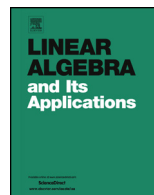




Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Linear Algebra and its Applications

journal homepage: www.elsevier.com/locate/laa



Accurate bidiagonal factorization of quantum Hilbert matrices [☆]



E. Mainar, J.M. Peña, B. Rubio ^{*}

Departamento de Matemática Aplicada/IUMA, Universidad de Zaragoza, Spain

ARTICLE INFO

Article history:

Received 16 May 2023

Received in revised form 27 October 2023

Accepted 27 October 2023

Available online 7 November 2023

Submitted by V. Mehrmann

MSC:

65F05

65F15

15A18

15A06

15A018

15A023

Keywords:

High relative accuracy

Bidiagonal decompositions

Totally positive matrices

Hilbert matrices

q-integers

Quantum Hilbert matrices

ABSTRACT

A bidiagonal decomposition of quantum Hilbert matrices is obtained and the total positivity of these matrices is proved. This factorization is used to get accurate algebraic computations with these matrices. The numerical errors due to imprecise computer arithmetic or perturbed input data in the computation of the factorization are analyzed. Numerical experiments show the accuracy of the proposed methods.

© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

[☆] This work was partially supported by Spanish research grants PGC2018-096321-B-I00 and RED2022-134176-T (MCI/AEI) and by Gobierno de Aragón (E41_23R).

^{*} Corresponding author.

E-mail address: brubio@unizar.es (B. Rubio).

<https://doi.org/10.1016/j.laa.2023.10.026>

0024-3795/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Hilbert matrices $H_n := (1/(i + j - 1))_{1 \leq i, j \leq n+1}$, were introduced by Hilbert in [8], obtaining the following expression for their determinant

$$\det H_n = \left(\prod_{k=1}^n (2k + 1) \binom{2k}{k} \right)^{-1}. \quad (1)$$

In Numerical Linear Algebra, Hilbert matrices are well-known Hankel (or catalecticant) matrices, that is, square matrices such that each ascending skew-diagonal from left to right takes a constant value. Their inverses H_n^{-1} have integer entries and their integer expression was provided in [3]. Moreover, the inverse of the one-parameter extension of the Hilbert matrices given by $H_n^{(\alpha)} := (\alpha/(i + j + \alpha - 2))_{1 \leq i, j \leq n+1}$, for $\alpha > 0$, was also obtained in [3].

In the literature, we can find combinatorial Hankel matrices, which are interesting extensions or analogues of Hilbert matrices, and are obtained by considering binomial coefficients, the Gaussian q -binomial coefficients or well-known sequences of integer numbers. Hankel matrices are well studied objects in mathematics with applications in various fields such as orthogonal polynomials, random matrices or operator theory. Hankel matrices are usually used to characterize the solution of classical moment problems. The Hilbert matrices are the Hankel matrices with respect to the moment sequence

$$s_n := \int_0^1 x^n dx,$$

and the corresponding orthogonal polynomials are the Legendre polynomials for the interval $[0, 1]$. In recent years q -calculus has been studied rigorously because of its latent application in Mathematics, Mechanics and Physics. By considering quantum integers, quantum Hilbert matrices

$$H_n^{(\alpha, q)} := \left(\frac{[\alpha]_q}{[i + j + \alpha - 2]_q} \right)_{1 \leq i, j \leq n+1}$$

were introduced in [1] as Hankel matrices with respect to a moment sequence obtained by considering a q -analogue of integration for a probability measure on $[0, 1]$, which for $q \rightarrow 1$ converges weakly to the Lebesgue measure on that interval. The corresponding orthogonal polynomials are certain little q -Jacobi polynomials. So, the quantum Hilbert matrix $H_n^{(1, q)}$ converges to the ordinary Hilbert matrices H_n when $q \rightarrow 1$. For the particular value $q = (1 - \sqrt{5})/(1 + \sqrt{5})$, quantum Hilbert matrices are closely related to Hilbert matrices (cf. [18]).

Hilbert matrices and these analogues are very ill-conditioned for moderate values of their dimension, although within structured perturbations better results can be expected

(cf. [9,12,10]). Under general perturbations, standard routines implementing best traditional numerical methods for computing their singular values, inverses or the solution of linear systems of equations with this Hilbert-type coefficient matrices do not obtain accurate results. So, the design and analysis of procedures to high relative accuracy, achieving relative errors of the order of the machine precision, regardless of the dimension or the conditioning of the considered algebraic problem, has attracted the interest of many researchers.

It is well known that Hilbert matrices H_n are strictly totally positive, that is, all their minors are positive. In [16], the pivots and the multipliers of the Neville elimination of H_n are explicitly derived and a bidiagonal factorization, accurately computed in $O(n^2)$ time, can be found in formulae (3.6) of Section 3 of [14]. Using this factorization, computations to high relative accuracy have been achieved for the resolution of algebraic problems with Hilbert matrices (cf. [14,16]). This paper describes the Neville elimination process for quantum Hilbert matrices. As a consequence, a bidiagonal factorization of these matrices is deduced and used to prove their total positivity and provide accurate computations in the resolution of algebraic problems related to these matrices.

In order to make this paper as self-contained as possible, Section 2 recalls basic concepts and results related to total positivity, Neville elimination and high relative accuracy. Section 3 provides the pivots and multipliers of the Neville elimination of quantum Hilbert matrices. As a result, it is obtained an expression for the determinant of quantum Hilbert matrices, generalizing the well-known formula (1) for the case of Hilbert matrices. Moreover, a bidiagonal factorization for quantum Hilbert matrices is also derived, allowing the analysis of their total positivity, as well as the resolution, for $q \in (0, 1]$, of algebraic problems with these matrices to high relative accuracy. The numerical errors appearing in the computation with a floating-point arithmetic of this factorization are studied in Section 4 and a structured condition number for quantum Hilbert matrices is deduced. Finally, Section 5 illustrates the numerical performed experimentation.

2. Notations and auxiliary results

An algorithm for the resolution of an algebraic problem is performed to high relative accuracy in floating-point arithmetic if the relative errors in the computations have the order of the unit round-off (or machine precision), without being affected by the dimension or the conventional conditionings of the problem. It is well known that algorithms to high relative accuracy are those avoiding subtractive cancellations, that is, only requiring the following arithmetics operations: products, quotients, and additions of numbers of the same sign (see page 52 in [4]). Moreover, if the floating-point arithmetic is well-implemented, the subtraction of initial data can also be done without losing high relative accuracy (see page 53 in [4]).

