

On the computation of eigenvectors of a symmetric tridiagonal matrix: comparison of accuracy improvements of Givens and inverse iteration methods

Stéphane Balac, Miloud Sadkane

▶ **To cite this version:**

Stéphane Balac, Miloud Sadkane. On the computation of eigenvectors of a symmetric tridiagonal matrix: comparison of accuracy improvements of Givens and inverse iteration methods. 2003. hal-00137149

HAL Id: hal-00137149
https://hal.science/hal-00137149

Preprint submitted on 16 Mar 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On the computation of eigenvectors of a symmetric tridiagonal matrix: comparison of accuracy improvements of Givens and inverse iteration methods

Stéphane Balac
Laboratoire de Mathématiques Appliquées de Lyon
INSA de Lyon, 69621 Villeurbanne cedex, France
stephane.balac@insa-lyon.fr

Miloud Sadkane
Laboratoire de Mathématiques
Université de Bretagne Occidentale, 29000 Brest, France
miloud.sadkane@univ-brest.fr

Keywords: eigenvalue problem, Sturm sequence, Givens method, inverse iteration method

Abstract

The aim of this paper is the comparison of the recent improvements of two methods to compute eigenvectors of a symmetric tridiagonal matrix once the eigenvalues are computed. The first one is the Givens method which is based on the use of Sturm sequences. This method suffers from a lack of accuracy for the computation of the eigenvector when an approximate value (even a very accurate one) of the eigenvalue is used in the computational process. In [3] the authors introduce a modification of Givens method to ensure the computation of an accurate eigenvector from a good approximation of the corresponding eigenvalue. The second improvement concerns the inverse iteration method. In [8] the authors present a way to determine the best initial vector to start the iterations. Although the two methods and their improvements seem to be very different from a computational point of view, there exists some striking analogies. For instance, in the two methods we look for an optimal index, we have to minimize a residual, *etc.* In the paper we briefly present the two methods and investigate the connections between them.

1 Introduction

This paper is concerned with the computation of the eigenvectors of a real symmetric tridiagonal matrix T once the eigenvalues λ are computed. Inverse iteration method is the most widely used method and is implemented in software libraries like LAPACK, see [1]. A critical problem in the inverse iteration method is the choice of the initial vector to start the iterations. It can be proved, see [11], that the best choice for the initial vector is the rth column of the identity matrix, where r is the largest component of the wanted eigenvector. Unfortunately, this information is not very useful for numerical purposes and for instance in the LAPACK library, a random vector is taken. In 1997, B. Parlett and I. Dhillon devised a way to compute this optimal index r using an LDU and UDL decompositions of the matrix $T - \lambda I$, see [8].
Another well known method for the computation of an eigenvector from an eigenvalue is the Givens method, see [11, p. 299]. This method is a very efficient for computing the eigenvalues of a real tridiagonal matrix using Sturm sequences and a bisection. This method can also be used to derive in a very simple way from the Sturm sequence the eigenvector associated to the

computed eigenvalue. Unfortunately the computation of the eigenvector in that method suffers from numerical instability. In [3] the authors present a way to circumvent this instability and to compute the eigenvector with accuracy from a good approximation of its eigenvalue.
This paper briefly describes the two methods and their improvements. In section 2 we first present Godunov and coworkers improvement of the Givens method. Then in section 3 we present Parlett and Dhillon improvement of the inverse iteration method. Although, from a computational point of view, the two methods seem to be very different, there exists some striking analogies between them. For instance, in the two methods we look for an optimal index, we have to minimize a residual, *etc.* We focus in section 4 on the connections between the two methods, and briefly compare their efficiency. We have implemented the improvements of the two methods under the software MATLAB. The code source can be obtained from the authors.
Let us introduce some notations. Let

$$T = \begin{pmatrix} d_1 & b_2 & & & \\ b_2 & d_2 & b_3 & & \\ & \ddots & \ddots & \ddots & \\ & & b_{N-1} & d_{N-1} & b_N \\ & & & b_N & d_N \end{pmatrix} \tag{1}$$

where $d_n, n = 1, \dots, N$ and $b_n, n = 2, \dots, N$ are given real numbers, with the convention $b_1 = b_{N+1} = 1$. We can assume that T is unreduced, i.e., $b_i \neq 0,\ i = 1, \dots, N$. Otherwise, the eigenvalue problem can be deflated. Let us decompose T as

$$T = U\Lambda U^t, \qquad \Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_N), \qquad U = (U_1, \cdots, U_N) \tag{2}$$

where U is orthogonal. The eigenvalues of T are real, distinct and the first or last component of any eigenvector of T cannot be zero.

2 Godunov and coworkers improvement of Givens method

Givens method is a very efficient method to compute the eigenvalues of a real symmetric tridiagonal matrix using Sturm sequences and the bisection method. Let μ be a real. The left Sturm sequence of first kind is defined from $P_0^+(\mu) = 0$ by the recurrence

$$\forall k \in [\![1, N]\!], \qquad P_k^+(\mu) = \begin{cases} +\infty & \text{if } P_{k-1}^+(\mu) = (d_k - \mu)/|b_k|, \\ 0 & \text{if } P_{k-1}^+(\mu) = +\infty, \\ \dfrac{|b_{k+1}|}{d_k - \mu - |b_k| P_{k-1}^+(\mu)} & \text{otherwise,} \end{cases} \tag{3}$$

whereas the right Sturm sequence of first kind is defined by $P_N^-(\mu) = +\infty$ and

$$\forall k \in [\![1, N]\!], \qquad P_{k-1}^-(\mu) = \begin{cases} \dfrac{(d_k - \mu)}{|b_k|} & \text{if } P_k^-(\mu) = +\infty, \\ +\infty & \text{if } P_k^-(\mu) = 0, \\ \dfrac{1}{|b_k|}\left((d_k - \mu) - \dfrac{|b_{k+1}|}{P_k^-(\mu)}\right) & \text{otherwise.} \end{cases} \tag{4}$$

It is well known that the Sturm sequences of first kind are related to the leading principal minors of T, see [11]. Here the subscript + indicates that the minors are taken with rows in increasing order whereas the subscript − indicates that the minors are with rows in decreasing order. Both the right and left Sturm sequences satisfy the same induction relations

$$\forall k \in [\![1, N-1]\!], \quad \begin{cases} P_k^{\pm}(\mu) = +\infty & \text{if } P_{k-1}^{\pm}(\mu) = (d_k - \mu)/|b_k|, \\ P_k^{\pm}(\mu) = 0 & \text{if } P_{k-1}^{\pm}(\mu) = +\infty, \\ (d_k - \mu) - |b_k| \; P_{k-1}^{\pm}(\mu) - \dfrac{|b_{k+1}|}{P_k^{\pm}(\mu)} = 0 & \text{otherwise,} \end{cases} \tag{5}$$

and differ only in the boundary condition for $k = 0$ and $k = N$. A Sturm sequence is termed two-sided when it satisfies the two boundary conditions $P_0^{\pm}(\mu) = 0$ and $P_N^{\pm}(\mu) = +\infty$. A straightforward calculation shows that the sequence $(P_k^{\pm}(\lambda_n))_{k\in[\![0,N]\!]}$ is a two-sided Sturm sequence.

