{i,j} &= 1 \Leftrightarrow i = j, & \sum_{j=1}^{0} &= 0. \end{aligned}$$

The matrix B is composed of similar pieces, namely the $\gamma_{ij}^{(r)}$ and at the end of each block (remember that the $\gamma_{ij}^{(r)}$ are one column too short) a column with mostly zeroes except for a 1 in position $\sigma + 1 + j + 1$ when adding the column to block number j where $j = 0, \ldots, \delta$:

$$B = \begin{pmatrix} \gamma_{00}^{(r)} & 0 & & \gamma_{0\delta}^{(r)} & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \gamma_{\sigma 0}^{(r)} & 0 & & \gamma_{\sigma\delta}^{(r)} & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & 0 & \ddots & \vdots & \vdots \\ & \vdots & & & 0 \\ 0 & 0 & & 0 & 1 \end{pmatrix}.$$

When a block row size u_t is even and a block column size v_k is odd, then for some integer i and j

$$v_k = \frac{u_t(2j+1)}{2i}, \qquad i = 0, \ldots, u_t - 1, \qquad j = 0, \ldots, v_k - 1,$$

and the block matrices $\bigoplus_{t=0}^{\sigma} D_{1,u_t}$ and $\bigoplus_{t=0}^{\delta} D_{-1,v_t}$ have common entries. In that case the particular block of u_t rows in (9) must be split into two blocks, respectively, of $u_t^{(1)}$ and $u_t^{(2)}$ rows, where $u_t^{(1)}$ and $u_t^{(2)}$ are odd. The consequence of this splitting technique, is that the size α of the factors G and B is incremented by one, each time we have to perform a split. Worst possible scenario is when all blocks of rows must be split into two subblocks and then the algorithm for the computation of the approximant (6) is $\mathcal{O}((\alpha + \sigma + 1)m^2)$ instead of $\mathcal{O}(\alpha m^2)$, where $\alpha = \sigma + \delta + 2$.

To allow a simple complexity comparison with Gaussian elimination, let us assume that we are dealing with a linear system of the form (9) where $\sigma = \delta$ and all $u_i = \ell = v_j$. Hence all blocks are square $\ell \times \ell$ blocks and the given system is immediately in lower block-triangular form. The fast algorithm detailed above then requires $\mathcal{O}\left((\sigma+1)^3\ell^2\right)$ operations. For a Toeplitz matrix we know that $\sigma = 0$, and that a fast algorithm executed on a Toeplitz block of size ℓ is indeed $\mathcal{O}(\ell^2)$. An optimal implementation of Gaussian elimination, only allowing pivoting per $\ell \times \ell$ block, needs $\mathcal{O}\left((\sigma+1)^2\ell^3\right)$ operations. In other words, when σ is too large, the block approach does not result in any complexity gain.

4.3. Block-Toeplitz-block case

To allow some simplicity in the notation, we only consider (13) for $s = 2$. Each $u_i \times v_j$ block $C_{n+i-j}^{(nm)}$ in (13) has dimensions

$$\begin{aligned} u_i &= nm + n + i + 1, \\ v_j &= nm + j + 1, \end{aligned}$$

and the displacement rank of the matrix given in (13) is at most $m + 1$. The displacement operators are as in the rectangular Toeplitz block case, with the only difference being the size of the Toeplitz blocks. But since now each Toeplitz block $C_{n+i-j}^{(nm+j)}$, where $i = 1, \ldots, m$ and $j = 0, \ldots, m$, is lower triangular, the resulting matrix $L\mathcal{T} - \mathcal{T}R$, where $\mathcal{T}$ now denotes the block-Toeplitz-block matrix (13), is even simpler. Its only nonzero entries occur in the last column of each block. Let us denote the last column of the transformed block by $\gamma_{n+i-j}^{(nm+j)}$. The factorization of $L\mathcal{T} - \mathcal{T}R$ then takes the form

