Simultaneous tridiagonalization of two symmetric matrices

Seamus D. Garvey[1,*,†], Françoise Tisseur[2], Michael I. Friswell[3],
John E. T. Penny[4] and Uwe Prells[3]

[1] *School of Mechanical, Materials, Manufacturing Engineering and Management, University of Nottingham, University Park, Nottingham NG7 2RD, U.K.*
[2] *Department of Mathematics, University of Manchester, Manchester M13 9PL, U.K.*
[3] *University of Wales Swansea, Singleton Park, Swansea SA2 8PP, U.K.*
[4] *University of Aston, Birmingham B4 7ET, U.K.*

SUMMARY

We show how to simultaneously reduce a pair of symmetric matrices to tridiagonal form by congruence transformations. No assumptions are made on the non-singularity or definiteness of the two matrices. The reduction follows a strategy similar to the one used for the tridiagonalization of a single symmetric matrix via Householder reflectors. Two algorithms are proposed, one using non-orthogonal rank-one modifications of the identity matrix and the other, more costly but more stable, using a combination of Householder reflectors and non-orthogonal rank-one modifications of the identity matrix with minimal condition numbers. Each of these tridiagonalization processes requires $O(n^3)$ arithmetic operations and respects the symmetry of the problem. We illustrate and compare the two algorithms with some numerical experiments. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: symmetric matrices; generalized eigenvalue problem; tridiagonalization; symmetric quadratic eigenvalue problem

1. INTRODUCTION

This paper concerns pairs of symmetric matrices (K, M) and the computation of a non-singular transformation Q that simultaneously tridiagonalizes the pair (K, M), that is

$$Q^\mathrm{T}KQ = T, \quad Q^\mathrm{T}MQ = S \tag{1}$$

where both T and S are symmetric tridiagonal. No assumptions are made on the non-singularity or definiteness of M and K.

---

*Correspondence to: Seamus D. Garvey, School of Mechanical, Materials, Manufacturing Engineering and Management, University of Nottingham, University Park, Nottingham NG7 2RD, U.K.
†E-mail: seamus.garvey@nottingham.ac.uk

1.1. Motivation

Our main motivation for reducing a symmetric pair (K, M) to symmetric tridiagonal form (T, S) arises from the modelling of undamped multi-degree-of-freedom second-order systems

$$Kq + M\ddot{q} = Bf$$
$$r = B^{\mathrm{T}}q \tag{2}$$

and multi-degree-of-freedom first-order systems

$$Kq + D\dot{q} = Bf$$
$$r = B^{\mathrm{T}}q \tag{3}$$

where q is the vector of displacements and r and f are the vectors of terminal displacements and forces, r and f generally being of much smaller dimension than the displacement q. The matrix B is a selection matrix relating the full-length vectors to the corresponding terminal quantities. Three types of analysis are commonly performed for systems such as (2) and (3).

1. Computation of the transient response q as a function of time τ.
2. Computation of the steady-state frequency response q as a function of the frequency ω.
3. Computation of the natural frequencies ω_n.

For the reminder of this section, these analyses will be discussed as though they pertained only to the system of (2). Equivalent statements apply in all cases to systems such as (3).

As long as we can determine the acceleration $\ddot{q}$ at any instant τ, the transient responses may be computed by conducting a time-marching integration. Note that

$$\ddot{q}(\tau) = M^{-1}(Bf(\tau) - Kq(\tau))$$

so that on each occasion where $\ddot{q}$ is determined, it is necessary to perform one matrix–vector multiplication with K and solve a system of equations with M. If M^{-1} is computed initially and then stored, the computation of each new $\ddot{q}$ requires $O(n^2)$ operations, where n is the dimension of the problem. Assume that (K, M) has been reduced to tridiagonal form (T, S) via the transformation Q and let p and C be defined by $q = Qp$ and $C = Q^{\mathrm{T}}B$. Then (2) becomes

$$Tp + S\ddot{p} = Cf$$
$$r = C^{\mathrm{T}}p \tag{4}$$

With this equivalent representation of the system, each computation of $\ddot{p}$ requires a matrix–vector product and the solution of a linear system with a tridiagonal matrix. Both operations can be done in $O(n)$ operations.

The advantage afforded by the reduction to tridiagonal form (1) in the computation of steady-state frequency response is even more striking. To obtain the steady-state frequency response of the original system (2) directly as a function of ω we have to compute

$$r(\omega) = B^{\mathrm{T}}(K - \omega^2 M)^{-1}Bf(\omega)$$

This involves the solution of a system of coupled equations at each distinct frequency and the associated computational burden is $O(n^3)$ operations. This burden is reduced to $O(n^2)$ when the tridiagonalizing transformation is applied so that (T, S) appears in place of (K, M).

Another motivation for this work is that the first step in most natural frequency or eigenvalue computations is the reduction, in a finite number of operations, to a simple form such as the tridiagonal reduction (1). Then an iterative procedure can be applied to compute the eigensystem efficiently.

1.2. Review of existing reductions

To put our work into perspective, we review some of the existing methods for reducing, in a finite number of steps, a pair of symmetric matrices (K, M) to some simple form and give the conditions that must be satisfied by K and M for the reduction to be possible. We concentrate on methods that preserve symmetry. All the methods we are aware of reduce K to tridiagonal form and M to diagonal form.

Generally, when M is positive definite $(M > 0)$, a Cholesky factorization of $M = LL^T$ is used to transform (K, M) into (A, I) with $A = L^{-1}KL^{-T}$ symmetric. Then A is tridiagonalized into T via a sequence of Householder transformations [1]. It is well known that when M is close to being singular, the computation of A may suffer from instability.

When M is positive semidefinite $(M \geqslant 0)$, with $r > 0$ zero eigenvalues, we can again transform (K, M) to tridiagonal–diagonal form (A, D) with

$$D = \begin{bmatrix} I_{n-r} & 0 \\ 0 & 0_r \end{bmatrix}$$

but the reduction is not as straightforward as in the previous case. It can be done by a Cholesky factorization for semidefinite matrices followed by one step of Bunse–Gerstner's MDR reduction for symmetric matrices [2] (to split the problem into two subproblems, one of them corresponding to the r zero eigenvalues), and then a judicially chosen sequence of Givens rotations.

If M is indefinite, that is, M has both positive and negative eigenvalues, (K, M) can be reduced to tridiagonal–diagonal form using one of the procedures described by Brebner and Grad [3] or by Zurmühl and Falk [4]. However, these reductions require M to be non-singular.

1.3. Objectives

The symmetric tridiagonal–diagonal reduction is the most compact form we can obtain in a finite number of steps and the price to pay for this may be numerical instability. In this paper, we consider a less compact form that allows the second matrix to be in tridiagonal form.