We say that a matrix is totally positive if all its minors are nonnegative and strictly totally positive if all its minors are positive. Computations to high relative accuracy with totally positive matrices can be achieved by means of a proper representation of

the matrices in terms of bidiagonal factorizations, which is in turn closely related to their Neville elimination (cf. [5–7]).

The essence of the Neville elimination procedure is to make zeros in a column of a given matrix $A \in \mathbb{R}^{(n+1) \times (n+1)}$ by adding to each row an appropriate multiple of the previous one. In every major step, the Neville elimination calculates a matrix $A^{(k+1)}$, $k = 2, \dots, n$, from the matrix $A^{(k)}$, previously obtained, with $A^{(1)} := A$. In more detail, $A^{(k+1)}$ is computed from $A^{(k)}$ according to the following formula

$$a_{i,j}^{(k+1)} := \begin{cases} a_{i,j}^{(k)}, & \text{if } 1 \leq i \leq k, \\ a_{i,j}^{(k)} - \frac{a_{i,k}^{(k)}}{a_{i-1,k}^{(k)}} a_{i-1,j}^{(k)}, & \text{if } k + 1 \leq i, j \leq n + 1, \text{ and } a_{i-1,j}^{(k)} \neq 0, \\ a_{i,j}^{(k)}, & \text{if } k + 1 \leq i \leq n + 1, \text{ and } a_{i-1,k}^{(k)} = 0. \end{cases} \quad (2)$$

The process finishes when $U := A^{(n+1)}$ is an upper triangular matrix. The entry

$$p_{i,j} := a_{i,j}^{(j)}, \quad 1 \leq j \leq i \leq n + 1, \quad (3)$$

is the (i, j) pivot and $p_{i,i}$ is called the i -th diagonal pivot of the Neville elimination of A . The Neville elimination of A can be done without row exchanges if all the pivots are nonzero. Then, the value

$$m_{i,j} := a_{i,j}^{(j)} / a_{i-1,j}^{(j)} = p_{i,j} / p_{i-1,j}, \quad 1 \leq j < i \leq n + 1, \quad (4)$$

is called the (i, j) multiplier. The complete Neville elimination of A consists of performing the Neville elimination to obtain the upper triangular matrix $U = A^{(n+1)}$ and next, the Neville elimination of the lower triangular matrix U^T .

Neville Elimination is a nice tool to deduce that a given matrix is STP, as shown in this characterization derived from Theorem 4.1, Corollary 5.5 of [5] and the arguments of p. 116 of [7].

Theorem 1. *A given nonsingular matrix A is STP (resp., TP) if and only if the Neville elimination of A and A^T can be performed without row exchanges, all the multipliers of the Neville elimination of A and A^T are positive (resp., nonnegative), and the diagonal pivots of the Neville elimination of A are all positive.*

In [7], it is shown that a nonsingular totally positive matrix $A \in \mathbb{R}^{(n+1) \times (n+1)}$ can be decomposed as follows,

$$A = F_n F_{n-1} \cdots F_1 D G_1 G_2 \cdots G_n, \quad (5)$$

where $F_i \in \mathbb{R}^{(n+1) \times (n+1)}$ (respectively, $G_i \in \mathbb{R}^{(n+1) \times (n+1)}$) is the TP, lower (respectively, upper) triangular bidiagonal matrix given by

Proof. Let $A = F_n \cdots F_1 D G_1 \cdots G_n$ the factorization (5) of A . Since $\det G_i = \det F_i = 1, i = 1, \dots, n$, we have $\det A = \det D = \prod_{i=1}^{n+1} p_{i,i}$. \square

3. Bidiagonal factorization of Quantum Hilbert matrices

Quantum calculus (see [13]) uses q -integers, q -binomial coefficients, and other q -analogues of classical calculus. Let us recall that the q -binomial coefficients $\begin{bmatrix} n \\ k \end{bmatrix}_q, k = 0, \dots, n$, are given by

$$\begin{bmatrix} n \\ k \end{bmatrix}_q := \frac{[n]_q!}{[k]_q! [n-k]_q!},$$

where, for any non-negative integer n , the q -factorial $[n]_q!$ is defined by

$$[0]_q! := 1, \quad [n]_q! := [n]_q [n-1]_q \cdots [1]_q, \quad n \in \mathbb{N},$$

and the q -integer $[n]_q$ is

$$[n]_q := \begin{cases} 1 + q + \cdots + q^{n-1} = \frac{1 - q^n}{1 - q}, & \text{if } q \neq 1 \\ n, & \text{if } q = 1. \end{cases} \tag{12}$$

Clearly, $[n]_q$ is a polynomial in q and $[n]_q > 0$, for any $q \in (0, 1], n \in \mathbb{N}$. Moreover, the q -binomial coefficients $\begin{bmatrix} n \\ k \end{bmatrix}_q, k = 0, \dots, n$, are also polynomials in q with integer polynomials, which are known as Gaussian polynomials.

It can be checked that the q -binomial coefficients satisfy the following useful identities

$$\text{a) } \frac{[\alpha]_q}{[n]_q} \begin{bmatrix} \alpha - 1 \\ n - 1 \end{bmatrix}_q = \begin{bmatrix} \alpha \\ n \end{bmatrix}_q, \tag{13}$$

$$\text{b) } \frac{[\alpha - n]_q}{[n]_q} \begin{bmatrix} \alpha - 1 \\ n - 1 \end{bmatrix}_q = \begin{bmatrix} \alpha - 1 \\ n \end{bmatrix}_q, \tag{14}$$

$$\text{c) } \frac{[\alpha]_q}{[\alpha - n + 1]_q} \begin{bmatrix} \alpha - 1 \\ n - 1 \end{bmatrix}_q = \begin{bmatrix} \alpha \\ n - 1 \end{bmatrix}_q. \tag{15}$$

In the sequel, we shall use the following result.