Using Givens theorem along with a bisection allow the computation of the eigenvalue of T in an accurate and stable way, see [11, p. 298]. Indeed, let the quantities $P_1^+(\mu), \ldots, P_N^+(\mu)$ be evaluated for some value μ; then the number of agreements in sign of consecutive numbers of this sequence is the number of eigenvalues of T which are strictly greater than μ (if $P_k^+(\mu) = 0$ then $P_k^+(\mu)$ is taken to have the opposite sign to that of $P_{k-1}^+(\mu)$, no two consecutive terms can be zero). Furthermore an eigenvector $U_n = (u_1, \cdots, u_N)^t$ associated with the eigenvalue λ_n can be computed from the Sturm sequence $P_k^{\pm}(\lambda_n)$ through the relations: $u_1 = 1$,

$$\forall k \in [\![2, N]\!], \quad \begin{cases} u_k = 0 & \text{if } P_{k-1}^{\pm}(\lambda_n) = +\infty, \\ u_k = -\dfrac{b_{k-1}}{b_k} u_{k-2} & \text{if } P_{k-1}^{\pm}(\lambda_n) = 0, \\ u_k = -\text{sign}(b_k) \dfrac{u_{k-1}}{P_{k-1}^{\pm}(\lambda_n)} & \text{otherwise.} \end{cases} \tag{6}$$

Although the $P_k^{\pm}(\lambda_n)$ determine the eigenvalue in a stable way, the explicit use of expressions (6) to compute components of the eigenvector does not necessarily lead to a good approximation of the eigenvector, see [3] for some examples. The reason for this lack of accuracy is that even if the eigenvalue λ_n is computed with a very good accuracy, the Sturm sequences $(P_k^+(\widetilde{\lambda}_n))_{k\in[\![0,N]\!]}$ or $(P_k^-(\widetilde{\lambda}_n))_{k\in[\![0,N]\!]}$ where $\widetilde{\lambda}_n$ is an approximation of λ_n, may not be a two-sided Sturm sequence. In practice, a two-sided Sturm sequence is extremely unlikely to occur even when $\widetilde{\lambda}_n$ is the closet machine number to the eigenvalue λ_n.

The idea of Godunov and coworkers to circumvent this drawback is to enforce the Sturm sequence in $\widetilde{\lambda}_n$ to be two-sided. This is obtained by joining the left Sturm sequence $(P_k^+(\widetilde{\lambda}_n))_{k\in[\![0,N]\!]}$ to the right Sturm sequence $(P_k^-(\widetilde{\lambda}_n))_{k\in[\![0,N]\!]}$ at a well chosen integer $k_0 \in [\![0, N]\!]$. The new sequence $(Q_k(\widetilde{\lambda}_n))_{k\in[\![0,N]\!]}$ automatically satisfies the boundary conditions $Q_0(\lambda) = 0$ and $Q_N(\lambda) = +\infty$. The sequence $(Q_k(\widetilde{\lambda}_n))_{k\in[\![0,N]\!]}$ is generally not a two-sided Sturm sequence for the matrix T. However, it can be proved (see proposition 1 below) that it is a two-sided Sturm sequence for a matrix $\widetilde{T}$ close to T and that the eigenvector $\widetilde{U}_n$ computed using relations (6) with the sequence $(Q_k(\widetilde{\lambda}_n))_{k\in[\![0,N]\!]}$ is a good approximation of the eigenvector U_n.

In order to prove the correctness of their approach, Godunov and coworkers introduce two more sequences, called "Sturm sequences of second kind". For $\lambda \in \mathbb{R}$, the left Sturm sequence of

second kind $(\phi_k^+(\lambda))_{k\in[\![0,N]\!]}$ is defined from the left Sturm sequence of first kind by the relation

$$\phi_j^+(\lambda) = \arctan(P_j^+(\lambda)) + \tau_j^+\pi, \tag{7}$$

where τ_j^+ is the number of non positive terms in the sequence $P_1^+(\lambda), \ldots, P_j^+(\lambda)$. Similarly for $\lambda \in \mathbb{R}$ and $m \in \mathbb{N}$, the right Sturm sequence of second kind $(\phi_k^-(\lambda))_{k\in[\![0,N]\!]}$ is defined from the right Sturm sequence of first kind by the relation

$$\phi_j^-(\lambda) = \arctan(P_j^-(\lambda)) + (m - 1 - \tau_{j+1}^-)\pi, \tag{8}$$

where τ_{j+1}^- is the number of non positive terms in the sequence $P_{j+1}^-(\lambda), \ldots, P_{N-1}^-(\lambda)$, and $\tau_N^- = 0$. Sturm sequences of second kind are central in Godunov and co-workers improvement of Givens method. We briefly summarize their main properties, in order to understand the result given in proposition 2 below and the connections between the two methods discussed in section 4.

First, we have for all $j \in [\![0, N]\!]$, $P_j^+(\lambda) = \tan(\phi_j^+(\lambda))$ and $P_j^-(\lambda) = \tan(\phi_j^-(\lambda))$. Each function ϕ_j^+ increases continuously and monotonically from 0 to $j\pi$ whereas each function ϕ_j^- decreases continuously and monotonically. For a fixed λ, the sequences $(\phi_k^+(\lambda))_{k\in[\![0,N]\!]}$ and $(\phi_k^-(\lambda))_{k\in[\![0,N]\!]}$ are not necessarily monotone. However, they can vary only in the following way:

$$\text{if} \quad \phi_k^+(\lambda) \in]\tau_k^+\pi - \pi/2, \tau_k^+\pi[\quad \text{then} \quad \begin{cases} \phi_{k-1}^+(\lambda) \in](\tau_k^+ - 1)\pi, \tau_k^+\pi - \pi/2[, \\ \phi_{k+1}^+(\lambda) \in]\tau_k^+\pi, \tau_k^+\pi + \pi/2[, \end{cases}$$

and

$$\text{if} \quad \phi_k^+(\lambda) \in]\tau_k^+\pi, \tau_k^+\pi + \pi/2[\quad \text{then} \quad \begin{cases} \phi_{k-1}^+(\lambda) \in](\tau_k^+ - 1)\pi, \tau_k^+\pi - \pi/2[, \\ \phi_{k+1}^+(\lambda) \in]\tau_k^+\pi, \tau_k^+\pi + \pi/2[. \end{cases}$$

$$\text{If} \quad \phi_k^-(\lambda) \in]p_k^-\pi - \pi/2, p_k^-\pi[\quad \text{then} \quad \begin{cases} \phi_{k-1}^-(\lambda) \in](p_k^- - 1)\pi - 3\pi/2, \tau_k^-\pi - \pi/2[, \\ \phi_{k+1}^-(\lambda) \in]p_k^-\pi, p_k^-\pi + \pi[, \end{cases}$$

and

$$\text{if} \quad \phi_k^-(\lambda) \in](p_k^- - 1)\pi, p_k^-\pi - \pi/2[\quad \text{then} \quad \begin{cases} \phi_{k-1}^-(\lambda) \in]p_k^-\pi - 3\pi/2, p_k^-\pi - \pi/2[, \\ \phi_{k+1}^-(\lambda) \in](p_k^- - 1)\pi, p_k^-\pi[. \end{cases}$$

¿From the relations (3) and (4) for the Sturm sequences of first kind, we can deduce relations for the Sturm sequences of second kind. We have

$$\phi_j^+(\lambda) = \omega(\phi_{j-1}^+(\lambda), |b_j|, d_j - \lambda, |b_{j+1}|) \qquad j \in [\![1, N]\!] \tag{9}$$

where for $c_1, c_3 \in \mathbb{R}_+^*$ and $c_2 \in \mathbb{R}$, the real function $\widehat{\omega} : x \in \mathbb{R} \longmapsto \omega(x, c_1, c_2, c_3)$ is continuously differentiable and strictly increasing. In the same way, we have

$$\phi_{j-1}^-(\lambda) = \gamma(\phi_j^-(\lambda), |b_j|, d_j - \lambda, |b_{j+1}|) \qquad j \in [\![1, N]\!], \tag{10}$$

where for $c_1, c_3 \in \mathbb{R}_+^*$ and $c_2 \in \mathbb{R}$, the real function $\widehat{\gamma} : x \in \mathbb{R} \longmapsto \gamma(x, c_1, c_2, c_3)$ is the inverse of $\widehat{\omega}$.