$$L\mathcal{T} - \mathcal{T}R = \begin{pmatrix} \gamma_{n+1}^{(nm)} & \cdots & \gamma_{n-m+1}^{(nm+m)} \\ \vdots & & \vdots \\ \gamma_{n+m}^{(nm)} & \cdots & \gamma_{n}^{(nm+m)} \end{pmatrix} \cdot \left(\Delta_{1,v_0} \ \cdots \ \Delta_{m+1,v_m} \right),$$

where

$$\begin{aligned} \Delta_{j,v_{j-1}} &= \left(\epsilon_{i,k}\right)_{(m+1)\times v_{j-1}}, \\ \epsilon_{ik} &= \begin{cases} 1, & i = j, k = v_{j-1}, \\ 0, & \text{elsewhere.} \end{cases} \end{aligned}$$

The same splitting of row blocks should be applied as soon as one block row size is even and another block column size is odd. When splitting all block rows, the complexity of the fast algorithm, for the case $s = 2$ that we are discussing in detail, doubles.

4.4. Stability and reliability

When transforming the block structured Toeplitz matrix $\mathcal{T}$ into the block Cauchy-like matrix $\hat{\mathcal{T}}$, the solution of the original linear problem $\mathcal{T}b = c$ requires the LU factorization of $\hat{\mathcal{T}}$ and some additional matrix–vector computations. The algorithm to obtain the LU factorization of the Cauchy-like matrix $\hat{\mathcal{T}}$, given the exact factorization $\nabla\mathcal{T} = GB$ is detailed in [14]. Next, knowing from $\hat{\mathcal{T}} = PLU$ (where P is a permutation matrix) that

$$\mathcal{T} = \left(\bigoplus_{t=1}^{u} Q_{1,t}^{H} \right) PLU \left(\bigoplus_{t=1}^{v} Q_{-1,t} \right),$$

one computes

$$\begin{aligned} PLy &= \left(\bigoplus_{t=1}^{u} Q_{1,t} \right) c, \\ Uz &= y, \\ b &= \left(\bigoplus_{t=1}^{v} Q_{-1,t}^{H} \right) z. \end{aligned}$$

We denote the computed solution for b by $\tilde{b}$. As already explained, the algorithm from Gohberg et al. [14] is selected for generalization to a rectangular block structure, because it can be combined with a simple interval arithmetic verification step. The importance of such a verification step is pointed out in [13] and is mainly motivated by the fact that the truncation errors in multivariate rational approximation can vary extensively. It is therefore important to be able to separate the round-off errors from the truncation errors, if desired.

Having the LU decomposition of $\hat{\mathcal{T}}$ at our disposal, an approximate inverse $\mathcal{W}$ of $\mathcal{T}$ can be computed and the following interval arithmetic verification step can easily be performed. This refinement might not be possible with so-called superfast algorithms. The fixpoint of the iteration function

$$f(e) = \mathcal{W}\left(c - \mathcal{T}\tilde{b}\right) + (I - \mathcal{W}\mathcal{T})e$$

is the defect vector $e = \hat{b} - \tilde{b}$ where $\hat{b}$ is the exact solution of $\mathcal{T}b = c$ [32]. If $\mathcal{F}(E)$ denotes its interval extension and if for some interval E,

$$\mathcal{F}(E) \subset \breve{E}$$

where $\breve{E}$ denotes the interior of the interval E, then the linear system $\mathcal{T}b = c$ has one and only one solution in the interval $\tilde{b} + \breve{E}$.

For classical Gaussian elimination with partial pivoting performed on the full matrix $\mathcal{T}$ instead of on the factors G and B, the error in $\tilde{b}$, say the width $\text{diam}(\breve{E})$ of $\breve{E}$, is typically of the order of the product of the condition number of $\mathcal{T}$ and the machine epsilon $\frac{1}{2}\beta^{-t+1}$ where β and t, respectively, denote the radix and precision of the floating-point system in use. In table 2 we illustrate that the fast Gaussian elimination with partial pivoting performed on the factors G and B enjoys the same property, under the condition that (21) is not too small [1]. This is in fact an optimal result for a fast linear system solver. Here the value $\text{diam}(E)$, reflecting the uncertainty in the computation of the coefficient vector b, with $E = (E_1, \ldots, E_{n+m+1})$ is defined by