Lemma 3. Given $n, p, r \in \mathbb{N}$,

$$[n + p]_q [n + r]_q - [n]_q [n + p + r]_q = q^n [p]_q [r]_q. \tag{16}$$

Proof. Using definition (12) for the q -integers, identity (16) trivially holds for $q = 1$. Moreover, for $q \neq 1$, we can write

$$\begin{aligned}
 [n+p]_q [n+r]_q - [n]_q [n+p+r]_q &= \frac{(1-q^{n+p})(1-q^{n+r}) - (1-q^n)(1-q^{n+p+r})}{(1-q)^2} \\
 &= \frac{q^n(1-q^p - q^r + q^{p+r})}{(1-q)^2} = q^n \frac{1-q^p}{1-q} \frac{1-q^r}{1-q} = q^n [p]_q [r]_q. \quad \square
 \end{aligned}$$

For $\alpha \in \mathbb{N}$, we shall consider the following generalization of Quantum Hilbert matrices $H_n^{(\alpha,q)} := (H_{i,j}^{(\alpha,q)})_{1 \leq i,j \leq n+1}$ with

$$H_{i,j}^{(\alpha,q)} := \frac{[\alpha]_q}{[i+j+\alpha-2]_q}, \quad 1 \leq i, j \leq n+1. \tag{17}$$

Theorem 4. Let $H_n^{(\alpha,q)}$ be the quantum Hilbert matrix (17). The multipliers $m_{i,j}$ of the Neville elimination of $H_n^{(\alpha,q)}$ are given by

$$m_{i,j} = \tilde{m}_{i,j} = q^{j-1} \frac{[i+\alpha-2]_q^2}{[i+j+\alpha-2]_q [i+j+\alpha-3]_q}, \quad 1 \leq j < i \leq n+1. \tag{18}$$

Moreover, the diagonal pivots $p_{i,i}$ of the Neville elimination of $H_n^{(\alpha,q)}$ are

$$p_{i,i} = q^{(i-1)(i+\alpha-2)} \frac{[\alpha]_q}{[2i+\alpha-2]_q [{}_{i-1}^{2i+\alpha-3}]_q^2}, \quad 1 \leq i \leq n+1, \tag{19}$$

and can be recursively computed as follows

$$\begin{aligned}
 p_{1,1} &= 1, \quad p_{2,2} = q^\alpha \frac{[\alpha]_q}{[2+\alpha]_q [1+\alpha]_q^2}, \\
 p_{i+1,i+1} &= q^{2i+\alpha-2} \frac{[i]_q^2 [i+\alpha-1]_q^2}{[2i+\alpha]_q [2i+\alpha-1]_q^2 [2i+\alpha-2]_q} p_{i,i}, \quad i = 1, \dots, n.
 \end{aligned} \tag{20}$$

Proof. Let $H^{(k)} := (h_{i,j}^{(k)})_{1 \leq i,j \leq n+1}$, $k = 1, \dots, n+1$, be the matrices obtained after $k-1$ steps of the Neville elimination procedure for $H_n^{(\alpha,q)}$. Now, by induction on k , we shall see that, for $k = 2, \dots, n+1$,

$$h_{i,j}^{(k)} = q^{(k-1)(i+\alpha-2)} [\alpha]_q \frac{[{}_{k-1}^{j-1}]_q}{[k]_q [{}_k^{i+j+\alpha-2}]_q [{}_{k-1}^{i+k+\alpha-3}]_q}, \quad k \leq j, i \leq n+1. \tag{21}$$

It can be easily checked that $h_{i,1}^{(1)}/h_{i-1,1}^{(1)} = [i+\alpha-2]_q/[i+\alpha-1]_q$ and then, from (2), we can write

$$h_{i,j}^{(2)} := h_{i,j}^{(1)} - \frac{h_{i,1}^{(1)}}{h_{i-1,1}^{(1)}} h_{i-1,j}^{(1)} = [\alpha]_q \left(\frac{1}{[i+j+\alpha-2]_q} - \frac{[i+\alpha-2]_q}{[i+\alpha-1]_q} \frac{1}{[i+j+\alpha-3]_q} \right). \tag{22}$$

From (22), and taking into account (16), with $n := i + \alpha - 2$, $p := 1$ and $r := j - 1$, we have the following identities

$$\begin{aligned} h_{i,j}^{(2)} &= [\alpha]_q \frac{[i + \alpha - 1]_q [i + j + \alpha - 3]_q - [i + \alpha - 2]_q [i + j + \alpha - 2]_q}{[i + j + \alpha - 2]_q [i + j + \alpha - 3]_q [i + \alpha - 1]_q} \\ &= q^{i+\alpha-2} [\alpha]_q \frac{[1]_q [j - 1]_q}{[2]_q \begin{bmatrix} i+j+\alpha-2 \\ 2 \end{bmatrix}_q \begin{bmatrix} i+\alpha-1 \\ 1 \end{bmatrix}_q} = q^{i+\alpha-2} [\alpha]_q \frac{[j - 1]_q}{[2]_q \begin{bmatrix} i+j+\alpha-2 \\ 2 \end{bmatrix}_q \begin{bmatrix} i+\alpha-1 \\ 1 \end{bmatrix}_q}, \end{aligned}$$

confirming (21) for $k = 2$. If (21) holds for some $k \in \{2, \dots, n\}$,

$$\frac{h_{i,k}^{(k)}}{h_{i-1,k}^{(k)}} = q^{k-1} \frac{[i + \alpha - 2]_q^2}{[i + k + \alpha - 2]_q [i + k + \alpha - 3]_q}, \quad i = k + 1, \dots, n + 1. \tag{23}$$

From (2), (23), and the following identity obtained from (15)

$$\frac{[i + k + \alpha - 3]_q \begin{bmatrix} i + k + \alpha - 4 \\ k - 1 \end{bmatrix}_q}{[i + \alpha - 2]_q} = \begin{bmatrix} i + k + \alpha - 3 \\ k - 1 \end{bmatrix}_q, \tag{24}$$

we deduce that

$$h_{i,j}^{(k+1)} = q^{(k-1)(i+\alpha-2)} [\alpha]_q \frac{\begin{bmatrix} j-1 \\ k-1 \end{bmatrix}_q}{[k]_q \begin{bmatrix} i+k+\alpha-3 \\ k-1 \end{bmatrix}_q} C_{i,j}^{(k)} \tag{25}$$

with

$$C_{i,j}^{(k)} := \frac{1}{\begin{bmatrix} i+j+\alpha-2 \\ k \end{bmatrix}_q} - \frac{[i + \alpha - 2]_q}{[i + \alpha + k - 2]_q} \frac{1}{\begin{bmatrix} i+j+\alpha-3 \\ k \end{bmatrix}_q},$$

for $k + 1 \leq j, i \leq n + 1$. Using in (25) the following identities derived from (14) and (13), respectively,