One feature of our algorithm not shared by any other algorithm reducing a pair of symmetric matrices to some simple symmetric form is that it is not necessary that either of the two matrices should be positive definite or even positive semidefinite. Although it is common that in structural vibration and other undamped second-order dynamic systems both matrices are at least positive semidefinite, the extension to damped second-order systems

$$M\ddot{q} + C\dot{q} + Kq = f$$

where $C = C^T$ is the damping matrix (often positive semidefinite), leads to first-order systems

$$\mathcal{K}p + \mathcal{M}\dot{p} = g, \quad p = \begin{bmatrix} q \\ \dot{q} \end{bmatrix}, \quad g = \begin{bmatrix} 0 \\ f \end{bmatrix}$$

where $\mathcal{K}$ and $\mathcal{M}$ are symmetric matrices neither of which is positive definite [5, 6]. Hence, our tridiagonal reduction works in the most general case and can be used when nothing is known about the pair (K, M) other than its symmetry.

The paper is organized as follows. We describe in Section 2 the simultaneous tridiagonalization of a symmetric pair (K, M) and define the transformations required at each step of the reduction. We explain in Section 3 how to compute these transformations efficiently. We present Algorithm 3.1 which is based on $n - 2$ successive transformations, each one being a non-orthogonal rank-one modification of the identity matrix. We show in Section 4 that in order to minimize the growth of inherent errors in the pair (K, M) it is crucial to keep the condition number of the transformations used during the tridiagonalization process small. We describe a new type of elementary transformation based on the product of a Householder reflector and a non-orthogonal rank-one modification of the identity matrix with minimal condition number. We use these new transformations to derive Algorithm 4.2. Both algorithms depend on a parameter γ. We give in Section 5 some heuristics for choosing this parameter. Section 6 is devoted to numerical experiments.

1.4. Notation

Generally we use capital letters for matrices, lower case letters for vectors and lower Greek letters for scalars. I is the identity matrix whose dimension is determined by the context, and the vector e_k denotes its kth column. We often use the rectangular matrix

$$V = [e_2, e_3, \ldots, e_\ell] \in \mathbb{R}^{\ell \times (\ell - 1)} \tag{5}$$

Premultiplication of an $\ell \times \ell$ matrix by V^{T} discards the first row of the matrix. We denote by

$$\kappa(A) = \|A\| \|A^{-1}\|$$

the condition number of a square matrix A, $\|\cdot\|$ being the 2-norm (Euclidean norm). Finally the colon notation $i = 1 : n$ means the same as $i = 1, \ldots, n$.

2. REDUCTION TO TRIDIAGONAL FORMS

In this section, we describe our technique for reducing a pair of symmetric $n \times n$ matrices (K, M) to tridiagonal form (T, S). Congruence transformations are used to preserve symmetry and eigenvalues.

2.1. Basic idea

Assume that there exists a non-singular $n \times n$ matrix G_1 that introduces zeros in the first column below the first subdiagonal and in the first row after the first superdiagonal of both $K_1 \equiv K$ and $M_1 \equiv M$. This gives

$$K_2 = G_1^{\mathrm{T}} K_1 G_1 = \begin{bmatrix} \kappa_1 & \tau_1 e_1^{\mathrm{T}} \\ \tau_1 e_1 & \tilde{K}_2 \end{bmatrix}, \quad M_2 = G_1^{\mathrm{T}} M_1 G_1 = \begin{bmatrix} \mu_1 & \sigma_1 e_1^{\mathrm{T}} \\ \sigma_1 e_1 & \tilde{M}_2 \end{bmatrix}$$

where κ_1, τ_1, μ_1 and σ_1 are real scalars. This causes K_2 and M_2 to be tridiagonal in their first rows and columns.

At the second step, the same idea is applied to $\tilde{K}_2$ and $\tilde{M}_2$. We denote by $\tilde{G}_2$ the non-singular matrix introducing zeros in the first column and first row of both $\tilde{K}_2$ and $\tilde{M}_2$,

$$\tilde{G}_2^{\mathrm{T}}\tilde{K}_2\tilde{G}_2 = \begin{bmatrix} \kappa_2 & \tau_2 e_1^{\mathrm{T}} \\ \tau_2 e_1 & \tilde{K}_3 \end{bmatrix}, \quad \tilde{G}_2^{\mathrm{T}}\tilde{M}_2\tilde{G}_2 = \begin{bmatrix} \mu_2 & \sigma_2 e_1^{\mathrm{T}} \\ \sigma_2 e_1 & \tilde{M}_3 \end{bmatrix}$$

and let

$$G_2 = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{G}_2 \end{bmatrix}$$

Note that for the two matrices

$$K_3 = G_2^{\mathrm{T}} K_2 G_2, \quad M_3 = G_2^{\mathrm{T}} M_2 G_2$$

to be tridiagonal in their two first columns, it is crucial that G_2 does not destroy the zeros already introduced in the first columns and rows of K_2 and M_2. As a consequence, $\tilde{G}_2$ must satisfy

$$\tilde{G}_2^{\mathrm{T}} e_1 = \beta e_1, \quad 0 \neq \beta \in \mathbb{R} \tag{6}$$

Hence,

$$K_3 = \begin{bmatrix} \kappa_1 & \beta\tau_1 & 0^{\mathrm{T}} \\ \beta\tau_1 & \kappa_2 & \tau_2 e_1^{\mathrm{T}} \\ 0 & \tau_2 e_1 & \tilde{K}_3 \end{bmatrix}, \quad M_3 = \begin{bmatrix} \mu_1 & \beta\sigma_1 & 0^{\mathrm{T}} \\ \beta\sigma_1 & \mu_2 & \sigma_2 e_1^{\mathrm{T}} \\ 0 & \sigma_2 e_1 & \tilde{M}_3 \end{bmatrix}$$

Finally, after $n-2$ such steps,

$$K_{n-1} = G_{n-2}^{\mathrm{T}} K_{n-2} G_{n-2} \equiv T, \quad M_{n-1} = G_{n-2}^{\mathrm{T}} M_{n-2} G_{n-2} \equiv S$$

are in tridiagonal form. Let

$$Q = G_1 G_2 \ldots G_{n-2}$$

Then Q simultaneously tridiagonalizes K and M,

$$T = Q^{\mathrm{T}} K Q, \quad S = Q^{\mathrm{T}} M Q$$

In what follows we describe how to construct the matrices G_k, $k = 1:n-2$. As the technique is the same at each step k of the reduction, we drop the subscript k.