$$\text{diam}(E) = \sqrt{\sum_{i=1}^{n+m+1} \text{diam}(E_i)^2}.$$

Let us denote the matrix elements in the factors G and B of (20) by $G = (\gamma_{ij})$ and $B = (\beta_{ij})$. According to Brent [1], instabilities can occur if the size of the matrix elements $\left|\sum_{k=1}^{\delta+\nu+2} \gamma_{ik}\beta_{kj}\right|$ is small compared to that of the elements $\sum_{k=1}^{\delta+\nu+2} |\gamma_{ik}| \cdot |\beta_{kj}|$. Therefore, in table 2, the value

$$\min_{i,j=1,\ldots,n+m+1} \frac{\left|\sum_{k=1}^{\delta+\nu+2} \gamma_{ik}\beta_{kj}\right|}{\sum_{k=1}^{\delta+\nu+2} |\gamma_{ik}| \cdot |\beta_{kj}|} \tag{21}$$

Table 2 Some characteristics of $[n/2]_H^f(x-1, y-1)$.

n	dim	(21)	$\|r\|_2$	$\|r_{\text{norm}}\|_2$	$\kappa_2(\mathcal{T})$	diam(E)
2	17	1.7e−01	5.9e−16	8.6e−16	4.4e+02	8.5e−14
3	23	1.1e−01	8.0e−15	1.1e−15	6.2e+03	5.7e−12
4	29	5.4e−02	6.0e−14	2.5e−15	1.3e+05	6.1e−11
5	35	4.0e−03	6.1e−14	1.8e−15	7.9e+05	6.8e−10
6	41	5.6e−02	1.1e−13	2.1e−15	1.0e+07	1.1e−08
7	47	3.9e−02	1.7e−13	2.8e−15	4.0e+08	2.5e−07
8	53	5.4e−02	2.2e−13	3.0e−15	2.1e+09	3.0e−06
9	59	1.7e−03	3.2e−13	4.0e−15	7.6e+09	3.8e−06
10	65	2.6e−02	4.5e−13	5.0e−15	3.3e+10	1.7e−05
11	71	8.9e−03	4.6e−13	4.8e−15	1.4e+11	8.3e−05
12	77	1.2e−02	5.1e−13	5.1e−15	6.2e+11	3.4e−04
13	83	8.2e−03	9.1e−13	8.7e−15	2.6e+12	1.5e−03
14	89	1.9e−02	6.1e−13	5.7e−15	1.1e+13	7.3e−03

is tabulated together with the norms of the residue $r = \|c - \mathcal{T}b\|_2$ and normalized residue $r_{\text{norm}} = \|c - \mathcal{T}b\|_2/(\|\mathcal{T}\|_2\|b\|_2)$, and the ℓ_2 condition number $\kappa_2(\mathcal{T})$ of the matrix $\mathcal{T}$.

The numerical computations are carried out in MATLAB. The interval verification is done using the INTLAB toolbox for MATLAB [33]. The guaranteed accuracy of the computed solution is given in the last column of table 2, namely the width of the interval solution. This number is more pessimistic than the residual. The value $\log(\kappa_2(\mathcal{T})) - \log(\text{diam}(E))$ is about 15 to 16 for all dimensions shown, which is fully in accordance with what can be expected in double precision: The logarithm of the condition number indicates how many of the double precision 15 to 16 digits are probably lost during the computation, while the logarithm of $\text{diam}(E)$ indicates the remaining accuracy of the computed solution. The latter is guaranteed and not an approximation and can never be delivered by a floating-point only computation. It is a useful add-on to the fast solver for the structured system under consideration. Superfast solvers for structured systems which do not compute an estimate of the matrix inverse, cannot be combined with such an interval verification step.