$$\begin{aligned} \begin{bmatrix} i + j + \alpha - 2 \\ k \end{bmatrix}_q &= \frac{[k + 1]_q}{[i + j - k + \alpha - 2]_q} \begin{bmatrix} i + j + \alpha - 2 \\ k + 1 \end{bmatrix}_q, \\ \begin{bmatrix} i + j + \alpha - 3 \\ k \end{bmatrix}_q &= \frac{[k + 1]_q}{[i + j + \alpha - 2]_q} \begin{bmatrix} i + j + \alpha - 2 \\ k + 1 \end{bmatrix}_q, \end{aligned}$$

and

$$[i + k + \alpha - 2]_q [i + j - k + \alpha - 2]_q - [i + \alpha - 2]_q [i + j + \alpha - 2]_q = q^{i+\alpha-2} [k]_q [j - k]_q,$$

which is deduced from (16), with $n := i + \alpha - 2$, $p := k$ and $r := j - k$, we obtain

$$\begin{aligned}
 h_{i,j}^{(k+1)} &= \frac{q^{(k-1)(i+\alpha-2)} [\alpha]_q \begin{bmatrix} j-1 \\ k-1 \end{bmatrix}_q}{[k]_q [k+1]_q \begin{bmatrix} i+j+\alpha-2 \\ k+1 \end{bmatrix}_q \begin{bmatrix} i+k+\alpha-3 \\ k-1 \end{bmatrix}_q} \times \\
 &\times \left([i+j-k+\alpha-2]_q - \frac{[i+\alpha-2]_q [i+j+\alpha-2]_q}{[i+k+\alpha-2]_q} \right) \\
 &= q^{k(i+\alpha-2)} [\alpha]_q \frac{\begin{bmatrix} j-k \\ k-1 \end{bmatrix}_q}{[k+1]_q \begin{bmatrix} i+j+\alpha-2 \\ k+1 \end{bmatrix}_q [i+k+\alpha-2]_q \begin{bmatrix} i+k+\alpha-3 \\ k-1 \end{bmatrix}_q},
 \end{aligned}$$

for $k+1 \leq j, i \leq n+1$. Finally, taking into account that, from (14) and (13), we can write

$$\frac{\begin{bmatrix} j-k \\ k \end{bmatrix}_q \begin{bmatrix} j-1 \\ k-1 \end{bmatrix}_q}{[k]_q \begin{bmatrix} k-1 \\ k-1 \end{bmatrix}_q} = \begin{bmatrix} j-1 \\ k \end{bmatrix}_q, \quad \frac{[i+k+\alpha-2]_q \begin{bmatrix} i+k+\alpha-3 \\ k-1 \end{bmatrix}_q}{[k]_q \begin{bmatrix} k-1 \\ k-1 \end{bmatrix}_q} = \begin{bmatrix} i+k+\alpha-2 \\ k \end{bmatrix}_q,$$

we conclude that

$$h_{i,j}^{(k+1)} = q^{k(i+\alpha-2)} [\alpha]_q \frac{\begin{bmatrix} j-1 \\ k \end{bmatrix}_q}{[k+1]_q \begin{bmatrix} i+j+\alpha-2 \\ k+1 \end{bmatrix}_q \begin{bmatrix} i+k+\alpha-2 \\ k \end{bmatrix}_q}, \quad k+1 \leq j, i \leq n+1,$$

and (21) holds for $k+1$.

Now, by (3) and (21), the pivots of the Neville elimination of $H_n^{(\alpha,q)}$ satisfy

$$p_{i,j} = h_{i,j}^{(j)} = q^{(j-1)(i+\alpha-2)} [\alpha]_q \frac{1}{[j]_q \begin{bmatrix} i+j+\alpha-2 \\ j \end{bmatrix}_q \begin{bmatrix} i+j+\alpha-3 \\ j-1 \end{bmatrix}_q}, \quad 1 \leq j < i \leq n+1. \quad (26)$$

For the particular case $i = j$, we have

$$p_{i,i} := q^{(i-1)(i+\alpha-2)} \frac{[\alpha]_q}{[i]_q \begin{bmatrix} 2i+\alpha-2 \\ i \end{bmatrix}_q \begin{bmatrix} 2i+\alpha-3 \\ i-1 \end{bmatrix}_q} = q^{(i-1)(i+\alpha-2)} \frac{[\alpha]_q}{[2i+\alpha-2]_q \begin{bmatrix} 2i+\alpha-3 \\ i-1 \end{bmatrix}_q},$$

corresponding to identity (19). It can be easily checked that $p_{i,i} = 1$ and

$$\frac{p_{i+1,i+1}}{p_{i,i}} = q^{2i+\alpha-2} \frac{[i]_q^2 [i+\alpha-1]_q^2}{[2i+\alpha]_q [2i+\alpha-1]_q^2 [2i+\alpha-2]_q},$$

confirming formula (20). Let us observe that, since the pivots of the Neville elimination of $H_n^{(\alpha,q)}$ are nonzero, this elimination can be performed without row exchanges.

Finally, using (4) and (21), the multipliers $m_{i,j}$ can be described as

$$m_{i,j} = \frac{p_{i,j}}{p_{i-1,j}} = q^{j-1} \frac{[i+\alpha-2]_q^2}{[i+j+\alpha-2]_q [i+j+\alpha-3]_q}, \quad 1 \leq j < i \leq n+1. \quad (27)$$

Since $H_n^{(\alpha,q)}$ is symmetric, using Remark 1, we deduce that $\tilde{m}_{i,j} = m_{i,j}$. \square

Taking into account Theorem 4, the decomposition (5) of $H_n^{(\alpha,q)}$ and (9) of $(H_n^{(\alpha,q)})^{-1}$, can be stored by means of $BD(H_n^{(\alpha,q)}) = (BD(H_n^{(\alpha,q)}))_{i,j} \mathbb{1}_{1 \leq i,j \leq n+1}$ with

$$BD(H_n^{(\alpha,q)})_{i,j} := \begin{cases} q^{j-1} \frac{[i+\alpha-2]_q^2}{[i+j+\alpha-2]_q [i+j+\alpha-3]_q}, & \text{if } i > j, \\ q^{(i-1)(i+\alpha-2)} \frac{[\alpha]_q}{[2i+\alpha-2]_q [i-1]_q^2}, & \text{if } i = j, \\ q^{i-1} \frac{[j+\alpha-2]_q^2}{[i+j+\alpha-2]_q [i+j+\alpha-3]_q}, & \text{if } i < j. \end{cases} \tag{28}$$

Using the previous result, the total positivity of quantum Hilbert matrices can be analyzed and their determinant derived. It can be also deduced that computations with these matrices can be performed to high relative accuracy.