2.2. Constructing G

We examine how to construct a non-singular matrix $G \in \mathbb{R}^{\ell \times \ell}$ satisfying

$$G^{\mathrm{T}} e_1 = e_1, \quad G^{\mathrm{T}} K G = \begin{bmatrix} \kappa & \tau e_1^{\mathrm{T}} \\ \tau e_1 & \tilde{K} \end{bmatrix}, \quad G^{\mathrm{T}} M G = \begin{bmatrix} \mu & \sigma e_1^{\mathrm{T}} \\ \sigma e_1 & \tilde{M} \end{bmatrix} \tag{7}$$

where κ, μ, σ, τ are real scalars and, K, M are symmetric matrices. The first constraint in (7) comes from (6), where without loss of generality we set $\beta = 1$.

We recall that the role of G is to introduce zeros in the first column below the first superdiagonal and in the first row after the first superdiagonal of both K and M. Here the $\ell \times \ell$ matrices G, K and M can be seen as the matrices G_k, $\tilde{K}_k$ and $\tilde{M}_k$ in Section 2.1 with $\ell = n - k + 1$.

An $\ell \times \ell$ Householder reflector has the form

$$H = I - 2\frac{vv^{\mathrm{T}}}{v^{\mathrm{T}}v}, \quad 0 \neq v \in \mathbb{R}^{\ell}$$

Householder reflectors are a powerful tool for introducing zeros into vectors. For any $u \in \mathbb{R}^{\ell}$, if

$$v = u + \mathrm{sign}(e_1^{\mathrm{T}}u)\|u\|e_1$$

then

$$Hu = -\mathrm{sign}(e_1^{\mathrm{T}}u)\|u\|e_1 \tag{8}$$

Householder reflectors are symmetric, orthogonal and they enjoy good numerical properties (see Reference [7, Chapter 19]). We denote by

$$v = \mathrm{house}(u)$$

any function (program) computing v so that (8) is satisfied.

Two situations are considered for the construction of G.

(i) If $V^{\mathrm{T}}Ke_1 = \alpha V^{\mathrm{T}}Me_1$ for some $\alpha \in \mathbb{R}$ with V as in (5) then taking $L = I$ and constructing a Householder matrix H such that

$$H^{\mathrm{T}}V^{\mathrm{T}}Ke_1 = \tau e_1$$

yields $H^{\mathrm{T}}V^{\mathrm{T}}Me_1 = \alpha\tau e_1$. Then

$$G = \mathrm{diag}(1, H)$$

satisfies (7).

(ii) Assume that $V^{\mathrm{T}}Ke_1 \neq \alpha V^{\mathrm{T}}Me_1$ for all $\alpha \in \mathbb{R}$, that is $V^{\mathrm{T}}Ke_1$ and $V^{\mathrm{T}}Me_1$ are linearly independent. If K and M can be transformed by congruence transformation with a non-singular matrix $L \neq I$ in a such way that

$$V^{\mathrm{T}}L^{\mathrm{T}}KLe_1 = \alpha V^{\mathrm{T}}L^{\mathrm{T}}MLe_1$$

for some $\alpha \in \mathbb{R}$, then we are back in case (i) with K and M replaced by $L^{\mathrm{T}}KL$ and $L^{\mathrm{T}}ML$, respectively. The role of L is to transform the first columns and rows of K and M so that a Householder reflector can subsequently create the requisite zeros simultaneously in both matrices. Hence, if $L^{\mathrm{T}}e_1 = e_1$ then

$$G = L\begin{bmatrix} 1 & 0 \\ 0 & H \end{bmatrix} \tag{9}$$

satisfies (7). Note that if there exists an L such that the two vectors $V^{\mathrm{T}}L^{\mathrm{T}}KLe_1$ and $V^{\mathrm{T}}L^{\mathrm{T}}MLe_1$ have zeros in all their components except the first then we can set $H = I$.

In the next section we show how to construct the matrix L of case (ii).

2.3. Constructing L

We describe a class of non-singular elementary matrices $L \neq I \in \mathbb{R}^{\ell \times \ell}$ of the form

$$L = I + xy^{\mathrm{T}}, \quad x, y \in \mathbb{R}^{\ell}, \quad x^{\mathrm{T}}y \neq -1 \tag{10}$$

such that

$$e_1^{\mathrm{T}}L = e_1^{\mathrm{T}}, \quad V^{\mathrm{T}}(L^{\mathrm{T}}KL)e_1 = \tilde{\kappa}w, \quad V^{\mathrm{T}}(L^{\mathrm{T}}ML)e_1 = \tilde{\mu}w \tag{11}$$

with $w \in \mathbb{R}^{\ell-1}$, $\tilde{\kappa}, \tilde{\mu}$ non-zero scalars, and with the assumption that the two vectors $V^{\mathrm{T}}Ke_1$ and $V^{\mathrm{T}}Me_1$ are linearly independent. If $V^{\mathrm{T}}Ke_1$ and $V^{\mathrm{T}}Me_1$ are linearly dependent then we can take $L = I$. The condition $x^{\mathrm{T}}y \neq -1$ ensures that L is non-singular.

The matrices L are rank-one modifications of the identity matrix. They have interesting mathematical and numerical properties. In particular their inverse is explicitly given by

$$L^{-1} = I - \frac{xy^{\mathrm{T}}}{1 + x^{\mathrm{T}}y}$$

and their condition number is readily available in terms of x and y (see Section 4). Also the transformation $L^{\mathrm{T}}AL$ with A symmetric can be done in $4\ell^2$ operations. This can be expressed in MATLAB-style pseudocode as follows.

```
function A = elm_apply(A, x, y)
% Apply elementary transformation L = I + xy^T to A.
z = Ax
u = z + (x^T z)y/2
A = A + uy^T + yu^T
```

In the rest of this section we concentrate on deriving the vectors x and y defining L. The first constraint in (11) implies that

$$e_1^{\mathrm{T}}x = 0 \tag{12}$$

Let us assume that $e_1^{\mathrm{T}}y = 0$. Then

$$L = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{L} \end{bmatrix}, \quad \tilde{L} = I - \tilde{x}\tilde{y}^{\mathrm{T}}, \quad \tilde{x}\tilde{y} \in \mathbb{R}^{\ell-1}$$

and the two last constraints in (11) become

$$V^{\mathrm{T}}[(L^{\mathrm{T}}KL)e_1 \ (L^{\mathrm{T}}KL)e_1] = \tilde{L}[V^{\mathrm{T}}Ke_1 \ V^{\mathrm{T}}Me_1] = [\tilde{\kappa}w \ \tilde{\mu}w]$$

