Computing, Suppl. 6, 59–68 (1988)

Computing

Errorbounds for Quadratic Systems of Nonlinear Equations Using the Precise Scalar Product

G. Alefeld, Karlsruhe

Abstract – Zusammenfassung

Errorbounds for Quadratic Systems of Nonlinear Equations Using the Precise Scalar Product. For nonlinear systems of quadratic equations we show how the precise scalar product can be used in order to compute and to improve inclusions for a solution. Our main interest is the special case which comes from the generalized eigenvalue problem.

Fehlerschranken für quadratische Systeme nichtlinearer Gleichungen unter Verwendung des genauen Skalarprodukts. Für Gleichungssysteme des angegebenen Typs berechnen wir Einschließungen für eine Lösung und verbessern diese unter Verwendung des genauen Skalarprodukts. Das Hauptinteresse gilt dem Spezialfall, welcher dem verallgemeinerten Eigenwertproblem entspricht.

1. Introduction

We consider the *quadratic equation*

$$y = u + Sy + Ty^2 \tag{1}$$

where $u=(u_i)$ is a real vector from $\mathbb{R}^n$, $S=(s_{ij})$ is a real (n, n) matrix and $T=(t_{ijk})$ is a real bilinear operator from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}^n$ defined by $Txy=\left(\sum_{j=1}^{n}\sum_{k=1}^{n} t_{ijk}\, x_k\, y_j\right)$ for $x=(x_i)$, $y=(y_i)\in\mathbb{R}^n$. Ty^2 is defined to be Tyy. The unknown vector is $y=(y_i)$. In general a solution of (1) can only be computed approximately by an iterative method. Furthermore if such a method is performed on a computer one has to take into account rounding errors. We show how this can be done for the quadratic (1) using the precise scalar product.

Our main interest is a special case of (1), namely the generalized eigenvalue problem. Note that quadratic equations have already been considered in [1]. The special case of an eigenvalue problem has been discussed in [2] and [3].

2. The Generalized Eigenvalue Problem as a Quadratic Equation of the Form (1)

We consider the generalized matrix eigenvalue problem

$$Ax = \lambda Bx \tag{2}$$

where A and B are real (n, n) matrices. We assume in this paper that B is nonsingular. Under practical aspects this is not very restrictive since in real life problems B is usually even symmetric and positive definite.

Assume now that after using some well known algorithm for computing eigenpairs (see [9], for example) we have given a real approximation λ for a simple real eigenvalue and an approximation x for the corresponding eigenvector. For the exact eigenpair $(\lambda+\tilde{\mu}, x+\tilde{y})$ the equation

$$A(x+\tilde{y})=(\lambda+\tilde{\mu})\,B(x+\tilde{y}) \tag{3}$$

holds. Let

$$\|x\|_\infty=|x_s|>0 \tag{4}$$

where s is some index for which the infinity norm is taken on. Since $x+\tilde{y}$ is not unique we normalize $x+\tilde{y}$ by setting

$$\tilde{y}_s=0 \tag{5}$$

where $\tilde{y}=(\tilde{y}_i)$. Equation (3) can be rewritten as

$$(A-\lambda B)\,\tilde{y}-\tilde{\mu}Bx=(\lambda B-A)\,x+\tilde{\mu}B\tilde{y}. \tag{6}$$

Defining the components y_i of the vector $y=(y_i)\in\mathbb{R}^n$ by

$$y_i=\begin{cases}\tilde{y}_i, & i\neq s\\ \tilde{\mu}, & i=s\end{cases}$$

the last equation can be written as

$$Cy=r+B(y_s\,\tilde{y}) \tag{7}$$

where

$$r=\lambda Bx-Ax \tag{8}$$

and where C is identical to $A-\lambda B$ with the exception of the s-th column which is replaced by $-Bx$. See [5].