Proposition 5. For any $\alpha \in \mathbb{N}$ and $q \in (0, 1]$, the quantum Hilbert matrix $H_n^{(\alpha,q)}$ in (17) is strictly totally positive and

$$\det H_n^{(\alpha,q)} = q^{\frac{1}{6}n(n+1)(2n+3\alpha-2)} [\alpha]_q^n \prod_{i=1}^n \left([2i + \alpha]_q \begin{bmatrix} 2i + \alpha - 1 \\ k \end{bmatrix}_q \right)^{-1}. \tag{29}$$

Moreover, $H_n^{(\alpha,q)}$ and its inverse $(H_n^{(\alpha,q)})^{-1}$ can be computed to high relative accuracy.

Proof. Let us observe that for any $\alpha \in \mathbb{N}$ and $q \in (0, 1]$, the multipliers in (18) as well as the diagonal pivots in (19) are positive. So, taking into account Theorem 1, the strict total positivity of $H_n^{(\alpha,q)}$ can be deduced. Moreover, $H_n^{(\alpha,q)}$ and its inverse can be obtained to high relative accuracy since the computation of the mentioned pivots and multipliers do not require subtractions. Finally, taking into account Lemma 2 and formula (19) for the diagonal pivots, we have

$$\det H_n^{(\alpha,q)} = \prod_{i=1}^n q^{i(i+\alpha-1)} \prod_{i=1}^n \frac{[\alpha]_q}{[2i + \alpha]_q [i]_q^{2i+\alpha-1}}. \tag{30}$$

Furthermore, using induction, it can be easily checked that

$$\sum_{i=1}^n i(i + \alpha - 1) = \frac{1}{6}n(n + 1)(2n + 3\alpha - 2),$$

and so, from (30), identity (29) for $\det H_n^{(\alpha,q)}$ readily follows. \square

Let us observe that, for the particular choice $\alpha = 1$ and $q = 1$ formula (29) coincides with formula (1) for the determinant of the Hilbert matrix $H_n = (1/(i+j-1))_{1 \leq i,j \leq n+1}$.

4. Error analysis and perturbation theory

Now, we shall analyze the numerical errors that occur during the computation of the bidiagonal factorization (5) of quantum Hilbert matrices (17) due to imprecise computer arithmetic or perturbed input data. For this purpose, let us first introduce some standard notations in error analysis.

For a given floating-point arithmetic and a real value $x \in \mathbb{R}$, the computed element is usually denoted by either $\text{fl}(x)$ or by \hat{x} . In order to study the effect of rounding errors, we shall use the well-known models

$$\text{fl}(x \text{ op } y) = (x \text{ op } y)(1 + \delta)^{\pm 1}, \quad |\delta| \leq u, \tag{31}$$

where u denotes the unit roundoff and op any of the elementary operations $+$, $-$, \times , $/$ (see [11], p. 40 for more details).

Following [11], when performing an error analysis, one usually deals with quantities θ_k such that

$$|\theta_k| \leq \gamma_k, \quad \gamma_k := \frac{ku}{1 - ku}, \tag{32}$$

for a given $k \in \mathbb{N}$ with $ku < 1$. Taking into account, Lemmas 3.3 and 3.4 of [11], the following properties of the values (32) hold:

- a) $(1 + \theta_k)(1 + \theta_j) = 1 + \theta_{k+j}$,
- b) $\gamma_k + \gamma_j + \gamma_k \gamma_j \leq \gamma_{k+j}$,
- c) $\gamma_k + u \leq \gamma_{k+1}$,
- d) if $\rho_i = \pm 1$, $|\delta_i| \leq u$, $i = 1, \dots, k$, then

$$\prod_{i=1}^k (1 + \delta_i)^{\rho_i} = 1 + \theta_k.$$

For example, statement a) above means that for any given two values θ_k and θ_j , bounded by γ_k and γ_j , respectively, there exists a number θ_{k+j} , bounded by γ_{k+j} , such that the above identity holds. Further use of the previous symbols must be intended in this respect.

Let us observe that, according the previous properties, relative errors and perturbations can be accumulated by means of the following counter

$$\langle k \rangle := \prod_{i=1}^k (1 + \delta_i)^{\rho_i}, \quad \rho_i = \pm 1, \quad |\delta_i| \leq u, \tag{33}$$

with the following rules

$$\langle k \rangle \langle j \rangle = \langle k + j \rangle, \quad \langle k \rangle / \langle j \rangle = \langle k + j \rangle, \tag{34}$$

(see Chapter 3 of [11]).

Let us note that the q -integers can be seen as polynomials in the variable q that take positive values for $q \in (0, 1]$. In fact, for $q \neq 1$, the evaluation of

$$[n]_q = 1 + q + \dots + q^{n-1}, \tag{35}$$

can be performed using Horner’s rule and nested multiplications (see Section 5.1 of [11]) with the following simple recurrence:

$$\begin{aligned} N[0] &= 1 \\ \text{for } i &= 1 : n - 1 \\ N[i] &= 1 + qN[i - 1] \\ \text{end} \end{aligned} \tag{36}$$

obtaining $N[n - 1] = [n]_q$.

Taking into account the rounding error analysis of Horner’s method in Section 5.1 of [11], the evaluation of a polynomial $p(x) = a_0 + a_1x + \dots + a_nx^n$ using Horner’s method has a small backward error in the sense that the computed value is the exact value at x of a polynomial obtained by making relative perturbations of size at most γ_{2n} to the coefficients of the polynomial p .

The following result adapts the mentioned error analysis to the computation of quantum integers, taking into account that the coefficients of $[n]_q$ in (35) are $a_i = 1$, $i = 0, \dots, n - 1$.

Lemma 6. *Given $q \in (0, 1)$, let $[n]_q$ be the q -integer (12) and $\text{fl}([n]_q)$ the value computed by recurrence (36). Then,*

$$\left| \frac{[n]_q - \text{fl}([n]_q)}{[n]_q} \right| \leq \gamma_{n-2}, \tag{37}$$

for the quantity γ_{n-2} defined in (32).