Godunov's method, which consists in joining a left Strum sequence to a right Sturm sequence at a well chosen index k_0 to obtain the required two-sided Sturm sequence to compute the eigenvector, is justified by the following proposition given in [3].

Proposition 1 *Let λ_n be the nth eigenvalue of T and $x_n, y_n \in \mathbb{R}$ be the upper and lower bound of the last interval in the bisection method used to compute the approximate eigenvalue (so that $x_n \leq \lambda_n \leq y_n$ and $\widetilde{\lambda}_n = \frac{x_n+y_n}{2}$ is the approximation of λ_n). Then the two following statements hold.*

- There exists an integer $k_0 \in [\![1, N]\!]$ such that

$$\phi^+_{k_0-1}(y_n) \leq \phi^-_{k_0-1}(x_n), \qquad \text{and} \qquad \phi^+_{k_0}(y_n) \geq \phi^-_{k_0}(x_n). \tag{11}$$

- There exists a real number $\tau \in [0, 1]$ such that the sequence $(\psi_k)_{k\in[\![0,N]\!]}$ defined by

$$\begin{aligned} \psi_k &= \phi^+_k(y_n), && \forall k = 0, \ldots, k_0 - 1, \\ \psi_k &= \phi^-_k(x_n), && \forall k = k_0, \ldots, N, \end{aligned} \tag{12}$$

is the two-sided Sturm sequence of second kind with parameters $(n, \widetilde{\lambda}_n)$ for the tridiagonal matrix $\widetilde{T}$ defined by

$$\widetilde{T} = \begin{pmatrix} \widetilde{d}_1 & b_2 & & & \\ b_2 & \widetilde{d}_2 & b_3 & & \\ & \ddots & \ddots & \ddots & \\ & & b_{N-1} & \widetilde{d}_{N-1} & b_N \\ & & & b_N & \widetilde{d}_N \end{pmatrix} \tag{13}$$

where

$$\widetilde{d}_k = \begin{cases} d_k - \frac{1}{2}(y_n - x_n) & \text{if } k = 1, \ldots, k_0 - 1, \\ (1-\tau)(d_{k_0+1} - \frac{1}{2}(y_n - x_n)) + \tau(d_{k_0+1} + \frac{1}{2}(y_n - x_n)) & \text{if } k = k_0, \\ d_k + \frac{1}{2}(y_n - x_n) & \text{if } k = k_0 + 1, \ldots, N. \end{cases} \tag{14}$$

Proposition 1 guarantees that the two sequences $(\phi^+_k(y_n))_{k\in[\![0,N]\!]}$ and $(\phi^-_k(x_n))_{k\in[\![0,N]\!]}$ cross so that the sequence $(\psi_k)_{k\in[\![0,N]\!]}$ is always defined. It expresses that $\widetilde{\lambda}_n$ is the nth eigenvalue of $\widetilde{T}$ and that:

$$\|T - \widetilde{T}\|_p = \frac{1}{2}|y_n - x_n|, \quad p = 1, 2, \infty. \tag{15}$$

Proposition 1 also gives the way to compute an approximation of the eigenvector U_n associated with the eigenvalue λ_n. We recall that relations (6) are unsuited for the computation of approximation to U_n because the Sturm sequence of the first kind $P^+_k(\widetilde{\lambda}_n)$ is not two-sided. Now, the sequence $(\psi_k)_{k\in[\![0,N]\!]}$ is a two-sided Sturm sequence of second kind for the matrix $\widetilde{T}$ which, according to (15), is closed to T. We can obtain a two-sided Sturm sequence of first kind $(Q_k(\widetilde{\lambda}_n))_{k\in[\![0,N]\!]}$ for $(\psi_k)_{k\in[\![0,N]\!]}$ by the relations:

$$Q_k(\widetilde{\lambda}_n) = \tan\psi_k, \quad \forall k \in [\![0, N]\!]. \tag{16}$$

It is then possible to use relations (6) with the two-sided Sturm sequence $(Q_k(\widetilde{\lambda}_n))_{k\in[\![0,N]\!]}$ to compute an eigenvector $\widetilde{U}_n$ of $\widetilde{T}$ associated with $\widetilde{\lambda}_n$ in an accurate and stable way. The eigenpair $(\widetilde{\lambda}_n, \widetilde{U}_n)$ of $\widetilde{T}$ is a good approximation of (λ_n, U_n) if $\widetilde{\lambda}_n$ is well separated from all eigenvalues

$\lambda_j \neq \lambda_n$ since from standard perturbation theory (Davis and Kahan theorem), see [9], and relation (15) we have

$$|\sin\angle(U_n, \widetilde{U}_n)| \leq \frac{\|T\ \widetilde{U}_n - \widetilde{\lambda}_n\ \widetilde{U}_n\|_2}{gap(\widetilde{\lambda}_n)} \leq \frac{|y_n - x_n|}{2\ gap(\widetilde{\lambda}_n)}, \tag{17}$$

$$|\lambda_n - \widetilde{\lambda}_n| \leq \frac{\|T\ \widetilde{U}_n - \widetilde{\lambda}_n\ \widetilde{U}_n\|_2^2}{gap(\widetilde{\lambda}_n)} \leq \frac{|y_n - x_n|^2}{4\ gap(\widetilde{\lambda}_n)}, \tag{18}$$

where $gap(\widetilde{\lambda}_n) = \min\{|\widetilde{\lambda}_n - \lambda_j|, j \neq n\}$. However the method does not guarantee that for closed eigenvalues, the corresponding computed vectors are orthogonal.

Of course the method seems very tedious from a computational point of view because of the use of Sturm sequences of second kind. In fact there is no need to compute them. The index k_0 for which the left and right Sturm sequences of first kind join can be characterized using the sequences $(P_k^-(x_n))_{[\![0,N]\!]}$ and $(P_k^+(y_n))_{[\![0,N]\!]}$ as stated in the following proposition given in [3].

