Multivariate Padé approximants: homogeneous or not, that's the question

Stefan Becuwe*† Annie Cuyt*‡

Key words: Padé approximation, Toeplitz-block matrix, displacement structure, linear equations, sparse matrices.

AMS subject classifications: 41A21, 15A06, 15A57, 65F20, 65F50.

1 Introduction

What we know about multivariate Padé approximants has been developed in the past 25 years. In [7] the results are reviewed and the various definitions for multivariate Padé approximants are grouped into four main categories (we only mention the main publications here):

- definitions based on an appropriate choice for the defining equations among the candidate equations [11, 13, 15, 17];
- definitions based on generalizations of the continued fraction concept [16];
- partly symbolic or symbolic-numeric approaches [1, 12];
- and last but not least a multivariate homogeneous definition [4, 18].

This last definition exhibits a great similarity with that of the univariate Padé approximant, in the sense that the traditional properties such as the existence and unicity of a multivariate irreducible form and several covariance properties remain true [7], the traditional ε- and qd-algorithms remain valid [2, 5], and classical convergence theorems such as 'de Montessus de Ballore' [9] and 'Nuttall-Pommerenke' [8] can be proved.

However, at the same time an intriguing question of 'insight' remains open with respect to the multivariate homogeneous Padé approximant. Which multivariate mechanism is responsable for the fact that, although overdetermined, its defining system of homogeneous linear equations can guarantee that it always delivers a nontrivial solution? The proof of this fact dates back to 1982 [4, 6], but the linear dependence among the defining equations has never been understood. We will clarify this open problem here and elaborate on the structure of the linear system which appears to be very sparse and have a low displacement rank.

2 The multivariate homogeneous definition

Given a univariate function $f(z)$ through its Taylor series expansion at a certain point in the complex plane (for simplicity we use the Taylor series at the origin),

$$f(z) = \sum_{i=0}^{\infty} c_i z^i,$$

*Universiteit Antwerpen (UIA), Departement Wiskunde en Informatica, Universiteitsplein 1, B-2610 Antwerpen, Belgium
†e-mail: Stefan.Becuwe@uia.ua.ac.be
‡Research Director FWO-Vlaanderen; e-mail: Annie.Cuyt@uia.ua.ac.be

16th IMACS World Congress (© 2000 IMACS)

the Padé approximant $[n/m]^f$ of degree n in the numerator and m in the denominator for f is defined by

$$p(z) = \sum_{i=0}^{n} a_i z^i, \qquad q(z) = \sum_{i=0}^{m} b_i z^i,$$
$$(fq - p)(z) = \sum_{i \geq n+m+1} d_i z^i,$$

with $[n/m]^f$ equal to the irreducible form of p/q. The conditions $d_i = 0$, $i = n+1, \ldots, n+m$ give rise to a Toeplitz linear system of equations:

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

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

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

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 a univariate term of degree k. 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. Interchanging the role of x and y in this approach does not necessarily lead to the same results. Since the 'continued fraction' approach does not compute its multivariate approximant from a defining system of equations for the numerator and denominator coefficients, we do not involve this generalization in the discussion.

The main difference between the 'equation lattice' and the 'symbolic' 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 which 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 will be explained and emphasized at several occasions in the sequel of the text.

The above distinction between the different approaches however does not imply that for the construction of similar multivariate Padé approximants the number N_d of data c_{ij} to be obtained from the function f is different. The informational usage is the same for all definitions and is computed in section 4 where we also deal in more detail with the redundancy of the problem. The fact that the informational usage remains equal between comparable multivariate Padé approximants implies that the homogeneous definition, that involves polynomials with more coefficients than

the traditional definitions, gives rise to a sparse system. To summarize, we will show that the homogeneous definition involves the solution of a very sparse block-Toeplitz-block system, while this is not the case for other definitions where traditionally $N_u = N_d + 1 = N_e + 1$.

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

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

are then computed from the conditions

$$(fq - p)(x,y) = \sum_{i+j \geq nm+n+m+1} d_{ij} x^i y^j \tag{2}$$

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

$$\begin{cases} C_{n+1}(x,y)B_0(x,y) + \cdots + C_{n+1-m}(x,y)B_m(x,y) \equiv 0 \\ \quad \vdots \\ C_{n+m}(x,y)B_0(x,y) + \cdots + C_n(x,y)B_m(x,y) \equiv 0 \end{cases} \tag{3}$$

with $C_k(x,y) \equiv 0$ if $k < 0$. This is exactly the system of defining equations (1) for univariate Padé approximants if the term $c_k x^k$ in the univariate definition is substituted by

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

3 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 a certain order:

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

When we arrange the conditions (3) in a similar (upward sloping diagonal) way and when we introduce the Toeplitz blocks

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

then the coefficient matrix of the system of equations (3) is

$$\begin{pmatrix} C_{n+1}^{(nm)} & C_n^{(nm+1)} & \ldots & C_{n-m+1}^{(nm+m)} \\ C_{n+2}^{(nm)} & \ldots & & \\ \vdots & & & \vdots \\ C_{n+m}^{(nm)} & \ldots & & C_n^{(nm+m)} \end{pmatrix} \tag{4}$$

which is very similar to (1). Behind each entry $C_i^{(j)}$ in this Toeplitz-structured matrix unfolds a simpler Toeplitz matrix.