5. Rational approximants for the beta function

Let us now apply this technique to the computation of some homogeneous as well as general bivariate Padé approximants to the beta function

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},$$

where Γ denotes the gamma function. The beta function is a nice example because of its meromorphy: The function has poles at $x = -k$ and $y = -k$ for $k \geq 1$. By means of the recurrence formulas

$$\begin{aligned}\Gamma(x+1) &= x\Gamma(x),\\ \Gamma(y+1) &= y\Gamma(y),\end{aligned}$$

for the gamma function, we can write

$$B(x, y) = \frac{1 + (x-1)(y-1)f(x-1, y-1)}{xy}. \tag{22}$$

Hence we can approximate f either by a homogeneous Padé approximant

$$f(x-1, y-1) \approx [n/m]_H^f(x-1, y-1),$$

or by a general lattice Padé approximant

$$f(x-1, y-1) \approx [N/D]_E(x-1, y-1),$$

and plug this approximant into (22). Because of the poles of $B(x, y)$ at $x = -1$ and $y = -1$, we are especially interested in the approximants $[n/2]_H^f$ and $[N_1/D]_{E_1}$ with

$$\begin{aligned}N_1 &= \{(i, j) \mid 0 \leq i + j \leq 2n - 1\},\\ D &= \{(i, j) \mid 0 \leq i, j \leq 1\},\\ E_1 &= N_1 \cup \{(2n, 0), (0, 2n), (n, n)\}.\end{aligned}$$

The set E_1 has been constructed such that it is symmetric because $B(x, y)$ is also a symmetric function. Since the computation of the latter approximants does not require the solution of a truly structured linear system, we have also computed the approximants $[N_2/N_2]_{E_2}$ with

$$\begin{aligned} N_2 &= \{(i, j) \mid 0 \le i + j \le 2n\}, \\ E_2 &= \{(i, j) \mid 0 \le i, j \le 2n\} \cup \{(i, 2n+1) \mid 0 \le i \le n-1\} \\ &\quad \cup \{(2n+1, j) \mid 0 \le j \le n-1\}. \end{aligned}$$

All approximants are evaluated in the points $(x, y) = (-0.75, -0.75)$ and $(x, y) = (-1.15, -1.15)$:

$$\begin{aligned} B(-0.75, -0.75) &= 9.88839827894065023\ldots, \\ B(-1.15, -1.15) &= -28.74143053667674\ldots. \end{aligned}$$

Note that the second evaluation point is outside the region of convergence of the Taylor series for $f(x, y)$, thereby illustrating a typical use of Padé approximants. We

Table 3 $B(-0.75, -0.75)$ and $B(-1.15, -1.15)$ using $\left[n/2\right]_H^f(x-1, y-1)$.

n	$\\|c - \mathcal{T}b\\|_2$	$\kappa_2(T)$	$x = y = -0.75$	s_1	$x = y = -1.15$	s_2
2	5.9e−16	4.4e+02	9.83346236474336	2	−39.03932482033827	1
3	8.0e−15	6.2e+03	9.87768132009373	3	−32.30502568791762	1
4	6.0e−14	1.3e+05	9.88693177128426	4	−29.54833276145216	2
5	6.1e−14	7.9e+05	9.88811820302491	5	−28.98652389451956	2
6	1.1e−13	1.0e+07	9.88829871186159	5	−28.86861999395085	3
7	1.7e−13	4.0e+08	9.88738384188594	4	−28.96366271894836	2
8	2.2e−13	2.1e+09	9.88840150418488	7	−28.72777552337503	4
9	3.2e−13	7.6e+09	9.88839845449305	8	−28.74046989041856	5
10	4.5e−13	3.3e+10	9.88839827723185	10	−28.74147377391598	6
11	4.6e−13	1.4e+11	9.88839827503527	10	−28.74149243021787	6
12	5.1e−13	6.2e+11	9.88839827789657	10	−28.74145492716109	6
13	9.1e−13	2.6e+12	9.88839827873497	11	−28.74143791214454	7
14	6.1e−13	1.1e+13	9.88839827890217	12	−28.74143252382255	7
15	1.2e−12	4.6e+13	9.88839827892984	12	−28.74143104924252	8
16	1.3e−12	1.9e+14	9.88839827894367	13	−28.74143066829773	9
17	1.1e−12	7.8e+14	9.88839827895234	12	−28.74143057152613	9
18	8.0e−13	3.2e+15	9.88839827892595	12	−28.74143054573386	10
19	3.4e−13	1.4e+16	9.88839827898111	12	−28.74143053903674	10
20	1.9e−14	7.0e+16	9.88839827899346	11	−28.74143053788137	10
21	8.1e−15	1.4e+17	9.88839827886541	11	−28.74143054416083	10
22	2.3e−15	9.1e+16	9.88839827921346	11	−28.74143053672243	11
23	2.4e−15	3.3e+17	9.88839827893152	12	−28.74143053745456	11
24	3.6e−16	1.5e+17	9.88839827890935	12	−28.74143053593324	11
25	2.0e−15	1.8e+17	9.88839827898606	11	−28.74143050379619	9
26	6.8e−16	1.4e+18	9.88839827889521	11	−28.74143053812084	10
27	4.9e−16	2.8e+17	9.88839827894266	13	−28.74143053604590	11
28	1.7e−16	3.2e+18	9.88839827893794	13	−28.74143053484292	10
29	2.2e−16	1.1e+18	9.88839827892942	12	−28.74143038973866	8
30	1.9e−16	8.7e+17	9.88839827894102	14	−28.74143056932070	9