Proposition 2 *For $k \in \{1, \ldots, N\}$ let us consider the integer p_k^+ and p_k^- defined respectively by $p_k^+ = \tau_k^+$ and $p_k^- = n - 1 - \tau_{k+1}^-$ where τ_k^+ is the number of non positive terms in the sequence $P_1^+(y_n), \ldots, P_k^+(y_n)$ and τ_k^- is the number of non positive terms in the sequence $P_k^-(x_n), \ldots, P_{N-1}^-(x_n)$. Let*

$$\mathcal{K} = \left\{k \in \{1, \cdots, N\} \mid \left(p_{k-1}^+ < p_{k-1}^-\right) \text{ or } \left(p_{k-1}^+ = p_{k-1}^- \text{ and } P_{k-1}^+(y_n) \leq P_{k-1}^-(x_n)\right)\right\}. \tag{19}$$

Then, the set $\mathcal{K}$ is not empty and $\ell = \max\{k \in \mathcal{K}\}$ satisfies:

$$\phi_{\ell-1}^+(y_n) \leq \phi_{\ell-1}^-(x_n) \qquad \text{and} \qquad \phi_\ell^+(y_n) \geq \phi_\ell^-(x_n).$$

Therefore the index k_0 coincides with $\ell = \max\{k \in \mathcal{K}\}$. This means that k_0 is the greatest integer k satisfying $p_{k-1}^+ < p_{k-1}^-$ or $p_{k-1}^+ = p_{k-1}^-$ and $P_{k-1}^+(y_n) \leq P_{k-1}^-(x_n)$.

We summarize the method to compute the eigenvectors of a symmetric tridiagonal matrix T in the following algorithm. It assumes that the eigenvalues $\lambda_n, n \in [\![1, N]\!]$ have been computed with accuracy by the bisection method and that $x_n, y_n \in \mathbb{R}$ are the upper and lower bound of the last interval.

compute the left Sturm sequence of first kind in y_n: $P_0^+(y_n), \ldots, P_N^+(y_n)$
compute the right Sturm sequence of first kind in x_n: $P_0^-(x_n), \ldots, P_N^-(x_n)$

for $k = 1$ to N do

compute τ_k^+ the number of non positive numbers in the sequence $P_1^+(y_n), \ldots, P_k^+(y_n)$
compute τ_k^- the number of non positive numbers in the sequence $P_k^-(x_n), \ldots, P_{N-1}^-(x_n)$

end-do

compute the greatest index k_0 for which
$\tau_{k_0-1}^+ < n - 1 - \tau_{k_0}^-$
or
$\tau_{k_0-1}^+ = n - 1 - \tau_{k_0}^-$ *and* $P_{k_0-1}^+(y_n) \leq P_{k_0-1}^-(x_n)$

Form the sequence $(P_k)_{k \in [\![0,N]\!]} = P_0^+(y_n), \cdots, P_{k_0-1}^+(y_n),\ P_{k_0}^-(x_n), \cdots, P_N^-(x_n)$
Set $U_n(1) = 1$
for $j = 1$ to $N - 1$ do

$$U_n(j+1) = -sign(T_{j+1,j})\ \frac{U_n(j)}{P_j}$$

end-do

end-do

All the relations mentioned so far hold in exact arithmetic. Godunov and coworkers show that the method guarantees accuracy even in finite precision arithmetic and that no overflow occurs if the data are normalized in a prescribed manner, see [3, chp. 5]. We have implemented the algorithm under MATLAB software in both cases.

3 Parlett and Dhillon improvement of the inverse iteration method

The basic idea of inverse iteration method to compute an eigenvector associated to a given eigenvalue. The eigenvector U_n associated with λ_n is defined as the solution to the linear system $(T-\lambda_n I)U_n = 0$. As the matrix $(T-\lambda_n I)$ is singular, $N-1$ equations from the system determine the eigenvector up to a scalar multiple. However, in practice we have a good approximation $\widetilde{\lambda}_n$ of the eigenvalue λ_n, which is often close to, but different from, λ_n. This implies that the matrix $T-\widetilde{\lambda}_n I$ is nonsingular and the only solution to the linear system $(T-\widetilde{\lambda}_n I)X = 0$ is the null vector. A way to get an approximation $\widetilde{U}_n$ of the eigenvector is to select $N-1$ equations from the linear system $(T-\widetilde{\lambda}_n I)X = 0$ (discarding say the rth) and to solve the resulting under-determined system. The discarded equation produces a residual $(T-\widetilde{\lambda}_n I)\widetilde{U}_n$ whose all components are zero except the rth. The central point in the process is to determine the best choice for the equation to discard and control the accuracy of the approximation $\widetilde{U}_n$. As presented in [11], let us consider the linear system:

$$(T-\widetilde{\lambda}_n I)X = b \tag{20}$$

where b is an arbitrary normalized vector. If b is expressed in the form

$$b = \sum_{j=1}^{N} \gamma_j U_j$$

then the solution X_0 to the system is

$$X_0 = \sum_{j=1}^{N} \frac{\gamma_j}{\lambda_j - \widetilde{\lambda}_n} U_j. \tag{21}$$

It follows that if $\widetilde{\lambda}_n$ is close to λ_n but not to any other λ_j then

$$\frac{\gamma_n}{\lambda_n - \widetilde{\lambda}_n} \lll \frac{\gamma_j}{\lambda_j - \widetilde{\lambda}_n} \qquad \forall j \neq n.$$

This means that X_0 is much richer in U_n than b is. We can repeat the process taking X_0 as right side term for the linear system. The solution X_1 will be even richer than X_0 in the vector U_n. This iterative process to approximate eigenvectors is known as the inverse iteration method. Thus the best choice for the equation to be omitted is the rth equation with r corresponding to the largest component of U_n. This means that the best starting vector in the inverse iteration method is e_r. The result is instructive but not useful at all since the index of the largest component of the eigenvector to be computed is not known a priori. In [8], Parlett and Dhillon give a practical way to determine the index r. Their approach is valid for normal triangular matrices that permit LDU and UDL factorizations. We summarize it as it is although we are only interested in the the symmetric case.

Proposition 3 *Assume that for all λ in a neighborhood of λ_n the matrices $J_\lambda = T - \lambda I$ are normal and permit triangular factorization $J_\lambda = L^+ D^+ U^+$ and $J_\lambda = U^- D^- L^-$ where $D^+ = \mathrm{diag}(D_1^+, \cdots, D_N^+)$ and $D^- = \mathrm{diag}(D_1^-, \cdots, D_N^-)$ are diagonal matrices, L^+ and L^- are lower triangular matrices, U^+ and U^- are upper triangular matrices (all these last four with 1's on the diagonal). For $j = 1, \cdots, N$ the solution $(Z_\lambda^{(j)}, \delta_\lambda^{(j)}) \in \mathbb{R}^N \times \mathbb{R}$ to the system*

$$\left\{ \begin{array}{rcl} J_\lambda \; Z_\lambda^{(j)} & = & \delta_\lambda^{(j)} \; e_j \\ Z_\lambda^{(j)}(j) & = & 1 \end{array} \right. \tag{22}$$

satisfies $\delta_\lambda^{(j)} = D_j^+ + D_j^- - J_{j,j}$. Moreover we have

$$\lim_{\lambda \to \lambda_n} \frac{\left(\delta_\lambda^{(j)}\right)^{-1}}{\displaystyle\sum_{l=1}^{N} \left(\delta_\lambda^{(l)}\right)^{-1}} = |u_j|^2. \tag{23}$$

Thus to determine the largest component of U_n it suffices to determine the index j for which $\delta_{\lambda_n}^{(j)}$ is minimum. As the exact value of the eigenvalue λ_n is unknown, we look for the index j for which $\delta_{\widetilde{\lambda}_n}^{(j)}$ is minimum where $\widetilde{\lambda}_n$ is an accurate approximation of λ_n. Each $\delta_{\widetilde{\lambda}_n}^{(j)}$ is computed from the relation

$$\delta_{\widetilde{\lambda}_n}^{(j)} = \widetilde{D}_j^+ + \widetilde{D}_j^- - \widetilde{J}_{j,j} \tag{24}$$

where the matrices $\widetilde{D}^+$ and $\widetilde{D}^-$ come from the triangular factorizations of $J_{\widetilde{\lambda}_n} = T - \widetilde{\lambda}_n I$.