As $V^{\mathrm{T}}Ke_1$ and $V^{\mathrm{T}}Me_1$ are linearly independent, $\tilde{L}$ transforms a rank-2 matrix into a rank-1 matrix, implying that $\tilde{L}$ and therefore also L is singular. So $e_1^{\mathrm{T}}y \neq 0$ and without loss of generality we take

$$e_1^{\mathrm{T}}y = 1 \tag{13}$$

The next two constraints in (11) can be rewritten as

$$V^{\mathrm{T}}K(x + e_1) + [(x + e_1)^{\mathrm{T}}Kx]V^{\mathrm{T}}y = \tilde{\kappa}w \tag{14}$$

$$V^T M(x + e_1) + [(x + e_1)^T Mx]V^T y = \tilde{\mu} w \tag{15}$$

Let γ be a fixed non-zero constant such that $K - \gamma M$ is non-singular. We now look for a solution x that satisfies the simplifying condition

$$(x + e_1)^T Kx = \gamma (x + e_1)^T Mx \tag{16}$$

We also impose that the scalars $\tilde{\kappa}$, $\tilde{\mu}$ in (11) satisfy

$$\tilde{\kappa} = \gamma \tilde{\mu} \tag{17}$$

Then, (14) $- \gamma \times$ (15) yields the underdetermined system

$$V^T (K - \gamma M)(x + e_1) = 0 \tag{18}$$

whose solutions are given by

$$(K - \gamma M)(x + e_1) = \delta_1 e_1, \quad \delta_1 \in \mathbb{R} \text{ arbitrary}$$

Since $K - \gamma M$ is non-singular, we have

$$x + e_1 = \delta_1 (K - \gamma M)^{-1} e_1$$

The condition $e_1^T x = 0$ in (12) implies

$$\delta_1 = (e_1^T (K - \gamma M)^{-1} e_1)^{-1}$$

Hence,

$$x = (e_1^T (K - \gamma M)^{-1} e_1)^{-1} (K - \gamma M)^{-1} e_1 - e_1 \tag{19}$$

It is easy to verify that with this expression for x, the simplifying condition in (16) is satisfied.

We now determine the vector y. Forming $\gamma \times$ (14) $+$ (15) gives

$$V^T (\gamma K + M)(x + e_1) + (\gamma (x + e_1)^T Kx + (x + e_1)^T Mx)V^T y = (\gamma \tilde{\kappa} + \tilde{\mu}) w$$

Note that because of (17), $\tilde{\kappa}$, $\tilde{\mu}$ are determined up to a constant. We now fix them completely with the normalization condition

$$\gamma \tilde{\kappa} + \tilde{\mu} = (x + e_1)^T (\gamma K + M)x \tag{20}$$

Hence

$$V^T y = -[(x + e_1)^T (\gamma K + M)x]^{-1} V^T (\gamma K + M)(x + e_1) + w \tag{21}$$

All the solutions to this underdetermined system are given by

$$y = -[(x + e_1)^T (\gamma K + M)x]^{-1} V V^T (\gamma K + M)(x + e_1) + Vw + \delta_2 e_1 \tag{22}$$

for some arbitrary $\delta_2 \in \mathbb{R}$. The condition $e_1^T y = 1$ in (13) yields $\delta_2 = 1$.

To summarize, given an arbitrary vector $w \in \mathbb{R}^{\ell-1}$ and a non-zero scalar γ such that $K - \gamma M$ is non-singular, the matrix $L = I + xy^T$ with

$$x = (e_1^T (K - \gamma M)^{-1} e_1)^{-1} (K - \gamma M)^{-1} e_1 - e_1$$

$$y = -[(x + e_1)^T (\gamma K + M)x]^{-1} V V^T (\gamma K + M)(x + e_1) + Vw + e_1 \tag{23}$$

satisfies

$$e_1^T L = e_1^T, \quad V^T(L^T K L)e_1 = \tilde{\kappa}w, \quad V^T(L^T M L)e_1 = \tilde{\mu}w$$

Note that L is non-singular for w such that $x^T y \neq -1$.

Assume that w and $z = (K - \gamma M)^{-1} e_1$ are given. The computation of x and y can be written in pseudocode as follows:

```
function [x, y] = elm_w_xy(K, M, γ, w, z)
% For w and z = (K − γM)⁻¹e₁ given, construct the vectors x, y ∈ ℝˡ
% so that L = I + xyᵀ satisfies (11).
k = K(2:ℓ, 1), m = M(2:ℓ, 1)
If |kᵀm| = ‖k‖₂‖m‖₂
      x = 0, y = 0
else
      x = z/z(1)
      u = (γK + M)x, x(1) = 0
      y = − u/(uᵀx)
      y(2:n) = y(2:n) + w, y(1) = 1
end
```

The condition checks to see whether k and m are linearly dependent. This algorithm requires about $2\ell^2$ operations.

3. COMPUTING THE TRANSFORMATIONS G_k EFFICIENTLY

We now return to the notation used in Section 2.1. Recall that at step k of the reduction to tridiagonal forms, we have to compute the vectors x_k, y_k and v_k defining the $n \times n$ transformation

$$G_k = \begin{bmatrix} I & 0 \\ 0 & \tilde{G}_k \end{bmatrix}, \quad \tilde{G}_k = L_k \begin{bmatrix} 1 & 0 \\ 0 & H_k \end{bmatrix} \in \mathbb{R}^{(n-k+1) \times (n-k+1)}$$

The main numerical difficulty is in the computation of

$$x_k = (e_1^T(\tilde{K}_k - \gamma\tilde{M}_k)^{-1} e_1)^{-1}(\tilde{K}_k - \gamma\tilde{M}_k)^{-1} e_1 - e_1$$

Forming all the $x_k \in \mathbb{R}^{n-k+1}$, $k = 1:n-2$ requires the solution of $n-2$ symmetric, possibly indefinite, systems of $n-k$ equations

$$(\tilde{K}_k - \gamma\tilde{M}_k)z_k = e_1, \quad k = 1:n-2 \tag{24}$$

and the whole tridiagonalization procedure requires $O(n^4)$ operations if systems (24) are solved as arbitrary systems. We describe in this section a way of solving the $n-2$ systems (24) in $O(n^3)$ operations.

At step k of the tridiagonalization procedure we have

$$K_k - \gamma M_k = \begin{bmatrix} \kappa_1 - \gamma\mu_1 & \tau_1 - \gamma\sigma_1 & & & & 0 \\ \tau_1 - \gamma\sigma_1 & \kappa_2 - \gamma\mu_2 & \tau_2 - \gamma\sigma_2 & & & \\ & \tau_2 - \gamma\sigma_2 & \ddots & & \ddots & \\ & & \ddots & \kappa_{k-1} - \gamma\mu_{k-1} & (\tau_{k-1} - \gamma\sigma_{k-1})e_1^{\mathrm{T}} \\ 0 & & & (\tau_{k-1} - \gamma\sigma_{k-1})e_1 & \tilde{K}_k - \gamma\tilde{M}_k \end{bmatrix}$$

By construction

$$\tau_j = -\tilde{\kappa}_j \, \mathrm{sign}(w_j(1))\|w_j\|, \quad \sigma_j = -\tilde{\mu}_j \, \mathrm{sign}(w_j(1))\|w_j\|, \quad j = 1:k-1$$

with $\tilde{\kappa}_j$, $\tilde{\mu}_j$ satisfying (17), that is, $\tilde{\kappa}_j = \gamma\tilde{\mu}_j$. This gives

$$\tau_j - \gamma\sigma_j = 0, \quad j = 1:k-1$$

Hence, at the kth step the matrix $K_k - \gamma M_k$ is diagonal in its first k rows and columns

$$K_k - \gamma M_k = \begin{bmatrix} D_{k-1} & 0 \\ 0 & \tilde{K}_k - \gamma\tilde{M}_k \end{bmatrix}$$

with $D_{k-1} = \mathrm{diag}(\kappa_1 - \gamma\mu_1, \ldots, \kappa_{k-1} - \gamma\mu_{k-1}) \in \mathbb{R}^{(k-1)\times(k-1)}$.