Proof. Using the properties (34) of the relative error counter and taking into account that $\text{fl}(1 + x) = 1 + x$ for all $x \in \mathbb{R}$, we can write

$$N[1] = 1 + q, \quad N[2] = 1 + q(1 + q) < 1 \rangle = 1 + q < 1 \rangle + q^2 < 1 \rangle,$$

and, by induction, it is very easy to see that

$$N[n - 1] = \sum_{k=0}^{n-2} q^k \langle k \rangle + q^{n-1} \langle n - 2 \rangle = \sum_{k=0}^{n-2} q^k (1 + \theta_k) + q^{n-1} (1 + \theta_{n-2}),$$

for values θ_k satisfying $|\theta_k| \leq \gamma_k, k = 1, \dots, n - 2$. Consequently,

$$|[n]_q - N[n - 1]| \leq \gamma_{n-2} \sum_{k=0}^{n-1} |q|^k = \gamma_{n-2}[n]_q,$$

and (37) holds. \square

Let us note that bound (37) improves the relative error bound γ_{2n-2} for a general polynomial of degree not greater than $n - 1$ evaluated with Horner’s method and illustrates that, using recursion (36), the computation in floating-point arithmetic of q -integers can be performed accurately.

Using Lemma 6, the following result analyzes the numerical error in the computation of the bidiagonal factorization (5) of a quantum Hilbert matrix $H_n^{(\alpha,q)}$.

Theorem 7. *For $\alpha \in \mathbb{N}$ and $q \in (0, 1)$, let $H_n^{(\alpha,q)}$ be the quantum Hilbert matrix (17). Let $BD(H_n^{(\alpha,q)}) = (b_{i,j})_{1 \leq i,j \leq n+1}$ be the matrix form of the bidiagonal decomposition (5) of $H_n^{(\alpha,q)}$ and $fl(BD(H_n^{(\alpha,q)})) = (fl(b_{i,j}))_{1 \leq i,j \leq n+1}$ be the matrix computed using the expression (18) and (20) for the multipliers and pivots, respectively and the recursion (36) for the computation of q -integers. Then*

$$\left| \frac{b_{i,j} - fl(b_{i,j})}{b_{i,j}} \right| \leq \gamma_{7n^2+O(n)}, \quad 1 \leq i, j \leq n + 1. \tag{38}$$

Proof. For $i > j$, the entry $b_{i,j}$ coincides with the multiplier $m_{i,j}$ that can be computed using (18). Taking into account the relative error bound for the computation of q -integers (see (37)), and accumulating relative errors in the style of Higham (see Chapter 3 of [11]), we can write

$$\left| \frac{b_{i,j} - fl(b_{i,j})}{b_{i,j}} \right| \leq \gamma_{4i+3j+4\alpha-15} \leq \gamma_{7n+4\alpha-11}, \quad 1 \leq j < i \leq n + 1. \tag{39}$$

For $i < j, b_{i,j} = m_{j,i}$ and then

$$\left| \frac{b_{i,j} - fl(b_{i,j})}{b_{i,j}} \right| \leq \gamma_{4j+3i+4\alpha-15} \leq \gamma_{7n+4\alpha-11}, \quad 1 \leq i < j \leq n + 1. \tag{40}$$

For $i = j$, by (20), $b_{1,1} = 1, b_{2,2} = q^\alpha \frac{[\alpha]_q}{[\alpha+2]_q [\alpha+1]_q^2}$ and $b_{i+1,i+1} = C_i p_{i,i}$ with

$$C_i := q^{2i+\alpha-2} \frac{[i]_q^2 [i + \alpha - 1]_q^2}{[2i + \alpha]_q [2i + \alpha - 1]_q^2 [2i + \alpha - 2]_q}, \quad i = 1, \dots, n. \tag{41}$$

From (37), the relative error in the computation of the factor C_i can be bounded as follows,

$$\left| \frac{C_i - \text{fl}(C_i)}{C_i} \right| \leq \gamma_{14i+7\alpha-17}, \quad i = 1, \dots, n, \tag{42}$$

and then,

$$\left| \frac{p_{i,i} - \text{fl}(p_{i,i})}{p_{i,i}} \right| \leq \gamma_{n_i}, \quad i = 1, \dots, n + 1, \tag{43}$$

with

$$n_1 = 0, \quad n_2 = 5\alpha - 1, \quad n_{i+1} = n_i + 14i + 7\alpha - 16, \quad i = 1, \dots, n.$$

It can be easily checked that

$$n_i = (i - 2)(7i + 7\alpha - 9) + 5\alpha - 1 \leq (n + 1)(7n + 7\alpha - 2) + 5\alpha - 1,$$

for $i = 2 \dots, n + 1$. \square

Now, we are going to analyze the sensitivity of the matrix representation of the bidiagonal factorization (10) of quantum Hilbert matrices with respect to perturbations in the value $q \in (0, 1)$. For this purpose, let us first study the sensitivity of q -integers.

Lemma 8. *Given $q \in (0, 1)$ and $q' = q(1 + \delta)$ such that $|\delta| < u$, let $[n]_q$ and $[n]_{q'}$ be the q -integers computed by recurrence (36). Then,*

$$\left| \frac{[n]_q - [n]_{q'}}{[n]_q} \right| \leq \gamma_{n-1}, \tag{44}$$

for the quantity γ_{n-1} defined in (32).

Proof. Let us observe that the perturbed q can be denoted by $q < 1 >$ and, using properties (34) of the relative error counter, we can write

$$\begin{aligned} N[1] &= 1 + q < 1 >, \\ N[2] &= 1 + q < 1 > (1 + q < 1 >) = 1 + q < 1 > + q^2 < 2 >. \end{aligned}$$

Furthermore, by induction, it is very easy to see that

$$[n]_{q'} = N[n - 1] = 1 + \sum_{k=1}^{n-1} q^k < k > = 1 + \sum_{k=1}^{n-1} q^k (1 + \theta_k),$$

for values θ_k satisfying $|\theta_k| \leq \gamma_k, k = 1, \dots, n - 1$. Then we can write

$$|[n]_q - [n]_{q'}| \leq \gamma_{n-1} \sum_{k=0}^{n-1} |q|^k = \gamma_{n-1} [n]_q,$$

and (44) holds. \square

Using Lemma (8), we can study the sensitivity of the matrix representation of the bidiagonal factorization of quantum Hilbert matrices.