4 Connections between the two methods

Although the methods appear at first sight to be very different, there exist various connections between them. Some of these connections were already mentioned by Parlett and Dhillon in [8]. We point out some others.

4.1 Connections concerning the computed terms

Godunov and coworkers method is based on the computation of Sturm sequences whereas Parlett and Dhillon method is based on the LDU decomposition of T. As mentioned in [8] the Sturm sequences can be obtained in a very straight manner from the LDU decomposition. Indeed, let $J_\lambda = T - \lambda I$ with T tridiagonal symmetric with decomposition $J_\lambda = L^+ D^+ (L^+)^T = L^- D^- (L^-)^T$. The matrices $L^\pm$ and $D^\pm$ can be computed explicitly. In particular, we obtain the following expression for the diagonal matrix $D = \mathrm{diag}(D_1^+, \cdots, D_N^+)$,

$$\left\{ \begin{array}{rcll} D_1^+ & = & d_1 - \lambda, & \\ D_k^+ & = & d_k - \lambda - \dfrac{b_k^2}{D_{k-1}^+} & k \in [\![2, N]\!]. \end{array} \right. \tag{25}$$

On the other hand, the left Sturm sequence of first kind $(P_k^+(\lambda))_{k \in [\![0,N]\!]}$ is defined, see relation (3), by

$$P_k^+(\lambda) = \frac{|b_{k+1}|}{d_k - \lambda - |b_k| P_{k-1}^+(\lambda)} = \frac{|b_{k+1}|}{d_k - \lambda - \dfrac{b_k^2}{|b_k|} \; P_{k-1}^+(\lambda)}. \tag{26}$$

It follows from (25) and (26) that $P_k^+(\lambda)$ and D_k^+ are connected by the relation

$$P_k^+(\lambda) = \frac{|b_{k+1}|}{D_k^+}. \tag{27}$$

Similarly, we show that the following relation between the right sequence of first kind $(P_k^-(\lambda))_{k\in[\![0,N]\!]}$ and the diagonal matrix D^- holds

$$P_k^-(\lambda) = \frac{D_{k+1}^-}{|b_{k+1}|} \quad \forall k \in [\![1,N]\!]. \tag{28}$$

Thus the basic tools in the two methods (Sturm sequences in Godunov method, and LDU decomposition of Parlett and Dhillon) are connected through the relations (27) and (28).

4.2 Connections between the optimal indices

Both methods look for a particular integer termed the optimal index. In one hand, in Parlett and Dhillon approach we look for an integer $j_0 \in [\![1,N]\!]$ such that $|\delta^{(j_0)}| = \min_{j\in[\![1,N]\!]} |\delta^{(j)}|$ where for all $j \in [\![1,N]\!]$, $(Z^{(j)}, \delta^{(j)})$ is the solution of the system

$$\left\{ \begin{array}{rcl} (T - \widetilde{\lambda}_n I)\ Z^{(j)} & = & \delta^{(j)}\ e_j \\ Z^{(j)}(j) & = & 1 \end{array} \right. . \tag{29}$$

On the other hand, in Godunov and coworkers approach we join together a left and right Sturm sequences at a well chosen index $k_0 \in [\![1,N]\!]$ to obtain a two-sided Sturm sequence. The question is then: are these two indices the same? To answer, let us first take an example (a random tridiagonal matrix) and determine the optimal indices j_0 and k_0. The result is depicted in figure 1. One can see that for approximatively half of the eigenvectors the two indices are the same. For the other ones even if the optimal index k_0 differs from j_0 (the index of the maximal component of the eigenvector under consideration) it is always (except in one case) among the 25% greatest components of this eigenvector. In the sequel, we will try to explain where this phenomenon originate from.

The linear system (29) has the following full expression (with an obvious modification when $j = 1$ and $j = N$),

$$\left\{ \begin{array}{lcll} (d_1 - \widetilde{\lambda}_n)z_1^{(j)} + b_2 z_2^{(j)} & = & 0, & \\ b_k z_{k-1}^{(j)} + (d_k - \widetilde{\lambda}_n)z_k^{(j)} + b_{k+1} z_{k+1}^{(j)} & = & 0, & \text{for } k = 2,\ldots,j-1, \\ b_j z_{j-1}^{(j)} + (d_j - \widetilde{\lambda}_n)z_j^{(j)} + b_{j+1} z_{j+1}^{(j)} & = & \delta^{(j)}, & \\ b_k z_{k-1}^{(j)} + (d_k - \widetilde{\lambda}_n)z_k^{(j)} + b_{k+1} z_{k+1}^{(j)} & = & 0, & \text{for } k = j+1,\ldots,N-1, \\ b_N z_{N-1}^{(j)} + (d_N - \widetilde{\lambda}_n)z_N^{(j)} & = & 0, & \\ z_j^{(j)} = 1. & & & \end{array} \right. \tag{30}$$

First assume for convenience that $z_k^{(j)} \neq 0, \forall k \in [\![1,N]\!]$ (the general case use the same ideas but

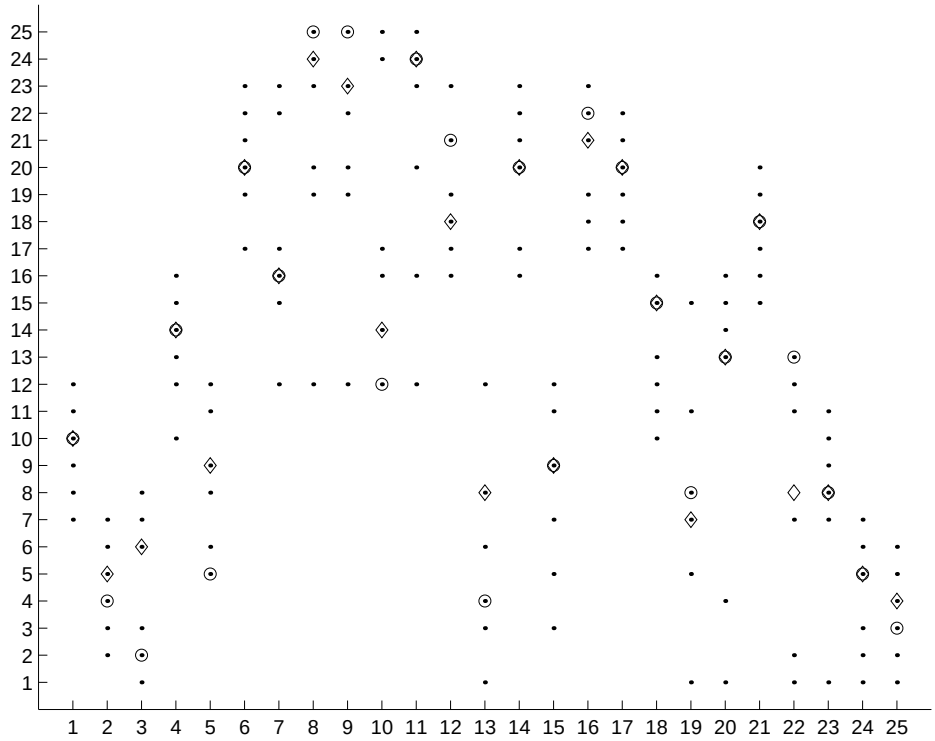


Figure 1: Position of indices j_0 ($\circ$), k_0 ($\diamond$) and j ($\cdot$) such that $|u_j|$ is among the 25% greatest components of the eigenvector u, for a random tridiagonal matrix of size 25.