This is only the bivariate case. Now let's 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 (1) is transformed into (4) which looks identical until we 'open' each entry $C_i^{(j)}$ and find that there's another Toeplitz matrix inside. We now describe the transition from two to more variables and focus on what happens if each of the $c_{k\ell}$ inside each $C_i^{(j)}$ is again 'opened'.

Let us denote the number of variables by s and let us denote by 0_t a sequence of t zero indices. The generalization of (2) and (3) 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 (3). 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's 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 (3) 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 $\binom{s+k-1}{k}$ coefficients $b_{i_1\ldots i_s}$. So the total of unknown denominator coefficients $b_{i_1\ldots i_s}$ equals

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

The k^{th} equation in (3) 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

$$N_e = \sum_{k=1}^{m} \binom{s+nm+n+k-1}{nm+n+k} \tag{5}$$

If $nm > 0$ then

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

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. Nevertheless it has been proven in [4] and [6, pp. 60–62] that a nontrivial solution always exists. It is therefore unnecessary to consider the linear conditions (2) in a least squares sense. The inherent dependence among the homogeneous Padé approximation conditions is however not well understood and will be analyzed in detail in the sequel.

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 columnn of each $C_i^{(j)}$ in (4) and expand $\left(c_{i,0,0_{s-2}}, \ldots, c_{0,i,0_{s-2}}\right)$ in the same way as the subvector $(b_{i,0,0_{s-2}}, \ldots, b_{0,i,0_{s-2}})$. During the unfolding process,

the size of each Toeplitz block at each step in the process can be determined from the following: the k^{th} term in N_e given by (5) is linked to the block entries in the k^{th} row of (4) 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 sub-blocks of row size $\binom{(s-1)+\ell-1}{\ell}$. Take for instance $n=1$ and $m=2$ when $s=3$ and construct the 3-dimensional analogue of the upper left block $C_{n+1}^{(nm)}$ of (4). 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 $(b_{200} \mid b_{110}, b_{101} \mid b_{020}, b_{011}, b_{002})$ which arranges the unknown coefficients of $B_0(x_1, x_2, x_3)$ when $nm = 2$. The size of each compartment in this last vector determines the column size of the rectangular Toeplitz blocks in the 3-dimensional analogue of (4) 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 will take the place of the entry on row 3 and column 2 of $C_2^{(2)}$ will have 3 rows because the entry on row 3 in the first column unfolded to $(c_{020}, c_{011}, c_{002})$. It will have 2 columns because in the vector of unknowns, that is multiplied with the coefficient matrix, the second compartment contains 2 elements. Also the total number of rows of the 3-dimensional analogue of $C_2^{(2)}$ is given by the first term ($k=1$) of (5) which equals 15. As explained these 15 rows split up in 5 smaller constructions according to

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

Hence we finally obtain for the 3-dimensional analogue of $C_{n+1}^{(nm)}$ with $n=1$ and $m=2$:

(6)

$c_{2,0,0}$	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	$c_{0,0,2}$

4 Redundancy, sparsity and displacement rank

When computing the actual size $N_e \times N_u$ of the coefficient matrix of (3), it is apparent that as the number of variables grows, the system is soon very much overdetermined. For instance for $s=4$, $n=3$ and $m=4$ we have $N_e = 4979$ and $N_u = 3480$. When inspecting the coefficient matrix it is also clear that it is very sparse and at the same time highly structured so that very efficient techniques for the solution of the linear system can be applied [10]. In a first attempt to get a grip on the redundant equations in (3), we tried to pinpoint the $N_e - N_u + 1$ linear dependent equations.

Although the structure is responsible for the redundancy, the linear dependent equations did not show up at specified entries in the matrix. For instance for $s = 3, n = 1$ and $m = 1$ the 10×9 symbolic 3-dimensional (symbolic entries c_{ijk}) homogeneous system was solved using the fraction free Gaussian elimination function of Maple V Release 5.1. The system has rank 8 and every combination (45 in total) of 8 equations out of the 10 imposed ones delivers rank 8. The experiment was repeated for $s = 3, n = 1$ and $m = 2$. It is not so that particular rows in the matrix constitute the linear dependent equations. In order to reduce the size of the overdetermined linear system, another strategy has to be followed.