Suppose that we have inverted the symmetric, possibly indefinite, matrix $K - \gamma M$. This can be done via a block LDL^{T} factorization

$$\Pi(K - \gamma M)\Pi^{\mathrm{T}} = LDL^{\mathrm{T}}$$

where L is unit lower triangular D is block diagonal with 1×1 or 2×2 diagonal blocks and Π is a permutation matrix. If a partial diagonal pivoting strategy is used for the determination of Π then the block LDL^{T} factorization is numerically stable as long as the growth factor is small [7, Chapter 11]. The whole inversion requires $O(n^3)$ operations.

At each new step of the tridiagonalization procedure, the inverse of $K_k - \gamma M_k$ is updated via

$$(K_k - \gamma M_k)^{-1} = G_{k-1}^{-1}(K_{k-1} - \gamma M_{k-1})^{-1}G_{k-1}^{-\mathrm{T}}$$

$$= \begin{bmatrix} D_{k-1}^{-1} & 0 \\ 0 & \tilde{G}_{k-1}^{-1}(K_{k-1} - \gamma M_{k-1})^{-1}\tilde{G}_{k-1}^{-\mathrm{T}} \end{bmatrix}$$

with

$$\tilde{G}_{k-1}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & H_{k-1} \end{bmatrix}\left(I - \frac{x_k y_k^{\mathrm{T}}}{1 + x_k^{\mathrm{T}} y_k}\right)$$

As explained in Section 2.3, the update at step k can be done in $O((n-k+1)^2)$ operations. Therefore, the computation of all the x_k requires $O(n^3)$ operations.

Note that

$$y_k = -[(x_k + e_1)^{\mathrm{T}}(\gamma \tilde{K}_k + \tilde{M}_k)x_k]^{-1}V V^{\mathrm{T}}(\gamma \tilde{K}_k + \tilde{M}_k)(x_k + e_1) + Vw_k + e_1$$

depends on the free vector w_k. If $w_k = e_1$ then the Householder transformation can be omitted. The complete algorithm with this particular choice for w_k is summarized below. We assume that a function computing a block LDL^{T} factorization is available. For example, we can use the MATLAB function `ldlt_symm` from Higham's Matrix Computation Toolbox [8].

Algorithm 3.1 (*Simultaneous tridiagonalization*)
Given two $n \times n$ symmetric matrices K, M and a non-zero scalar γ such that $K - \gamma M$ is non-singular, the following algorithm overwrites K and M with the tridiagonal matrices $Q^{\mathrm{T}}KQ$ and $Q^{\mathrm{T}}MQ$, where the non-singular matrix Q is the product of elementary transformations of the form $I + xy^{\mathrm{T}}$.

$\quad Q = I$
$\quad [L, D, \Pi] = \text{ldlt\_symm}(K - \gamma M)$
$\quad N = \Pi^{\mathrm{T}}L^{-\mathrm{T}}D^{-1}L^{-1}\Pi$
$\quad \text{for } k = 1 : n - 2$
$\qquad [x, y] = \text{elm\_w\_xy}(K(k{:}n, k{:}n), M(k{:}n, k{:}n), \gamma, e_1, N(k{:}n, k))$
$\qquad K(k{:}n, k{:}n) = \text{elm\_apply}(K(k{:}n, k{:}n), x, y)$
$\qquad M(k{:}n, k{:}n) = \text{elm\_apply}(M(k{:}n, k{:}n), x, y)$
$\qquad N(k{:}n, k{:}n) = \text{elm\_apply}(N(k{:}n, k{:}n), -y/(1 + x^{\mathrm{T}}y), x)$
$\qquad Q(:, k{:}n) = Q(:, k{:}n) + (Q(:, k{:}n)x)y^{\mathrm{T}}$
$\quad \text{end}$

This algorithm requires about $8n^3$ operations.

Note that the choice $w_k = e_1$ at each step of the reduction avoid the need to apply the Householder transformation with H_k, saving therefore some computation. However, in the next section we show that this choice for w_k is not generally the best for the numerical stability of the reduction.

4. MINIMIZING THE CONDITION NUMBER OF $L = I + xy^{\mathrm{T}}$

To minimize the growth of inherent errors in the pair (K, M) it is crucial to keep the condition number of the transformations used during the tridiagonalization process small. This can be explained as follows. Assume that a symmetric matrix A is affected by some error $E = E^{\mathrm{T}}$ and let Q be non-singular. We have

$$Q^{\mathrm{T}}(A + E)Q = \tilde{A} + F, \quad \tilde{A} = Q^{\mathrm{T}}AQ, \quad F = Q^{\mathrm{T}}EQ$$

To find a bound for $\|F\|$ we make use of a theorem of Ostrowski [9, p. 224]. Here the eigenvalues of a symmetric $n \times n$ matrix N are ordered $\lambda_n \leqslant \cdots \leqslant \lambda_1$.

Theorem 4.1
Let $A \in \mathbb{R}^{n \times n}$ be symmetric and $Q \in \mathbb{R}^{n \times n}$ non-singular. Then for each $k = 1 : n$,

$$\lambda_k(Q^{\mathrm{T}}EQ) = \theta_k \lambda_k(E)$$

where $\lambda_n(Q^{\mathrm{T}}Q) \leqslant \theta_k \leqslant \lambda_1(Q^{\mathrm{T}}Q)$.

From this result we have

$$\sigma_{\min}(Q)^2 \|E\| \leqslant \|F\| \leqslant \sigma_{\max}(Q)^2 \|E\| \tag{25}$$

where $\sigma_{\min}(Q)$ and $\sigma_{\max}(Q)$ denote the smallest and largest singular values of Q. Hence, pre- and post-multiplication of a matrix A by Q magnifies inherent errors in A by a factor at least $\sigma_{\min}(Q)^2$ and at most $\sigma_{\max}(Q)^2$. From (25) we obtain

$$\frac{\|F\|}{\|\tilde{A}\|} \leqslant \kappa(Q)^2 \frac{\|E\|}{\|A\|}, \quad \kappa(Q) = \frac{\sigma_{\max}(Q)}{\sigma_{\min}(Q)} \geqslant 1 \tag{26}$$

Hence, the error in A can be magnified by as much as $\kappa(Q)^2$ in passing to $\tilde{A}$.