is much more cumbersome to handle). We have,

$$\left\{\begin{array}{rcll} (d_1-\widetilde{\lambda}_n)+b_2\dfrac{z_2^{(j)}}{z_1^{(j)}} & = & 0, & \\ b_k\dfrac{z_{k-1}^{(j)}}{z_k^{(j)}}+(d_k-\widetilde{\lambda}_n)+b_{k+1}\dfrac{z_{k+1}^{(j)}}{z_k^{(j)}} & = & 0, & k=2,\dots,j-1,\\ b_jz_{j-1}^{(j)}+(d_j-\widetilde{\lambda}_n)+b_{j+1}z_{j+1}^{(j)} & = & \delta^{(j)}, & \\ b_k\dfrac{z_{k-1}^{(j)}}{z_k^{(j)}}+(d_k-\widetilde{\lambda}_n)+b_{k+1}\dfrac{z_{k+1}^{(j)}}{z_k^{(j)}} & = & 0, & k=j+1,\dots,N-1,\\ b_N\dfrac{z_{N-1}^{(j)}}{z_N^{(j)}}+(d_N-\widetilde{\lambda}_n) & = & 0. & \end{array}\right. \tag{31}$$

Then introduce the sequence $(Q_k(\widetilde{\lambda}_n))_{k\in[\![0,N]\!]}$ defined by $Q_0(\widetilde{\lambda}_n)=0$, $Q_N(\widetilde{\lambda}_n)=+\infty$ and for $k\in[\![1,N-1]\!]$,

$$Q_k(\widetilde{\lambda}_n)=-\mathrm{sign}(b_{k+1})\frac{z_k^{(j)}}{z_{k+1}^{(j)}}. \tag{32}$$

It follows from (31) that the sequence $(Q_k(\widetilde{\lambda}_n))_{k\in[\![1,N-1]\!]}$ satisfies

$$\left\{\begin{array}{l}(d_1-\widetilde{\lambda}_n)-\dfrac{|b_2|}{Q_1(\widetilde{\lambda}_n)}=0,.\\ (d_k-\widetilde{\lambda}_n)-|b_k|\ Q_{k-1}(\widetilde{\lambda}_n)-\dfrac{|b_{k+1}|}{Q_k(\widetilde{\lambda}_n)}=0, \qquad k\in[\![2,j-1]\!],\\ (d_j-\widetilde{\lambda}_n)-|b_j|\ Q_{j-1}(\widetilde{\lambda}_n)-\dfrac{|b_{j+1}|}{Q_j(\widetilde{\lambda}_n)}=\delta^{(j)}, \\ (d_k-\widetilde{\lambda}_n)-|b_k|\ Q_{k-1}(\widetilde{\lambda}_n)-\dfrac{|b_{k+1}|}{Q_k(\widetilde{\lambda}_n)}=0, \qquad k\in[\![j+1,N-1]\!].\\ |b_N|Q_{N-1}(\widetilde{\lambda}_n)-(d_N-\widetilde{\lambda}_n)=0.\end{array}\right. \tag{33}$$

Now, for $k\in[\![0,j-1]\!]$ we consider

$$\varphi_k(\widetilde{\lambda}_n)=\arctan(Q_k(\widetilde{\lambda}_n))+\tau_k^+\pi, \tag{34}$$

where τ_k^+ is the number of non positive terms in the sequence $Q_1(\widetilde{\lambda}_n),\ \dots\ ,\ Q_k(\widetilde{\lambda}_n)$. Clearly $\varphi_0(\widetilde{\lambda}_n),\ \dots\ ,\ \varphi_{j-1}(\widetilde{\lambda}_n)$ are the jth first terms of the left Sturm sequence of second kind $(\phi_k^+(\widetilde{\lambda}_n))_k$. In a similar way, for $k=j,\cdots,N-1$ we consider

$$\varphi_k(\widetilde{\lambda}_n)=\arctan(Q_k(\widetilde{\lambda}_n))+(n-1-\tau_{k+1}^+)\pi \tag{35}$$

where τ_{k+1}^+ is the number of non positives terms in the sequence $Q_{k+1}(\widetilde{\lambda}_n),\dots,Q_{N+1}(\widetilde{\lambda}_n)$. Clearly $\varphi_j(\widetilde{\lambda}_n),\dots,\varphi_N(\widetilde{\lambda}_n)$ are the $N-j+1$th last terms of the right Sturm sequence of second kind $(\phi_k^-(\widetilde{\lambda}_n))_k$.
¿From (7), (8) and (9) we deduce that the sequence $(\varphi_k(\widetilde{\lambda}_n))_k$ satisfies

$$\left\{\begin{array}{rcll}\varphi_k(\widetilde{\lambda}_n) &=& \omega(\varphi_{k-1}(\widetilde{\lambda}_n),|b_k|,d_k-\widetilde{\lambda}_n,|b_{k+1}|) & k=1,\dots,j-1,\\ \varphi_j(\widetilde{\lambda}_n) &=& \omega(\varphi_{j-1}(\widetilde{\lambda}_n),|b_j|,d_j-\delta^{(j)}-\widetilde{\lambda}_n,|b_{j+1}|), & \\ \varphi_k(\widetilde{\lambda}_n) &=& \omega(\varphi_{k-1}(\widetilde{\lambda}_n),|b_k|,d_k-\widetilde{\lambda}_n,|b_{k+1}|) & k=j+1,\dots,N.\end{array}\right. \tag{36}$$

The sequence $(\varphi_k(\widetilde{\lambda}_n))_k$ is not a Sturm sequence of second kind for T but is the linkage between the left and right Sturm sequences $(\phi_k^+(\widetilde{\lambda}_n))_{k\in[\![0,N]\!]}$ and $(\phi_k^-(\widetilde{\lambda}_n))_{k\in[\![0,N]\!]}$. We have

$$\begin{array}{rcl}\varphi_j(\widetilde{\lambda}_n)-\phi_j^+(\widetilde{\lambda}_n) &=& \phi_j^-(\widetilde{\lambda}_n)-\phi_j^+(\widetilde{\lambda}_n)\\ &=& \omega(\phi_{j-1}^+(\widetilde{\lambda}_n),|b_j|,d_j-\delta^{(j)}-\widetilde{\lambda}_n,|b_{j+1}|)\\ && -\omega(\phi_{j-1}^+(\widetilde{\lambda}_n),|b_j|,d_j-\widetilde{\lambda}_n,|b_{j+1}|).\end{array} \tag{37}$$

Using Taylor formula we deduce that

$$\phi_j^-(\widetilde{\lambda}_n)-\phi_j^+(\widetilde{\lambda}_n)=-\delta^{(j)}\partial_3\omega(\phi_{j-1}^+(\widetilde{\lambda}_n),|b_j|,d_j-\widetilde{\lambda}_n,|b_{j+1}|)+O({\delta^{(j)}}^2). \tag{38}$$