Since the coefficient matrix is highly structured, one of course does not want to eliminate equations that destroy 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 linear dependent equations. In the case $s = 4$, $n = 3$ and $m = 4$ one for instance has to remove 1500 equations from the overdetermined system before it can be passed to a solver. When inspecting the coefficient matrix, one counts 1420 trailing zero entries in the first column. This part could be cut away, but another 80 equations will have to be eliminated higher up, with minimal influence on the structure. Here minimal effect on the structure means without increasing the displacement rank.

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. From (3) 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

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

we obtain $N_u > N_d$. Hence it is always possible to cut away equations without cutting away data. From the structure of (6) it is also possible to compute the sparsity of the block-Toeplitz-block matrix. The full matrix is of size $N_e \times N_u$ with at most $N_d \times N_u$ nonzero entries. Hence only a fraction of at most N_d/N_e in the matrix is filled. After removing any redundant equations, still a fraction of less than $N_d/(N_u - 1)$ in the matrix is nonzero. For $s = 4, n = 3$ and $m = 4$ the actual ratio is for instance 3.4% while for $s = 6, n = 5$ and $m = 5$ it further reduces to 0.18%.

The concept of displacement rank was first introduced in [14]. We use the definition given in [10] where the displacement rank of a matrix T is defined as the rank of the matrix $LT - TR$ with L and R being so-called displacement operators. For a Toeplitz-block matrix with u block rows and v block columns and rectangular Toeplitz blocks of size $u_i \times v_j$ the displacement operators

$$L = \bigoplus_{k=1}^{u} Z_{u_k}^{(1)} \qquad R = \bigoplus_{k=1}^{v} Z_{v_k}^{(-1)}$$

are used, where $\oplus W_k$ denotes the block diagonal matrix of which the k^{th} block is given by W_k and where

$$Z_k^{(\alpha)} = \begin{pmatrix} 0 & \ldots & \ldots & 0 & \alpha \\ 1 & 0 & \ldots & \ldots & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 & 0 \end{pmatrix}$$

When this definition is applied to a Toeplitz matrix ($u = 1 = v$) then the resulting matrix $LT - TR$ only consists of the first row and last column of T. Hence the displacement rank of a Toeplitz matrix equals 2. When applied to the coefficient matrix of (3) the resulting matrix consists of the first row and the last column of each Toeplitz 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

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

For $s = 2$, the displacement rank of (4) equals at most $m + 1$.

References

[1] C. Chaffy. (Padé)y of (Padé)x approximants of $F(x, y)$. In A. Cuyt, editor, *Nonlinear numerical methods and rational approximation*, pages 155–166, 1988.

[2] A. Cuyt. The epsilon-algorithm and multivariate Padé approximants. *Numer. Math.*, 40:39–46, 1982.

[3] A. Cuyt. A comparison of some multivariate Padé approximants. *SIAM J. Math. Anal.*, 14(1):195–202, 1983.

[4] A. Cuyt. Multivariate Padé approximants. *J. Math. Anal. Appl.*, 96:283–293, 1983.

[5] A. Cuyt. The qd-algorithm and multivariate Padé approximants. *Numer. Math.*, 42:259–269, 1983.

[6] A. Cuyt. *Padé approximants for operators: theory and applications*, volume 1065 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1984.

[7] A. Cuyt. How well can the concept of Padé approximant be generalized to the multivariate case? *J. Comput. Appl. Math.*, 105:25–50, 1999. Proceedings CONFUN, Trondheim, June 1998.

[8] A. Cuyt, K. Driver, and D. Lubinsky. Nuttall-Pommerenke theorems for homogeneous Padé approximants. *J. Comput. Appl. Math.*, 67:141–146, 1996.

[9] A. Cuyt and D. Lubinsky. A de Montessus theorem for multivariate homogeneous Padé approximants. *Ann. Numer. Math.*, 4:217–228, 1997.

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

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

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

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

[14] T. Kailath, S.-Y. Kung, and M. Morf. Displacement rank of a matrix. *Bull. Amer. Math. Soc.*, 1:769–773, 1979.

[15] J. Karlsson and H. Wallin. Rational approximation by an interpolation procedure in several variables. In E. Saff et al., editors, *Padé and rational approximation*, pages 83–100, 1977.

[16] K. Kuchminskaya. Corresponding and associated branching continued fractions for the double power series. *Dokl. Akad. Nauk Ukrain. SSR Ser. A*, 7:613–617, 1978.

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

[18] H. Werner. Multivariate Padé approximation. *Numer. Math.*, 48:429–440, 1986.