Recall that in our simultaneous tridiagonalization

$$Q = G_1 G_2 \cdots G_{n-2}$$

where $G_k = \mathrm{diag}(I_k, \tilde{G}_k)$ with $\tilde{G}_k = L_k H_k$. Householder matrices have unity 2-norm condition number since they are orthogonal. Hence,

$$\kappa(Q) \leqslant \prod_{k=1}^{n-2} \kappa(L_k)$$

This suggests that in order to minimize the growth of errors we should keep $\kappa(L_k)$ small at each step of the tridiagonalization procedure. In the rest of this section, we show how to choose the vector w in (23) so that $\kappa(L)$ is minimized.

By definition,

$$\kappa(L) = \frac{\sigma_{\max}(L)}{\sigma_{\min}(L)} = \left(\frac{\lambda_{\max}(L^{\mathrm{T}}L)}{\lambda_{\min}(L^{\mathrm{T}}L)} \right)^{1/2}$$

Let y_k, $k = 1 : \ell - 2$ be $\ell - 2$ linearly independent vectors orthogonal to $[x, y]$. Then

$$L^{\mathrm{T}} L y_k = (I + yx^{\mathrm{T}})(I + xy^{\mathrm{T}}) y_k = y_k$$

which implies that $L^{\mathrm{T}}L$ has $\ell - 2$ eigenvalues equal to 1. We denote by λ_1, λ_2 the remaining two eigenvalues. We have

$$\lambda_1 \lambda_2 = \det(L^{\mathrm{T}}L) = \det(L)^2 = (1 + y^{\mathrm{T}}x)^2$$

and

$$(\ell - 2) + \lambda_1 + \lambda_2 = \mathrm{trace}(L^{\mathrm{T}}L)$$
$$= \mathrm{trace}(I + xy^{\mathrm{T}} + yx^{\mathrm{T}} + (x^{\mathrm{T}}x)yy^{\mathrm{T}})$$
$$= \ell + 2y^{\mathrm{T}}x + \|x\|^2 \|y\|^2$$

Let

$$p = 2(1 + y^{\mathrm{T}}x), \quad q = \|x\|^2 \|y\|^2 \tag{27}$$

Then λ_1 and λ_2 are the roots of

$$\lambda^2 - (p+q)\lambda + \frac{p^2}{4} = 0$$

and hence

$$\kappa(L) = \left(\frac{p + q + \sqrt{q(q+2p)}}{p + q - \sqrt{q(q+2p)}} \right)^{1/2} \tag{28}$$

Defining

$$r(y) \equiv \frac{q}{p} = \frac{\|x\|^2 \|y\|^2}{2(1 + y^{\mathrm{T}}x)}$$

we have

$$\min_{y \in \mathbb{R}^\ell} \kappa(L) = \min_{y \in \mathbb{R}^\ell} \frac{1 + r(y) + \sqrt{r(y)(r(y)+2)}}{1 + r(y) - \sqrt{r(y)(r(y)+2)}}$$

and the optimal y minimizes

$$\min_{y \in \mathbb{R}^\ell} \sqrt{r(y)(r(y)+2)}$$

and hence also

$$\min_{y \in \mathbb{R}^\ell} r(y)$$

We first show that $r(y)$ is minimized for $y \in \mathrm{span}(x, e_1)$, that is, $y = \alpha x + e_1$ for some $\alpha \in \mathbb{R}$ (recall that $e_1^{\mathrm{T}}x = 0$ and $e_1^{\mathrm{T}}y = 1$) and then determine the scalar α. Let $y_1 = y + z$ with $z \in \mathrm{span}(x, e_1)^\perp$. We have

$$p_1 \equiv 2(1 + y_1^{\mathrm{T}}x) = p, \quad q_1 \equiv \|x\|^2 \|y_1\|^2 = \|x\|^2 (1 + \|x\|^2) > p$$

so that $r(y) < r(y_1)$. Then it is easy to show that the function

$$g(\alpha) \equiv r(y) = \|x\|^2 (1 + \alpha^2 \|x\|^2) / (2(1 + \alpha \|x\|^2))$$

is minimized for

$$\alpha = -\frac{1 + \sqrt{1 + \|x\|^2}}{\|x\|^2}$$

yielding

$$y = e_1 - \frac{1 + \sqrt{1 + \|x\|^2}}{\|x\|^2} x$$

and

$$w = V^{\mathrm{T}}y + [(x + e_1)^{\mathrm{T}}(\gamma K + M)x]^{-1} V^{\mathrm{T}}(\gamma K + M)(x + e_1) \tag{29}$$

Note that with this expression for y, we have $x^{\mathrm{T}}y \neq -1$ so that L is non-singular.

The optimal condition number for L is given by

$$\kappa_{\mathrm{opt}}(L) = \frac{\sqrt{1 + \|x\|^2} + \|x\|}{\sqrt{1 + \|x\|^2} - \|x\|}$$

and is reasonably small as long as $\|x\| = O(1)$.

Given $z = (K - \gamma M)^{-1} e_1$, the computation of x and y requires 9ℓ operations and can be written in pseudocode as follows:

```
function [x, y] = elm_xy(K, M, z)
% Construct the vectors x, y ∈ ℝ^ℓ so that L = I + xyᵀ satisfies (11) with w given by (29)
k = K(2:ℓ, 1),  m = M(2:ℓ, 1)
If |kᵀm| = ‖k‖₂‖m‖₂
      x = 0,  y = 0
else
      x = z/z(1), x(1) = 0
      y = −(1 + √(1 + ‖x‖²))x/‖x‖²,  y(1) = 1
end
```

This yields the following algorithm.