Table 4 $B(-0.75, -0.75)$ and $B(-1.15, -1.15)$ using $\left[N_1/D\right]_{E_1}(x-1, y-1)$.

n	$\|c - \mathcal{T}b\|_2$	$\kappa_2(\mathcal{T})$	$x = y = -0.75$	s_1	$x = y = -1.15$	s_2
2	1.2e−17	3.5e+00	10.18357300993597	2	−187.56480604351407	0
3	3.3e−18	2.6e+00	9.90310445661817	3	−23.94480772366405	1
4	4.3e−19	3.7e+00	9.89243920147501	4	−21.08433539555264	1
5	8.8e−20	3.1e+01	9.88918069326081	4	−20.61206791620240	1
6	2.2e−20	4.1e+02	9.88853031217559	5	1.35867575264281	0
7	5.8e−21	7.9e+03	9.88841931455667	6	−30.93338729208869	1
8	1.5e−21	2.1e+05	9.88840152374231	7	−29.15422184026320	2
9	2.7e−22	7.1e+06	9.88839876874271	8	−28.84734775250110	3
10	1.2e−22	3.1e+08	9.88839838781188	8	−28.75863359997508	3
11	8.1e−23	1.8e+09	9.88839826005179	9	−28.74363873660080	4
12	2.2e−22	6.6e+08	9.88839827474591	10	−28.74224179067835	5
13	1.7e−24	7.5e+06	9.88839827859541	11	−28.74166653238182	5
14	3.6e−24	3.3e+06	9.88839827885322	11	−28.74152008089442	6
15	1.0e−23	7.5e+05	9.88839827892942	12	−28.74145962997683	6
16	2.6e−26	5.0e+03	9.88839827893958	13	−28.74143931551998	7

also display the condition number $\kappa_2(\mathcal{T})$ of the structured linear systems (13) and (9), the residues and the number of significant digits, respectively, denoted by s_1 and s_2.

The factorization (17) introduces small imaginary parts in the computation that are the result of round-off error accumulation. In none of the output tables these imaginary parts are displayed. In table 3 the magnitude of these imaginary parts is at most 10^{-11}, in table 4 at most 10^{-14} and in table 5 it increases to 10^{-6} and then decreases again, further down the table.