If B is nonsingular — this was our general assumption — and for sufficiently good approximations λ and x it can be shown that the matrix C is nonsingular (see [8], for example). Assume now that this is the case and let L be some approximation to the inverse of C. Then (7) can be rewritten as

$$y=Lr+(I-LC)\,y+L(B(y_s\,\tilde{y})). \tag{9}$$

This equation has the form (1) where $u=Lr$ and $S=I-LC$. The bilinear operator T is defined by $Ty^2=L(By_s\,\tilde{y}))$. We omit to express the elements t_{ijk} of T explicitly by the elements of the matrices L and B because this is not important in the sequel. Note, however, that the last term in (9) could also be written as

$$L(B(y_s\,\tilde{y}))=(LB)(y_s\,\tilde{y})=y_s(LB)\,\tilde{y}$$

since the associative law holds and since y_s is a scalar. The reason why we use the first one of these equivalent expressions becomes clear in Chapters 3 and 4.

3. Computing an Enclosing Interval Vector

We now try to compute an interval vector $[y]=([y]_i)$ for which

$$u+Sy+Ty^2\in[y] \quad \text{for all } y\in[y]. \tag{10}$$

Then by the Brouwer fixed-point theorem the equation (1) has at least one solution in $[y]$. We try to find $[y]$ in the form

$$[y]=[-\beta,\beta]\,e \tag{11}$$

where $\beta>0$ and $e=(1, 1, \ldots, 1)^T\in\mathbb{R}^n$ (This approach is motivated by the fact that $y=0$ is nearly a solution of (1) if u is "small"). By inclusion monotonicity (see [4], Chapter 1, Theorem 5) we have for $y\in[y]$

$$u+Sy+Ty^2\in u+S[y]+T[y]^2.$$

Hence

$$[w]:=u+S[y]+T[y]^2\subseteq[y] \tag{12}$$

is sufficient for (10).

Following the laws of interval arithmetic we have

$$S[y]=\left(\sum_{j=1}^{n} s_{ij}[-\beta,\beta]\right)=[-\beta,\beta]\left(\sum_{j=1}^{n}|s_{ij}|\right)$$

and

$$\begin{aligned}T[y]^2&=\left(\sum_{j=1}^{n}\left(\sum_{k=1}^{n} t_{ijk}\,y_k\right)y_j\right)\\&=\left(\sum_{j=1}^{n}\left(\sum_{k=1}^{n} t_{ijk}[-\beta,\beta]\right)[-\beta,\beta]\right)\\&=[-\beta^2,\beta^2]\left(\sum_{j=1}^{n}\left(\sum_{k=1}^{n}|t_{ijk}|\right)\right)\end{aligned}$$

and therefore

$$u+Sy+Ty^2=u+[-\beta,\beta]|S|e+[-\beta^2,\beta^2]|T|e^2$$

where $|S|=(|S_{ij}|)$ and $|T|=(|t_{ijk}|)$.

(12) holds iff

$$|m[w]-m[y]|+\tfrac{1}{2}d[w]\leq\tfrac{1}{2}d[y] \tag{13}$$

where m denotes the center, d the diameter and $|\cdot|$ the absolute value of an interval vector (see [4], Chapter 10).

We have

$$\begin{aligned}&m[w]=u, \quad m[y]=0,\\&d[w]=2\beta|S|e+2\beta^2|T|e^2,\\&d[y]=2\beta e.\end{aligned}$$

Hence (13) holds iff

$$|u|+\beta|S|e+\beta^2|T|e\leq\beta e. \tag{14}$$

Defining

$$\rho = \|u\|_\infty, \tag{15}$$

$$\kappa = \|S\|_\infty, \tag{16}$$

$$\ell = \max_i \left\{ \sum_{j=1}^{n} \sum_{k=1}^{n} |t_{ijk}| \right\} \tag{17}$$

the last vector inequality holds if

$$\rho + \beta\kappa + \beta^2 \ell \leq \beta.$$

(For the bilinear operator $T=(t_{ijk})$ a norm can be defined by

$$\|T\|_\infty = \sup_{\|x\|_\infty = \|y\|_\infty = 1} \|Txy\|_\infty.$$

It is easy to prove that $\|T\|_\infty \leq \ell$ where ℓ is defined by (17). However, in general equality does not hold. See [7], for example).

Hence we have the following result.