Algorithm 4.2 (*Simultaneous tridiagonalization*)

Given two $n \times n$ symmetric matrices K, M and a non-zero scalar γ such that $K - \gamma M$ is non-singular, the following algorithm overwrites K and M with the tridiagonal matrices $Q^{\mathrm{T}}KQ$ and $Q^{\mathrm{T}}MQ$, where the non-singular matrix Q is the product of Householder matrices and elementary matrices of the form $I + xy^{\mathrm{T}}$ having minimal condition number.

```
Q = I
[L, D, Π] = ldlt_symm(K − γM)
N = ΠᵀL⁻ᵀD⁻¹L⁻¹Π
for k = 1:n − 2
      [x, y] = elm_xy(K(k:n, k:n), M(k:n, k:n), N(k:n, k))
      K(k:n, k:n) = elm_apply(K(k:n, k:n), x, y)
      M(k:n, k:n) = elm_apply(M(k:n, k:n), x, y)
      N(k:n, k:n) = elm_apply(N(k:n, k:n), −y/(1 + xᵀy), x)
      Q(:, k:n) = Q(:, k:n) + (Q(:, k:n)x)yᵀ

      v = house(K(k + 2:n, k + 1))
      x₁ = [0, −2vᵀ/(vᵀv)]ᵀ,  y₁ = [0, vᵀ]ᵀ
      K(k:n, k:n) = elm_apply(K(k:n, k:n), x₁, y₁)
      M(k:n, k:n) = elm_apply(M(k:n, k:n), x₁, y₁)
      N(k:n, k:n) = elm_apply(N(k:n, k:n), x₁, y₁)
      Q(:, k:n) = Q(:, k:n) + Q(:, k:n)x₁)y₁ᵀ
end
```

This algorithm requires about $13n^3$ operations which is about 50% more than Algorithm 3.1.

5. COMMENTS ON HOW TO CHOOSE γ

A bad choice for $\gamma \neq 0$ may affect the numerical stability of the reduction. Ideally, we would like to choose γ to minimize

$$\kappa(K - \gamma M) = \frac{\max |\lambda_j(\gamma)|}{\min |\lambda_j(\gamma)|}$$

where the $\lambda_j(\gamma)$ are the eigenvalues of $K - \gamma M$. This is a difficult non-linear optimization problem.

If one knows about the location of the eigenvalues of the pair (K, M) then one can choose γ to be outside the region where the eigenvalues lie. For example if (K, M) comes from a first-order system which is known to be stable then all the eigenvalues lie in the left-half plane and it is advisable to take $\gamma > 0$.

If no special information is known about the problem, we suggest to take

$$\gamma = \pm \frac{\|K\|_1}{\|M\|_1} \tag{30}$$

where $\|\cdot\|_1$ denotes the matrix 1-norm, and to choose the sign for γ that maximizes $\|K - \gamma M\|_1$. This choice of sign minimizes the risk of cancellation in computing $K - \gamma M$ but does not guarantee that the resulting γ is far from an eigenvalue.

A simple test to check for a 'bad γ' is to compute the condition number of the block diagonal matrix D in the LDL^{T} factorization of $K - \gamma M$. If $\kappa(D)$ is large then one can choose another value of γ, compute the LDL^{T} factorization of the new $K - \gamma M$ and check $\kappa(D)$ again. It is unlikely this process will need to be repeated many times. The LDL^{T} factorization costs about $n^3/3$ operations, which is small compared with the total number of operations needed for the complete simultaneous tridiagonalization process. Recall that Algorithm 3.1 requires about $8n^3$ operations and Algorithm 4.2 requires about $13n^3$ operations. This is an example of a trade-off between numerical stability and computational cost.

6. NUMERICAL EXPERIMENTS

Our aim in this section is to investigate the numerical properties of the simultaneous tridiagonalization procedures just described. In our tests following quantities are computed:

- the two normalized residual errors

$$\mathscr{R}_K = \frac{\|Q^{\mathrm{T}} K Q - T\|}{\|K\| \|Q\|^2}, \quad \mathscr{R}_M = \frac{\|Q^{\mathrm{T}} M Q - S\|}{\|M\| \|Q\|^2} \tag{31}$$

- the condition number $\kappa(Q)$ of the transformation Q,
- the largest condition number $\kappa(G) = \max_{1 \leqslant k \leqslant n-2} \kappa(G_k)$ of the individual transformations G_k used during the tridiagonalization process.

All our tests have been performed with MATLAB. We named our MATLAB implementations:

- `trd_L`: tridiagonalization by non-orthogonal rank-one modifications of the identity matrix (Algorithm 3.1).
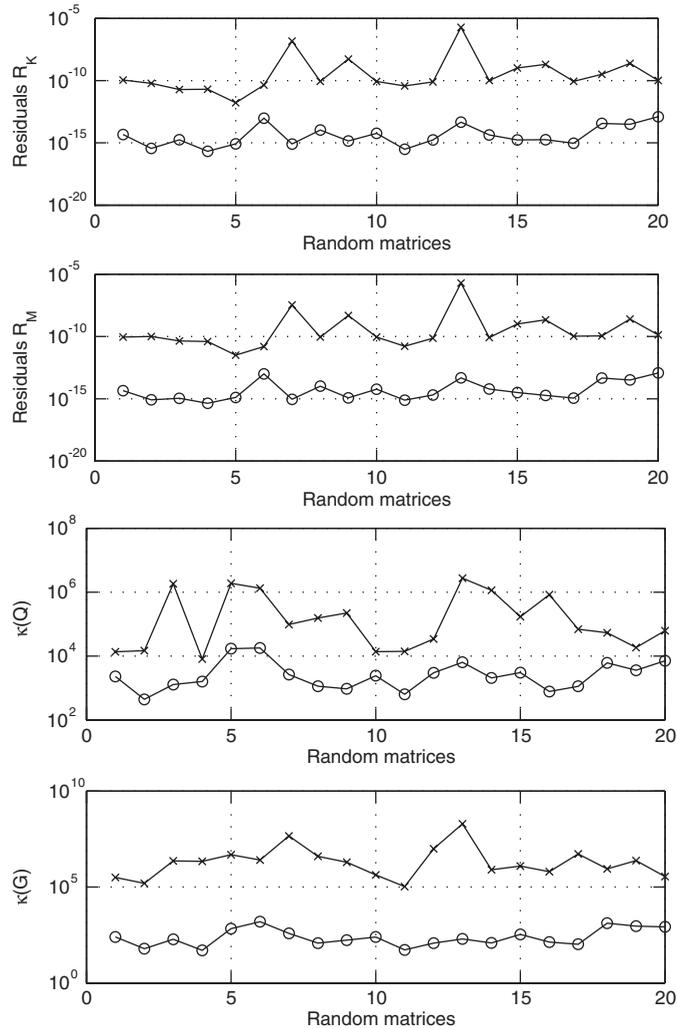
Figure 1. Residuals and condition numbers for 20 random matrices. Results from `trd_L` are marked with '×' and results from `trd_LH` are marked with 'o'.

- `trd_LH`: tridiagonalization by mixed Householder reflectors and non-orthogonal rank-one modifications of the identity matrix with minimal condition numbers (Algorithm 4.2).