Theorem 9. For $\alpha \in \mathbb{N}$ and $q \in (0, 1)$, $q' = q(1 + \delta)$ such that $|\delta| < u$, let $H_n^{(\alpha, q)}$ and $H_n^{(\alpha, q')}$ be the quantum Hilbert matrices (17). The matrices $BD(H_n^{(\alpha, q)}) = (b_{i,j})_{1 \leq i, j \leq n+1}$ and $(BD(H_n^{(\alpha, q')})) = (b'_{i,j})_{1 \leq i, j \leq n+1}$ computed using the expression (18) and (20) for the multipliers and pivots, respectively and the recursion (36) for the computation of q -integers satisfy

$$\left| \frac{b_{i,j} - b'_{i,j}}{b_{i,j}} \right| \leq \gamma_{7n^2+O(n)}, \quad 1 \leq i, j \leq n + 1. \tag{45}$$

Proof. For $i > j$, the entries $b_{i,j}$ and $b'_{i,j}$ coincide with the multipliers in (18). Taking into account Lemma 8 and accumulating relative errors, we derive

$$\left| \frac{b_{i,j} - b'_{i,j}}{b_{i,j}} \right| \leq \gamma_{4i+3j+4\alpha-14} \leq \gamma_{7n+4\alpha-10}, \quad 1 \leq j < i \leq n + 1. \tag{46}$$

From the symmetry and (46), for $i < j$, we derive

$$\left| \frac{b_{i,j} - b'_{i,j}}{b_{i,j}} \right| \leq \gamma_{4j+3i+4\alpha-14} \leq \gamma_{7n+4\alpha-10}, \quad 1 \leq i < j \leq n + 1. \tag{47}$$

Now, we study the sensitivity with respect to perturbations on the value q of the factor C_i such that $p_{i+1,i+1} = C_i p_{i,i}$ (see (20)). Using Lemma 8, the relative error can be bounded as follows,

$$\left| \frac{C_i - \text{fl}(C_i)}{C_i} \right| \leq \gamma_{14i+7\alpha-16}, \tag{48}$$

and then

$$\left| \frac{b_{i,i} - b'_{i,i}}{b_{i,i}} \right| \leq \gamma_{n_i}, \quad i = 1, \dots, n + 1, \tag{49}$$

with

$$n_1 = 0, \quad n_2 = 4\alpha + 1, \quad n_{i+1} = n_i + 14i + 7\alpha - 16, \quad i = 1, \dots, n.$$

It can be easily checked that

$$\begin{aligned} n_i &= (i - 2)(7i + 7\alpha - 9) + 4\alpha + 1 \\ &\leq (n + 1)(7(n + 1) + 7\alpha - 9) + 4\alpha + 1 = (n + 1)(7n + 7\alpha - 2) + 4\alpha + 1, \end{aligned} \tag{50}$$

for $i = 2, \dots, n + 1$. From (46), (47) and (50), the result holds. \square

Finally, let us note that quantity $7n^2$ can be seen as an appropriate condition number adapted to this problem with the quantum Hilbert matrix $H_n^{(\alpha, q)}$.

5. Numerical experiments

Some numerical tests are presented in this section supporting the obtained theoretical results. We have implemented different Matlab functions in $O(n^2)$ time for computing in the matrix form (10) the bidiagonal factorizations (5) for quantum Hilbert matrices $H_n^{(\alpha, q)}$ (see (28)).

We have considered different strictly totally positive quantum Hilbert matrices $H_n^{(\alpha, q)}$, for $(\alpha, q) = (1, 4/5)$ and $(\alpha, q) = (4, 4/5)$, with dimension $n + 1 = 10, \dots, 30$. In the rest of the section, for the sake of brevity, all the considered quantum Hilbert matrices and their corresponding bidiagonal decomposition will be denoted as H and $BD(H)$, respectively.

We have performed several matrix computations using the routines available in [15] with the matrix form (10) of the bidiagonal factorization (5) as an input argument. The obtained approximations have been compared with the respective approximations obtained by traditional methods provided in Matlab *R2022b*. In this context, the values provided by Wolfram Mathematica 13.1 with 100-digit arithmetic have been taken as the exact solution of the considered algebraic problem.

The relative error of each approximation has also been computed in Mathematica with 100-digit arithmetic as $e := |y - \tilde{y}|/|y|$, where y denotes the exact solution and \tilde{y} the computed approximation.

As we shall see, the proposed procedure exploits the structure of totally positive matrices, achieving computations to high relative accuracy for quantum Hilbert matrices. Then, it is possible to carry out virtually calculations with these matrices almost as if no rounding errors occur in the computation process, meaning that the uncertainty in the output results is about the same as in the input data.

Computation of the inverse matrix. For all considered matrices, we have compared the inverse obtained using the proposed bidiagonal decompositions with the function `TNInverseExpand` and the inverse computed with the Matlab command `inv`. As shown in Fig. 1, our procedure provides very accurate results. On the contrary, the results obtained with Matlab reflect poor accuracy.

Resolution of linear systems. Further to this, for all considered matrices, we have compared the solution of the linear systems $Hc = d$ where $d = ((-1)^{i+1}d_i)_{1 \leq i \leq n+1}$ and d_i , $i = 1, \dots, n + 1$, are random nonnegative integer values, obtained using the obtained bidiagonal decompositions with the function `TNSolve` and the solutions computed with the Matlab command `\`. As opposed to the results obtained with the command `\`, the proposed procedure preserves the accuracy for all the dimensions which have been taken into account. Fig. 2 illustrates the relative errors.

Computation of singular values. We have also compared the tenth singular value of the considered quantum Hilbert matrices computed with `svd` and `TNSingularValues`

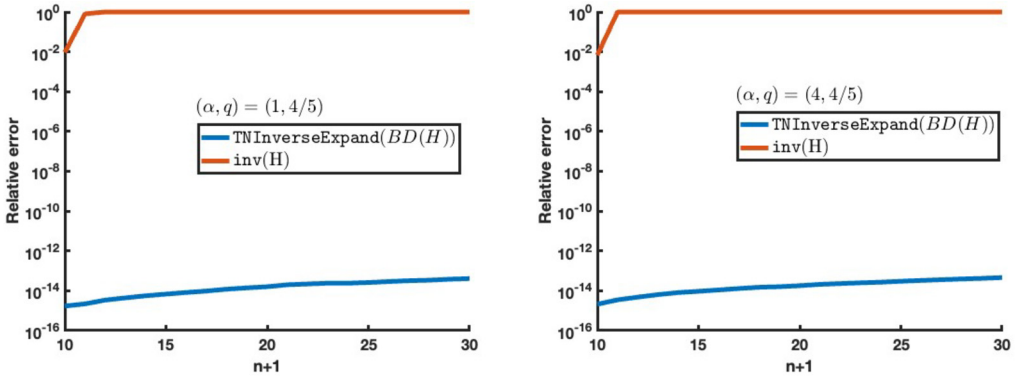


Fig. 1. Relative error of the approximations to the inverse of quantum Hilbert matrices.