Theorem 1. *Let ρ, κ, ℓ be defined by* (15)–(17). *Assume that $\kappa < 1$, $(1-\kappa)^2 - 4\rho\ell \geq 0$ and let*

$$\beta_{1/2} = \frac{1 - \kappa \mp ((1-\kappa)^2 - 4\rho\ell)^{1/2}}{2\ell} \tag{18}$$

be the solutions of the quadratic equation

$$\beta^2 \ell + (\kappa - 1)\beta + \rho = 0. \tag{19}$$

If $\beta \in [\beta_1, \beta_2]$ then the equation (1) *has at least one solution y^* in the interval vector* (11). □

Please note that Theorem 1 gives only sufficient conditions for (10).

In the special case of the equation (9) the preceding theorem holds for

$$\rho = \|Lr\|_\infty \tag{20}$$

$$\kappa = \|I - LC\|_\infty \tag{21}$$

$$\ell = \| |L| |B| \|_\infty. \tag{22}$$

If we choose $L := C^{-1}$ in (9) then the equation (9) reads

$$y = C^{-1}r + C^{-1}(B(y_s \tilde{y})) \tag{23}$$

and the condition (10) can be written as

$$C^{-1}(r + B([y]_s [\tilde{y}]) \subseteq [y] \tag{24}$$

(Note that for a real matrix M and interval vectors $[x]$ and $[y]$ it holds that $M([x]+[y]) = M[x] + M[y]$).

If we would have written the equation (9) in the form

$$y = C^{-1}r + (C^{-1}B)(y_s \tilde{y})$$

then we could not rewrite the condition (10) in the form (24) since in the set of interval vectors and interval matrices the associative law does not hold in general.

4. Improving an Enclosing Interval Vector Iteratively

After applying the preceding theorem one can try to improve the computed inclusion using the following iteration method:

$$\left.\begin{aligned} [y]^0 &= [-\beta, \beta]e \\ [y]^{k+1} &= g([y]^k), \quad k=0, 1, 2, \ldots \end{aligned}\right\} \tag{25}$$

where

$$g[y] = u + S[y] + T[y]^2. \tag{26}$$

This iteration method computes a sequence $\{[y]^k\}_{k=0}^{\infty}$ of interval vectors for which the following result holds.

Theorem 2. *Let* $\kappa < 1$, $(\kappa-1)^2 - 4\rho\ell > 0$ *and let* β_1, β_2 *be defined by* (18). *Then if*

$$\beta_1 \leq \beta < \frac{\beta_1 + \beta_2}{2} \tag{27}$$

it holds that

a) $y^* \in [y]^k$

and

b) $\lim\limits_{k \to \infty} [y]^k = y^*$.

Furthermore y^* *is unique in* $[y]^0$. □

We omit the details of a proof.

Please note again that the result also holds for

$$g([y]) = Lr + (I - LC)[y] + L(B([y]_s [\tilde{y}])) \tag{28}$$

if ρ, κ and ℓ are defined by (20)–(22).

In the special case $L := C^{-1}$ we can rewrite (28) as

$$\begin{aligned} g([y]) &= C^{-1}r + C^{-1}(B[y]_s [\tilde{y}]) \\ &= C^{-1}(r + B([y]_s [\tilde{y}])). \end{aligned} \tag{29}$$

At first glance it does not make much sense to set $L := C^{-1}$ since normally C^{-1} can not be represented exactly on a computer. However, there are some good reasons why this makes sense. In the first place the term $(I - LC)[y]$ becomes zero, which means that it has not to be computed in every iteration step. Furthermore the fact that C^{-1} can not be represented exactly has not to bother us. It is well known that there exist algorithms using the precise scalar product which deliver an optimal inclusion of C^{-1}. Hence we only have to care for the accurate computation of $r + B([y]_s [\tilde{y}])$ in each iteration step. How this can be done is described in Chapter 6 for the general quadratic (1).

5. A Heuristic Procedure for Computing an Enclosing Interval Vector

Assume that the assumptions of Theorem 1 hold such that, using this theorem, we can compute an including interval vector for a solution of (1). From a practical point of view one is interested in a very good inclusion. Therefore the choice $\beta := \beta_1$ suggests itself. However, as the proof of Theorem 1 shows this choice is only sufficient for $g([y]^0) \subseteq [y]^0$. One could try to find $[y]^0$ with a smaller diameter such that $g([y]^0) \subseteq [y]^0$ by the following iteration method

$$\begin{aligned} &\text{Set } [z]^0 := 0; \\ &[z]^{k+1} := \operatorname{conv}\{u + S[z]^k + T([z]^k)^2, [z]^k\} \\ &\qquad \text{until } [z]^{k+1} \subseteq [z]^k; \\ &[y]^0 := [z]^k; \end{aligned} \tag{30}$$

where $\operatorname{conv}\{\cdot, \cdot\}$ denotes the convex hull of two interval vectors.