We ran a set of tests with random matrices of the form

$$K = \mathtt{randn(n)}; K = K + K'; \quad M = \mathtt{randn(n)}; M = M + M'$$

We took $n = 50$ and chose γ as in (30). The residuals $\mathscr{R}_M$ and $\mathscr{R}_K$ in (31) and the corresponding condition numbers $\kappa(Q)$ and $\kappa(G)$ are plotted in Figure 1 for 20 random matrices. The lines with '×' corresponds to residuals obtained with `trd_L` and the lines with 'o' corresponds to residuals obtained with `trd_LH`. Clearly, the reduction using a combination of Householder
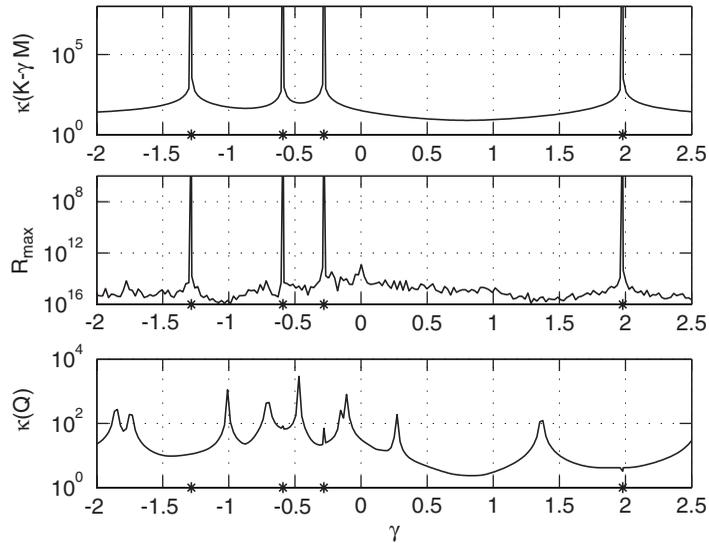
Figure 2. Influence of γ on $\kappa(K - \gamma M)$, $\mathcal{R}_{\max} = \max(\mathcal{R}_K, \mathcal{R}_M)$ and $\kappa(Q)$ for a 6×6 random pair of symmetric matrices (K, M).

reflectors and rank-one modifications of the identity matrix with minimal condition numbers yields smaller residuals and condition numbers. The residuals $\mathcal{R}_M$ and $\mathcal{R}_K$ corresponding to `trd_L` all lie between 10^{-6} and 10^{-12} and are on average of the order 10^{-7}. For `trd_LH` the residuals $\mathcal{R}_M$ and $\mathcal{R}_K$ lie between 10^{-13} and 10^{-16} and are on average of the order 10^{-14}. This shows a large improvement of the quality of the results over those obtained with `trd_L`. Note that for this set of random problems and for a particular algorithm, $\mathcal{R}_M$ and $\mathcal{R}_K$ have the same magnitude. Typically, $\kappa(Q)$ is of the order 10^5 if Q if computed with `trd_L` and of the order 10^3 if computed from `trd_LH`. The last plot in Figure 1 shows that a large condition number for one of the G_k directly affects the residuals and confirms the fact that it is crucial to keep the condition number of each transformation small.

We ran similar tests with random singular matrices M and tests with positive definite K and M and obtained similar conclusions: `trd_LH` behaves more stably than `trd_L`.

We studied the numerical behaviour of `trd_LH` with respect to γ on several random test matrices by letting γ vary between the smallest and largest real eigenvalues of (K, M). The plots in Figure 2 show $\kappa(K - \gamma M)$, $\mathcal{R}_{\max} = \max(\mathcal{R}_K, \mathcal{R}_M)$ and $\kappa(Q)$ for a 6×6 random pair (K, M) and values of γ between -2 and 2.5. The real eigenvalues of (K, M) are marked with '$*$' on the γ-abscisse. As expected, values of γ in the neighbourhood of these real eigenvalues yields large values for $\kappa(K - \gamma M)$. This directly affects the quality of the computed tridiagonal matrices but does not seem to interfere with the value of $\kappa(Q)$.

7. CONCLUSION

A new method has been presented whereby a pair of symmetric matrices (K, M) is successively transformed using elementary transformations (rank-one modifications of the identity matrix)

in such a way that the resulting two matrices (T, S) are both tridiagonal. The total number of operations needed in this reduction process is $O(n^3)$. No assumption on the positivity or non-singularity of K and M is required. However, it is important to choose γ so that $K - \gamma M$ is well conditioned.

We provide two algorithms: Algorithm 3.1, which uses only non-orthogonal elementary transformations and Algorithm 4.2, more costly but with better numerical stability properties, which uses a combination of orthogonal transformations and non-orthogonal elementary transformations with minimal condition numbers. All our numerical experiments confirm the numerical superiority of Algorithm 4.2.

Our simultaneous tridiagonalizing process has particular relevance to the quadratic eigenvalue problem

$$(\lambda^2 M + \lambda D + K)x = 0$$

for two reasons. Firstly, symmetric linearizations of that problem usually result in a generalized eigenvalue problem in which neither matrix is positive definite [6]. More importantly, however, this process may serve as a prototype for an analogous procedure for the quadratic eigenvalue problem in which all three symmetric matrices are modified in steps using elementary structure-preserving co-ordinate transformations [10] such that the eigenvalues of the system are preserved but the structure of all three system matrices becomes progressively more tridiagonal.

REFERENCES

1. Golub GH, Van Loan CF. *Matrix Computations*. (3rd edn). Johns Hopkins University Press: Baltimore, MD, USA, 1996.
2. Bunse-Gerstner A. An algorithm for the symmetric generalized eigenvalue problem. *Linear Algebra and its Applications* 1984; **58**:43–68.
3. Brebner MA, Grad J. Eigenvalues of $Ax = \lambda Bx$ for real symmetric matrices A and B computed by reduction to a pseudosymmetric form and the HR process. *Linear Algebra and its Applications* 1982; **43**:99–118.
4. Zurmühl R, Falk S. *Matrizen und ihre Anwendungen für angewandte Mathematiker, Physiker und Ingenieure. Teil* 2, (5th edn). Springer: Berlin, 1986.
5. Garvey SD, Friswell MI, Prells U. Coordinate transformations for second-order systems. Part I: General transformations. *Journal of Sound and Vibration* 2002; **258**(5):885–909.
6. Tisseur F, Meerbergen K. The quadratic eigenvalue problem. *SIAM Review* 2001; **43**:235–286.
7. Higham NJ. *Accuracy and Stability of Numerical Algorithms*, (2nd edn). Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2002.
8. Higham NJ. The matrix computation toolbox. http://www.ma.man.ac.uk/ higham/mctoolbox.
9. Horn RA, Johnson CR. *Matrix Analysis*, (2nd edn). Cambridge University Press: Cambridge, MA, 1985.
10. Garvey SD, Friswell MI, Prells U. Coordinate transformations for second-order systems. Part II: Elementary structure preserving transformations. *Journal of Sound and Vibration* 2002; **258**(5):911–930.