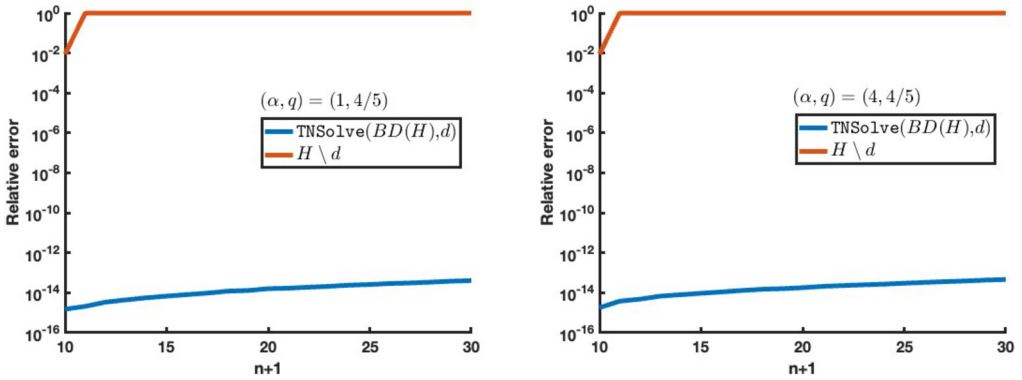


Fig. 2. Relative error of the approximations to the solution of $Hc = d$, with $d = ((-1)^{i+1}d_i)_{1 \leq i \leq n+1}$ and $d_i, i = 1, \dots, n + 1$, random nonnegative integer values.

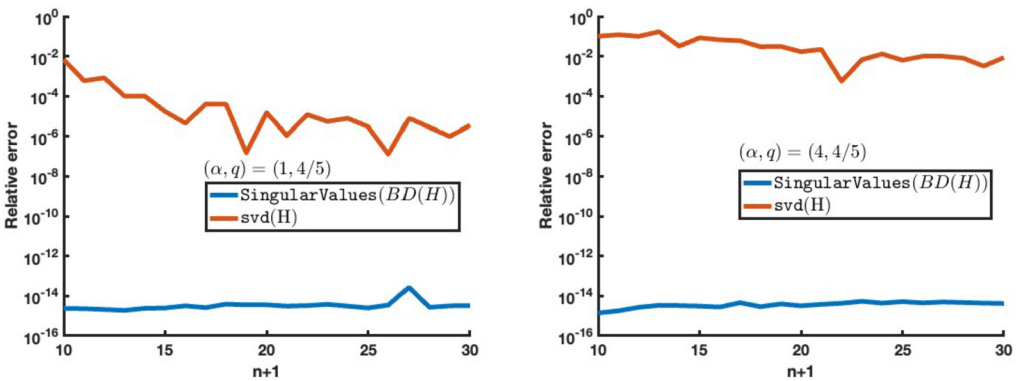


Fig. 3. Relative error of the approximations to the tenth singular value of quantum Hilbert matrices H .

with the proposed bidiagonal decomposition as input argument. The relative errors are shown in Fig. 3. Note that our approach computes accurately the tenth singular value. In contrast, the Matlab command `svd` return results that are not accurate at all.

Declaration of competing interest

None declared.

Data availability

Data will be made available on request.

Acknowledgement

We thank the comments and suggestions of an anonymous referee that have helped to improve this paper.

References

- [1] J.E. Andersen, C. Berg, Quantum Hilbert matrices and orthogonal polynomials, *J. Comput. Appl. Math.* 233 (2009) 723–729.
- [2] A. Barreras, J.M. Peña, Accurate computations of matrices with bidiagonal decomposition using methods for totally positive matrices, *Numer. Linear Algebra Appl.* 20 (2013) 413–424.
- [3] A.R. Collar, On the reciprocation of certain matrices, *Proc. R. Soc. Edinb.* 59 (1939) 195–206.
- [4] J. Demmel, Accurate singular value decompositions of structured matrices, *SIAM J. Matrix Anal. Appl.* 21 (1999) 562–580.
- [5] M. Gasca, J.M. Peña, Total positivity and Neville elimination, *Linear Algebra Appl.* 165 (1992) 25–44.
- [6] M. Gasca, J.M. Peña, A matricial description of Neville elimination with applications to total positivity, *Linear Algebra Appl.* 202 (1994) 33–53.
- [7] M. Gasca, J.M. Peña, On factorizations of totally positive matrices, in: M. Gasca, C.A. Micchelli (Eds.), *Total Positivity and Its Applications*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1996, pp. 109–130.
- [8] D. Hilbert, Ein Beitrag zur Theorie des Legendreschen Polynoms, *Acta Math.* 18 (1894).
- [9] I. Gohberg, I. Koltracht, On the inversion of Cauchy matrices, in: M.A. Kaashoek, J.H. van Schuppen, A.C.M. Ran (Eds.), *Proc. of Internat. Symposium MTNS-89*, vol. III, Birkhäuser, Basel, 1990, pp. 381–392.
- [10] I. Gohberg, I. Koltracht, Mixed, componentwise and structured condition numbers, *SIAM J. Matrix Anal. Appl.* 14 (1993) 688–704.
- [11] N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.
- [12] D.J. Higham, N.J. Higham, Backward error and condition of structured linear systems, *SIAM J. Matrix Anal. Appl.* 13 (1992) 162–175.
- [13] V. Kac, P. Cheung, *Quantum Calculus*, Springer, New York, 2002.
- [14] P. Koev, Accurate eigenvalues and SVDs of totally nonnegative matrices, *SIAM J. Matrix Anal. Appl.* 27 (2005) 1–23.
- [15] P. Koev, <http://math.mit.edu/~plamen/software/TNTTool.html>. (Accessed 16 December 2022).
- [16] E. Mainar, J.M. Peña, B. Rubio, Accurate computations with matrices related to bases $\{t^i e^{\lambda t}\}$, *Adv. Comput. Math.* 48 (38) (2022).
- [17] A. Marco, J.J. Martínez, Accurate computation of the Moore–Penrose inverse of strictly totally positive matrices, *J. Comput. Appl. Math.* 350 (2019) 299–308.
- [18] T.M. Richardson, The Filbert matrix, *Fibonacci Q.* 39 (3) (2001) 268–275.