By inclusion monotonicity of interval arithmetic we have $[z]^k \subseteq [-\beta_1, \beta_1]e$ for all k provided Theorem 1 applies.

Practical experience shows that applying Theorem 1 and the iteration method (30), respectively, nearly needs the same computing time. Using (30), however, has the advantage that it may still work if Theorem 1 is not applicable.

Of course these remarks hold also in the special case of the equation (9) which comes from the generalized eigenvalue problem (2).

6. Using the Precise Scalar Product

In order to get inclusions with small diameters on a computer one should try to keep the rounding errors as small as possible when performing the iteration method (25).

Our quadratic equation is an example for which the so-called precise scalar product can be applied (see [6]) to achieve this.

We show that the components of the right-hand side of the equation

$$y = u + Sy + Ty^2$$

can be computed by a single scalar product.

The only problem which has to be explained is how to reduce Ty^2 to one scalar product. This can be done in the following manner:

We have by definition

$$Ty^2 = \left(\sum_{j=1}^{n} \left(\sum_{k=1}^{n} t_{ijk}\, y_k \right) y_j \right).$$

Define now

$$r_{ijk}=f\ell(t_{ijk}\,y_k)$$

where $f\ell(\cdot)$ denotes the floating point multiplication.

Furthermore let

$$p_{ijk}=\begin{pmatrix}t_{ijk}\\ t_{ijk}\end{pmatrix}^T\cdot\begin{pmatrix}y_k\\ -1\end{pmatrix}$$

be the precise scalar product. Then it holds exactly

$$t_{ijk}\,y_k=r_{ijk}+p_{ijk}.$$

Therefore we can write

$$\begin{aligned}Ty^2&=\left(\sum_{j=1}^{n}\left(\sum_{k=1}^{n}(r_{ijk}+p_{ijk})\right)y_j\right)\\&=\left(\sum_{j=1}^{n}\left(\sum_{k=1}^{n}r_{ijk}\,y_j+\sum_{k=1}^{n}p_{ijk}\,y_j\right)\right)\end{aligned}$$

and our problem is solved. Note that when performing the iteration method (25) we have interval vectors on the right-hand side. In this case a similar idea can be applied.

If we specialize our quadratic equation (1) to the eigenvalue problem (2) then u, S and T are in general not exactly representable on the machine and the situation becomes much more complicated. We do not discuss the general case but only the choice $L:=C^{-1}$. As we have seen at the end of Chapter 4 we have in this case

$$g[y]=C^{-1}(r+B([y]_s\,[\tilde{y}])).$$

Using the same ideas as before the components of the expression $r+B([y]_s\,[\tilde{y}])$ can be computed each by one scalar product. We do not repeat the details.

We close with a final comment. Since it is clear that in general we can not compute C^{-1} exactly it seems to be more favourable to perform the iteration method

$$\begin{aligned}[y]^{k+1}&=\mathrm{IGA}(C,r+B([y]_s^k\,[\tilde{y}]^k)),\\ &k=0,1,2,\ldots\end{aligned}$$

where $\mathrm{IGA}(\cdot,\cdot)$ denotes the result of the Gaussian algorithm. Note, however, that for a real (n, n) matrix M and an interval vector $[b]$ one can prove that

$$M^{-1}[b]\subseteq\mathrm{IGA}(M,[b]).$$

See [4], Chapter 15. If one assumes that this also holds if rounding errors are taken into account — this has not been proved — then it is clear that the iteration method (25) with $g[y]$ defined by (29) has to be preferred in order to get narrow inclusions.