Table 5 $B(-0.75, -0.75)$ and $B(-1.15, -1.15)$ using $\left[N_2/N_2\right]E_2(x-1, y-1)$.

n	$\|c - \mathcal{T}b\|_2$	$\kappa_2(\mathcal{T})$	$x = y = -0.75$	s_1	$x = y = -1.15$	s_2
1	2.2e−15	1.0e+02	9.29971366877699	1	−65.90763165486295	0
2	1.8e−16	2.8e+03	10.09253897188655	2	−53.08797811334378	0
3	2.9e−16	6.4e+06	9.88866906226073	5	−28.36916023709587	2
4	1.7e−15	1.9e+09	9.88839828126046	10	−28.74112996396373	5
5	1.4e−15	1.9e+12	9.88839827830668	10	−28.74144115453159	7
6	5.4e−14	5.1e+16	9.88839827840003	10	−28.74143971365216	7
7	1.5e−13	3.2e+17	9.88839827864923	11	−28.74143650209453	7
8	3.9e−13	7.9e+17	9.88839827889932	12	−28.74143183384669	8
9	1.5e−12	1.0e+18	9.88839827891820	12	−28.74143131999405	8
10	5.5e−12	3.3e+17	9.88839827894547	12	−28.74143017513286	8
11	9.8e−13	6.5e+18	9.88839827894271	12	−28.74143057484249	7
12	7.4e−14	1.6e+19	9.88839827894081	14	−28.74143049588934	9
13	3.2e−15	3.1e+19	9.88839827894061	15	−28.74143055690920	9
14	4.4e−15	6.2e+22	9.88839827894066	15	−28.74143055029435	10
15	1.1e−15	9.7e+22	9.88839827894063	15	−28.74143055349034	9
16	0.0e+00	2.1e+23	9.88839827894057	14	−28.74143238435482	7

The informational usage N_d of the three approximation techniques is comparable for qualitatively equal results. For instance: The computation of

- $[n/2]_H^f$ for $n = 27$,
- $[N_1/D]_{E_1}$ for $2n - 1 = 29$,
- $[N_2/N_2]_{E_2}$ for $2n = 20$,

needs 465, 468 and 461 c_{ij}-coefficients, respectively, varying in magnitude from 10^0 to 10^{-19}. Evaluated in $(x, y) = (-0.75, -0.75)$ these approximants all yield 12 to 13 significant digits.

References

[1] Brent, R.P.: Stability of fast algorithms for structured linear systems. In: Kailath, T., Sayed, A.H. (eds.) Fast Reliable Algorithms for Matrices with Structure, pp. 103–116. SIAM, Philadelphia, Pennsylvania (1999)

[2] Bultheel, A.: Epsilon and qd algorithms for matrix- and 2D-Padé approximation. Technical Report TW57, Department of Computer Science, K.U. Leuven (1982)

[3] Bultheel, A.: Low displacement-rank problems in 2-D Padé approximation. In: Outils et modèles mathématiques pour l'automatique, l'analyse de systèmes et le traitement du signal, vol. 2, pp. 563–576. Editions du CNRS, Paris (1982)

[4] A. Bultheel. Recursive computation of triangular 2D-Padé approximants. Technical report, Department of Computer Science, K.U. Leuven (1982)

[5] Bunch, J.R.: Stability of methods for solving Toeplitz systems of equations. SIAM J. Sci. Statist. Comput. **6**(2), 349–364 (1985)

[6] Chaffy, C.: $(\text{Padé})_y$ of $(\text{Padé})_x$ approximants of $F(x, y)$. In: Cuyt, A. (ed.) Nonlinear Numerical Methods and Rational Approximation (Wilrijk, 1987), pp. 155–166. Reidel, Dordrecht (1988)

[7] Chisholm, J.S.R.: Rational approximants defined from double power series. Math. Comput. **27**, 841–848 (1973)

[8] Critchfield, C.L., Gammel, J.L.: Rational approximants for inverse functions of two variables. Rocky Mt. J. Math. **4**, 339–349 (1974)

[9] Cuyt, A.: Padé Approximants for Operators: Theory and Applications, vol. 1065 of Lecture Notes in Mathematics. Springer, Berlin Heidelberg New York (1984)

[10] Cuyt, A.: How well can the concept of Padé approximant be generalized to the multivariate case? J. Comput. Appl. Math. **105**(1–2), 25–50 (1999)

[11] Cuyt, A., Verdonk, B.: General order Newton–Padé approximants for multivariate functions. Numer. Math. **43**(2), 293–307 (1984)

[12] Cuyt, A., Verdonk, B.: A review of branched continued fraction theory for the construction of multivariate rational approximants. Appl. Numer. Math. **4**(2–4), 263–271 (1988)