Therefore looking for the integer j such that $|\delta^{(j)}|$ is minimum amounts to finding j such that $|\phi_j^-(\widetilde{\lambda}_n)-\phi_j^+(\widetilde{\lambda}_n)|$ is minimum. As ϕ_k^+ is increasing and ϕ_k^- is decreasing we have for $k=1,\dots,N$

$$\left\{\begin{array}{lcl}\phi_k^+(x_n) &\le& \phi_k^+(\widetilde{\lambda}_n)\ \le\ \phi_k^+(y_n),\\ \phi_k^-(y_n) &\le& \phi_k^-(\widetilde{\lambda}_n)\ \le\ \phi_k^-(x_n),\end{array}\right. \tag{39}$$

and therefore

$$\phi_k^-(y_n) - \phi_k^+(y_n) \leq \phi_k^-(\widetilde{\lambda}_n) - \phi_k^+(\widetilde{\lambda}_n) \leq \phi_k^-(x_n) - \phi_k^+(x_n), \tag{40}$$

and

$$\left|\phi_k^-(\widetilde{\lambda}_n) - \phi_k^+(\widetilde{\lambda}_n)\right| \leq \max\left(|\phi_k^+(y_n) - \phi_k^-(y_n)|, |\phi_k^+(x_n) - \phi_k^-(x_n)|\right). \tag{41}$$

The optimal index $k_0 \in [\![1, N]\!]$ in Godunov and coworkers approach satisfies, see proposition 1,

$$\phi_{k_0-1}^+(y_n) - \phi_{k_0-1}^-(x_n) \leq 0, \qquad \text{et} \qquad \phi_{k_0}^+(y_n) - \phi_{k_0}^-(x_n) \geq 0. \tag{42}$$

We cannot deduce that $|\phi_k^+(y_n) - \phi_k^-(x_n)|$ becomes minimum for $k = k_0$ (indeed the Sturm sequences $(\phi_k^-(x_n))_k$ and $(\phi_k^-(y_n))_k$ can be closer for a given index k than they are for the index k_0 where their values cross) but may explain why we often observe that the two indices j_0 and k_0 coincide, see figure 1.
If we consider the example of the tridiagonal matrix Wilkinson of size 21, see [11], the index k_0 coincide with j_0 only for two eigenvectors, see figure 2. However, in this example as well, the index k_0 is always (except in one case) among the greatest components of the eigenvector.

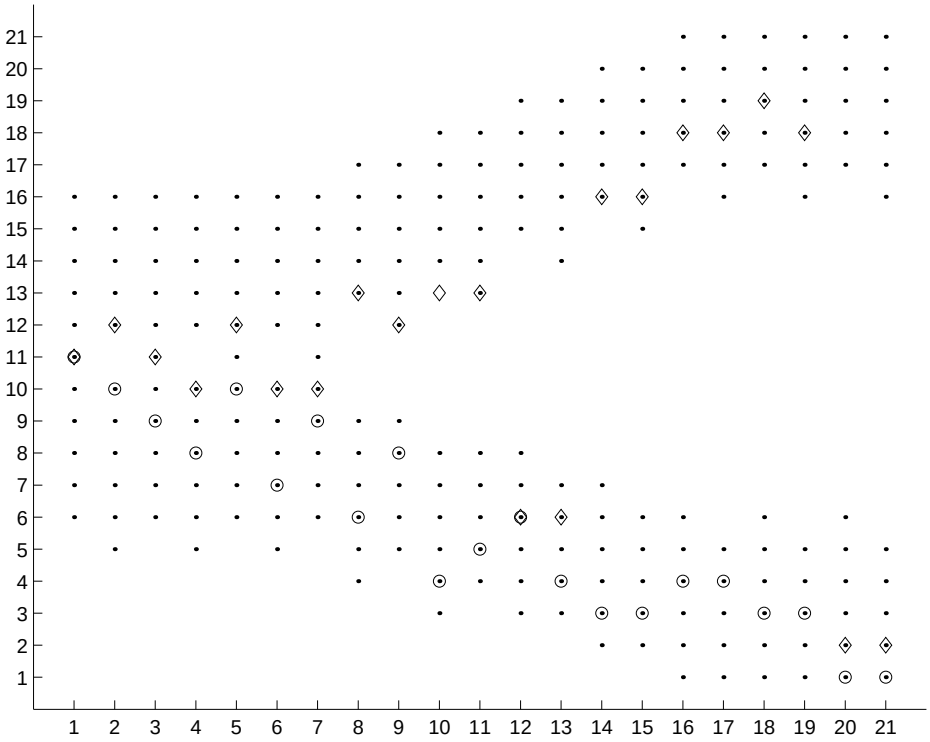


Figure 2: Position of indices j_0 ($\circ$), k_0 ($\diamond$) and j ($\cdot$) such that $|u_j|$ is among the 50% greatest components of the eigenvector u, for the tridiagonal matrix Wilkinson of size 21.

The conclusion is that in general the indices k_0 and j_0 are equal even if in some cases they can differ. However even in this later case $|\delta^{(k_0)}|$ is always small compared to the average value of the $|\delta^{(k)}|$.

4.3 Further connections

Let us consider the two-sided Sturm sequence of first kind $(P_k(\widetilde{\lambda}_n))_k$ obtained by joining the left Sturm sequence in y_n and the right Sturm sequence in x_n at the index k_0. It is composed of the following terms

$$0 = P_0^+(y_n), \ldots, P_{k_0-1}^+(y_n), P_{k_0}^-(x_n), \ldots, P_N^-(x_n) = +\infty. \tag{43}$$

This sequence is the two-sided Sturm sequence corresponding to the eigenvalue $\widetilde{\lambda}_n$ of $\widetilde{T}$. We therefore have the following recurrence for $k = 1, \ldots, N$ (we omit to distinguish the case when $P_k = 0$ or $P_k = \infty$ for simplicity), see (5),

$$(\widetilde{d}_k - \widetilde{\lambda}_n) - |b_k| P_{k-1}(\widetilde{\lambda}_n) - \frac{b_{k+1}}{P_k(\widetilde{\lambda}_n)} = 0. \tag{44}$$

Using (14) relation (44) can be written

$$\left\{ \begin{array}{lcll} (d_k - \widetilde{\lambda}_n) - |b_k| P_{k-1}(\widetilde{\lambda}_n) - \dfrac{b_{k+1}}{P_k(\widetilde{\lambda}_n)} & = & \frac{1}{2}h, & k = 1, \ldots, k_0 - 1, \\ (d_{k_0} - \widetilde{\lambda}_n) - |b_{k_0}| P_{k_0-1}(\widetilde{\lambda}_n) - \dfrac{b_{k_0+1}}{P_{k_0}(\widetilde{\lambda}_n)} & = & \frac{1}{2}h - \tau h, & \\ (d_k - \widetilde{\lambda}_n) - |b_k| P_{k-1}(\widetilde{\lambda}_n) - \dfrac{b_{k+1}}{P_k(\widetilde{\lambda}_n)} & = & -\frac{1}{2}h, & k = k_0 + 1, \ldots, N, \end{array} \right. \tag{45}$$

where $h = y_n - x_n$. An eigenvector corresponding to $\widetilde{\lambda}_n$, which is an exact eigenvalue for $\widetilde{T}$ and an approximate eigenvalue for T, is computed from the values of the two-sided Sturm sequence of first kind $(P_k(\widetilde{\lambda}_n))_{k=1,\ldots,N}$ by the recurrence: $u_1 = 1$,

$$\forall k = 2, \ldots, N, \qquad \left\{ \begin{array}{ll} u_k = 0 & \text{if } P_{k-1}(\lambda_n) = +\infty, \\ u_k = -\dfrac{b_{k-1}}{b_k} u_{k-2} & \text{if } P_{k-1}(\widetilde{\lambda}_n) = 0, \\ u_k = -\text{sign}(b_k) \dfrac{u_{k-1}}{P_{k-1}(\widetilde{\lambda}_n)} & \text{otherwise .} \end{array} \right. \tag{46}$$