7. Numerical Examples

Consider the symmetric matrices

$$A=\begin{bmatrix} 12 & 1 & -1 & 2 & 1 \\ 1 & 14 & 1 & -1 & 1 \\ -1 & 1 & 16 & -1 & 1 \\ 2 & -1 & -1 & 12 & -1 \\ 1 & 1 & 1 & -1 & 11 \end{bmatrix}$$

and

$$B=\begin{bmatrix} 10 & 2 & 3 & 1 & 1 \\ 2 & 12 & 1 & 2 & 1 \\ 3 & 1 & 11 & 1 & -1 \\ 1 & 2 & 1 & 9 & 1 \\ 1 & 1 & -1 & 1 & 15 \end{bmatrix}$$

from [9], p. 313.

As λ we choose the first six digits of an approximation to an eigenvalue given in [9]:

$$\lambda=0.231060\times 10^{+1}.$$

Analogously we choose the corresponding eigenvector approximation

$$x=\begin{pmatrix} -0.204587 \\ 0.931721\times 10^{-1} \\ 0.240022507111 \\ -0.166395 \\ 0.630418\times 10^{-1} \end{pmatrix}.$$

(The third component which has the largest absolute value is exactly the approximation given in [9]. This component is not changed by our algorithm).

Setting $C=L^{-1}$ we get for β_1 from Theorem 1

$$\beta_1=0.432158741876\,3\times 10^{-5}.$$

After two iteration steps of method (25) with g defined by (29) we get the following inclusions for the eigenvalue $\lambda+\tilde{\mu}$ and for the components of the corresponding eigenvector $x+\tilde{y}$:

$$\lambda+\mu\in[0.231060432134_9^8\times 10^1]$$

$$x+\tilde{y}\in\begin{pmatrix} [-0.204586718184_4^5] \\ [0.931720977435_5^4\times 10^{-1}] \\ [0.240022507111\,0] \\ [-0.166395354479_7^8] \\ [0.630417653106_8^7\times 10^{-1}] \end{pmatrix}.$$

The computation of β_1 needs 6063 ms. The overall computing time is 10022 ms. Using (30) we have $[z]^{k+1} \subseteq [z]^k$ after 4 steps. This needs 7630 ms. After three iteration steps of (25) with (29) we get the same final inclusions. The overall computing time is in this case 10012 ms.

Practical experience shows that for good approximations λ and x both approaches need nearly the same computing time. If, however, Theorem 1 is not applicable then the second approach still works if the approximations are not too bad.

References

[1] Alefeld, G.: Componentwise Inclusion and Exclusion Sets for Solutions of Quadratic Equations in Finite Dimensional Spaces, Numer. Math. *48*, 391–416 (1986)

[2] Alefeld, G.: Berechenbare Fehlerschranken für ein Eigenpaar unter Einschluß von Rundungsfehlern bei Verwendung des genauen Skalarprodukts. Z. angew. Math. Mech. *67*, 3, 145 – 152 (1987).

[3] Alefeld, G.: Berechenbare Fehlerschranken für ein Eigenpaar beim verallgemeinerten Eigenwertproblem. Z. angew. Math. Mech. *68*, 3, 181 ff. (1988).

[4] Alefeld, G., Herzberger, J.: Introduction to Interval Computations. Academic Press, New York, 1983 (ArNu).

[5] Dongarra, J.J., Moler, C.B., Wilkinson, J.: Improving the accuracy of computed eigenvalues and eigenvectors. SIAM J. Numer. Anal. *20*, 23 – 45 (1983).

[6] Kulisch, U., Miranker, W. (Eds.): A New Approach to Scientific Computation. Academic Press, 1983.

[7] Platzöder, L.: Einige Beiträge über die Existenz von Lösungen nichtlinearer Gleichungssysteme und Verfahren zu ihrer Berechnung. Dissertation. Technische Universität Berlin (1981).

[8] Symm, H.J., Wilkinson, J.H.: Realistic error bounds for a simple eigenvalue and its associated eigenvector. Numer. Math. *35*, 113 – 126 (1980).

[9] Wilkinson, J.H., Reinsch, C.: Handbook for Automatic Computation. Volume 2: Linear algebra. Springer Verlag (1971).

Prof. Dr. G. Alefeld
Institut für Angewandte Mathematik
Universität Karlsruhe
Kaiserstrasse 12
D-7500 Karlsruhe 1
Federal Republic of Germany