[13] Cuyt, A., Verdonk, B.: The need for knowledge and reliability in numeric computation: Case study of multivariate Padé approximation. Acta Appl. Math. **33**, 273–302 (1993)

[14] Gohberg, I., Kailath, T., Olshevsky, V.: Fast Gaussian elimination with partial pivoting for matrices with displacement structure. Math. Comput. **64**(212), 1557–1576 (1995)

[15] Gohberg, I., Olshevsky, V.: Complexity of multiplication with vectors for structured matrices. Linear Algebra Appl. **202**, 163–192 (1994)

[16] Gončar, A.A.: A local condition for the single-valuedness of analytic functions of several variables. Math. USSR Sb. **22**(2), 305–322 (1974)

[17] Graves-Morris, P.R., Hughes Jones, R., Makinson, G.J.: The calculation of some rational approximants in two variables. J. Inst. Math. Appl. **13**, 311–320 (1974)

[18] Guillaume, P.: Nested multivariate Padé approximants. J. Comput. Appl. Math. **82**(1–2), 149–158 (1997)

[19] Heinig, G.: Inversion of generalized Cauchy matrices and other classes of structured matrices. In: Linear algebra for signal processing (Minneapolis, MN, 1992), vol. 69 of IMA Vol. Math. Appl., pp. 63–81. Springer, Berlin Heidelberg New York(1995)

[20] Hughes Jones, R.: General rational approximants in N-variables. J. Approx. Theory **16**(3), 201–233 (1976)

[21] Hughes Jones, R., Makinson, G.J.: The generation of Chisholm rational polynomial approximants to power series in two variables. J. Inst. Math. Appl. **13**, 299–310 (1974)

[22] Kailath, T., Kung, S.Y., Morf, M.: Displacement ranks of matrices and linear equations. J. Math. Anal. Appl. **68**(2), 395–407 (1979)

[23] Karan, B.M., Srivastava, M.C.: A new multidimensional rational approximant. J. Indian Inst. Sci. **67**(9–10), 351–360 (1987)

[24] Karlsson, J., Wallin, H.: Rational approximation by an interpolation procedure in several variables. In: Saff, E.B., Varga, R.S. (eds.) Padé and Rational Approximation: Theory and Applications (Proc. Internat. Sympos., Univ. South Florida, Tampa, Fla., 1976), pp. 83–100. Academic Press, Berlin Heidelberg New York (1977)

[25] Kuchminskaya, K.I.: Corresponding and associated branching continued fractions for a double power series. Dokl. Akad. Nauk Ukr. SSR, Ser. A **7**, 613–617, 669 (1978)

[26] Levin, D.: General order Padé-type rational approximants defined from double power series. J. Inst. Math. Appl. **18**(1), 1–8 (1976)

[27] Levy, B., Morf, M., Kung, S.-Y.: Some algorithms for the recursive input-output modeling of 2-d systems. Technical Report LIDS-P-962, MIT, Cambridge (1979)

[28] Lutterodt, C.H.: A two-dimensional analogue of Padé approximant theory. J. Phys. A **7**, 1027–1037 (1974)

[29] Lutterodt, C.H.: Addendum to: "A two-dimensional analogue of Padé approximant theory" (J. Phys. A **7** (1974), 1027–1037). J. Phys. A **8**, 427–428 (1975)

[30] Lutterodt, C.H.: Rational approximants to holomorphic functions in n-dimensions. J. Math. Anal. Appl. **53**(1), 89–98 (1976)

[31] Paraskevopoulos, P.N.: Padé-type order reduction of two-dimensional systems. IEEE Trans. Circuits Syst. **27**, 413–416 (1980)

[32] Rump, S.M.: Kleine Fehlerschranken bei Matrixproblemen. PhD thesis, Universität Karlsruhe (1980)

[33] Rump, S.M.: INTLAB – INTerval LABoratory. In: Csendes, T. (ed.) Developments in Reliable Computing, pp. 77–104. Kluwer, Dordrecht (1999)