¿From (45) and (46), it follows that the components u_k of the eigenvector U_n satisfy

$$\left\{ \begin{array}{lcll} (d_k - \widetilde{\lambda}_n) u_k - b_k u_{k-1} - b_{k+1} u_{k+1} & = & \dfrac{h}{2} u_k, & k = 1, \ldots, k_0 - 1, \\ (d_{k_0} - \widetilde{\lambda}_n) u_{k_0} - b_{k_0} u_{k_0-1} - b_{k_0+1} u_{k_0+1} & = & h(\dfrac{1}{2} - \tau) u_{k_0}, & \\ (d_k - \widetilde{\lambda}_n) u_k - b_k u_{k-1} - b_{k+1} u_{k+1} & = & -\dfrac{h}{2} u_k, & k = k_0 + 1, \ldots, N. \end{array} \right. \tag{47}$$

In a matrix form the linear system (47) reads

$$(T - \widetilde{\lambda}_n) U_n = \delta, \tag{48}$$

where $\delta = (\frac{h}{2} u_k, \ldots, (\frac{1}{2} - \tau) h u_{k_0}, \ldots, -\frac{h}{2} u_k)^t$.

Thus, the approach of Godunov and coworkers connecting two Sturm sequences of second kind amounts to minimizing the global residual from the linear system

$$(T - \widetilde{\lambda}_n) X = 0.$$

In Parlett and Dhillon approach we look for the index j_0 for which the residual produced by discarding equation j_0 is minimum.

4.4 Comparison of the computational cost

Let us compare the computational cost to obtain the approximate eigenvector with the two methods. We assume that the eigenvalue has been computed to machine accuracy, $|\lambda_n - \widetilde{\lambda}_n| \approx \epsilon_{mach} |\lambda_n|$.
In Godunov and coworkers method, the computation of the eigenvector requires the computation of the left Sturm sequence of first kind in y_n: $P_0^+(y_n), \ldots, P_N^+(y_n)$ and the computation of the right Sturm sequence of first kind in x_n: $P_0^-(x_n), \ldots, P_N^-(x_n)$ using relations (3) and (4). Each term $P_k^{\pm}$ necessitates one multiplication, one division and two additions to be evaluated. This requires $2N$ multiplications, $2N$ divisions et $4N$ additions. ¿From the joined Sturm sequence, the computation of the eigenvector from relation (6) requires $N-1$ divisions. The total cost to get one eigenvector with Godunov and coworkers method is $2N$ multiplications, $3N$ divisions et $4N$ additions. Moreover, the determination of the index k_0 requires N^2 sign tests.
In Parlett and Dhillon variant of the inverse iteration method, the determination of the optimal index requires the computation of the LDU and ULD decomposition for the matrix $(T - \widetilde{\lambda}_n I)$. Since the matrix is tridiagonal, the cost for each decomposition is $N-2$ multiplications, $2N-4$ divisions and $N-2$ additions. Then the solution of the triangular systems $Lv = u$ and $Uz = v$ necessitate $N-1$ multiplications and $N-1$ additions for the first one and $N-1$ multiplications, N divisions and $N-1$ additions for the second one. The total cost to get one eigenvector with Parlett and Dhillon variant of the inverse iteration method is therefore $4N$ multiplications, $3N$ divisions and $4N$ additions.

5 Conclusion

This paper has compared the improvement of two classical methods for computing eigenvectors of symmetric tridiagonal matrices. Namely, the improvement of Givens method by Godunov and coworkers, see [3] and the improvement of the inverse iteration method by Parlett and Dhillon, see [8]. Godunov and coworkers improvement of Givens method ensures that the Sturm sequence used to compute the eigenvector is two-sided which guarantees a stable and accurate computation. This is not always the case with the standard Givens method. Moreover the extra-cost for this modification of Givens method remains low. Parlett's and Dhillon improvement of the inverse iteration method consists of establishing the best initial vector to start the iterations. This guarantees the inverse iteration method to converge to the sought-after eigenvector. With the standard inverse iteration method, convergence is unlikely to occur if the chosen initial vector is orthogonal to eigenvector. Although it is interesting to be sure to have an initial vector that guarantees convergence of inverse iteration, from a practical point of view, the overall cost due to the determination of the best initial vector is usually dissuasive. We can quote Peters and Wilkinson, see [10, p. 360]:

> the ordinary process of inverse iteration will almost always succeed in one iteration; if it does not do so one has only to restart with an initial vector orthogonal to the first. This process can be continued until one reaches an initial vector which gives success in one iteration. It is rare for the first vector to fail and the average number of iterations is unlikely to be as high as 1.2.

Thus, it is, perhaps, more economical to use 2 iterations with the standard process than one iteration with the best initial vector.

References

[1] Anderson, E. and all. *Lapack users' guide.* SIAM, Philadelphia, 1995.

[2] Godunov, S.K. Adaptation de la methode de Sturm pour le calcul des vecteurs propres des matrices Jacobiennes. In *Analyse mathématique et applications, Contrib. Honneur Jacques-Louis Lions.* Gauthier-Villard, Paris, 1988.

[3] Godunov, S.K. and Antonov, A.G. and Kirilyuk, O.P. and Kostin, V.I. *Guaranteed accuracy in numerical linear algebra. Updated and rev. transl. of the original work 'The guaranteed precision of linear equations solutions in Euclidean spaces', publ. by Nauka 1992.* Mathematics and its Applications, Kluwer Academic Publishers, 1993.

[4] Godunov, S.K., Kostin, V.I. and Mitchenko, A.D. . Computation of an eigenvector of a symmetric tridiagonal matrix. *Sib. Math. J.*, 26:684–696, 1985.

[5] Ipsen, I.C.F. A history of inverse iteration. In B. Huppert and Schneider H., editors, *Helmut Wielandt, mathematische werke, vol 2, matrix theory and analysis.* Walter de Guyter, Berlin, 1995.

[6] Ipsen, I.C.F. Computing an eigenvector with inverse iteration. *SIAM Rev.*, 39:254–291, 1997.

[7] Jessup, E.R. and Ipsen, I.C.F. Improving the accuracy of inverse iteration. *SIAM J. Sci. Stat. Comput.*, 13:550–572, 1992.

[8] Parlett, B. N. and Dhillon, I. S. Fernando's solution to Wilkinson's problem: An application of double factorization. *Linear Algebra Appl.*, 267:247–279, 1997.

[9] Parlett, B.N. *The symmetric eigenvalue problem.* Prentice-Hall Series in Computational Mathematics. Englewood Cliffs, New Jersey: Prentice-Hall, 1980.

[10] Peters, G. and Wilkinson, J.H. Inverse iterations, ill-conditioned equations and Newton's method. *SIAM Rev.*, 21:339–360, 1979.

[11] Wilkinson, J.H. *The algebraic eigenvalue problem.* Monographs on Numerical Analysis. Oxford Clarendon Press, 1965.