A RAPID ITERATIVE DIAGONALIZATION METHOD FOR LARGE MATRICES

MIKLÓS. I. BÁN

Institute of Physical Chemistry, Attila József University

P.O.Box 105., H–6701 Szeged, Hungary

(Received November 7, 1989)

AN ACCELERATED ITERATIVE MATRIX DIAGONALIZATION TECHNIQUE REDUCING THE COMPUTATIONAL WORK IS DESCRIBED. THE METHOD IS USEFUL WHEN THE RATE OF CONVERGENCE IS NOTORIOUSLY SLOW AND THE ITERATIVE PROCEDURE CAN NOT BE AVOIDED.

1. Introduction

Iterative procedures used to determine the eigenvectors (and the spectrum) of matrices yield the eigenvector belonging to the eigenvalue largest in absolute sense . The overall rate of convergence depends on the structure of the spectrum very strongly and the usual procedures require in most cases too many iterations to be of practical value. In order to determine the rest of eigenvectors (and eigenvalues) several methods reducing the rank of the matrix, or projective elimination techniques are generally used.

The diagonalization of large matrices is an acute problem extending to many fields of computational sciences (molecular physics, quantum chemistry, *etc.*). A typical computational task in recent molecular physics and quantum chemistry is the determination of some (*e.g.* generally the largest) eigenvalues and eigenvectors of the Hamiltonian at the post–SCF level.

In this paper an effective acceleration procedure reducing substantially the number of iteration steps necessary to reach convergence will be described. The procedure is especially useful in cases where the separation of eigenvalues is small.

2. Method

Let us consider a positive (semi–)definite matrix A with real elements and denote the ordered set of its eigenvalues by λ_j:

$$\lambda_j := (\lambda_1 > \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n) \in R, \qquad j=1,..,n \tag{1}$$

and its eigenvectors by e_i, where R is the assembly of real numbers.

As is well known, the existence of the following limit values[1] can be verified[2][1–4]:

$$\lim_{k\to\infty} \frac{1}{\lambda_1^k} A^k {}^0u = \alpha_1 e_1, \qquad {}^0u = {}^0\alpha_1 e_1 + {}^0\alpha_2 e_2 + \ldots + {}^0\alpha_n e_n, \tag{2}$$

$$\min_{(u)} Q = \frac{\bar{u}\ A\ u}{\bar{u}\ u} = \lambda_n, \tag{3}$$

and

$$\max_{(u)} Q = \frac{\bar{u}\ A\ u}{\bar{u}\ u} = \lambda_1, \tag{4}$$

where k is the sequential number of the iteration step, ${}^0\alpha_i$ is the 0^{th} coefficient vector and the vector[3] u in the formulas (3) and (4) run over the whole configuration space, while Q is the so called Rayleigh quotient [5]. The classic iterative procedure, the "power method"[5] consists of consecutive multiplications by a trial vector. In the following section a modification of this method will be described.

Let us examine the following schemes (in the left and right columns the steps of the power method and the modified procedure respectively, are compared):

[1]The superscripts on the left and right of symbols refer to the sequential number of iteration and the exponents, *resp.*

[2]Theorem of von MISES.

[3]$\bar{u}$ is the transpose of u.

$$
\begin{array}{lll}
{}^{1}u = A\,{}^{0}u, & {}^{1}v = {}^{1}u, & \\
{}^{2}u = A\,{}^{1}u, & {}^{2}v = {}^{2}u, & \\
\vdots & \vdots & \\
\vdots & \vdots & \\
{}^{s}u = A\,{}^{s-1}u, & {}^{s}v = {}^{s}u, & {}^{s}w = \hat{B}_s\,{}^{s}v, \\
 & {}^{s+1}v = A\,{}^{s}w, & \\
\vdots & \vdots & \\
\vdots & \vdots & \\
{}^{s+r}u = A\,{}^{s+r-1}u, & {}^{s+r}v = A\,{}^{s+r-1}v, & {}^{s+r}w = \hat{B}_{s+r}\,{}^{s+r}v.
\end{array}
$$

The operator $\hat{B}_s$ performs a non–linear transformation of the trial vector. This step is embedded into the basic procedure and it will repeatedly be applied by using the actual vectors stored in the given cycle.

Now we try to construct the operator $\hat{B}_s$ in such a manner that the vector ${}^{s+1}v = A^s w$ approximates a vector ${}^{k}u$ of the basic procedure where $k>s+1$ (or even more favourably: $k\gg s+1$).

3. Details of the procedure

To define the operator $\hat{B}_s$ we express ${}^{0}u$, ${}^{k}u$, *etc.* by the eigenvectors $\{e_i\}$ in the space used for the computations[4]:

$$
{}^{0}u = \alpha_1 e_1 + \alpha_2 e_2 + \ldots + \alpha_n e_n, \tag{5}
$$

$$
{}^{0}u_i = \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_n x_{in}, \tag{6}
$$

$$
{}^{k}u = \alpha_1 \lambda_1^k e_1 + \alpha_2 \lambda_2^k e_2 + \ldots + \alpha_n \lambda_n^k e_n, \tag{7}
$$

$$
{}^{k}u_i = \alpha_1 \lambda_1^k x_{i1} + \alpha_2 \lambda_2^k x_{i2} + \ldots + \alpha_n \lambda_n^k x_{in}. \tag{8}
$$

[4]The coordinates of the eigenvectors in this system are: $\{e_j\} = \{x_{ij}\}$.

By constructing the sequence (2) we obtain:

$$^{k}\tilde{u}_i = \alpha_1 x_{i1} + \alpha_2\left[\frac{\lambda_2}{\lambda_1}\right]^k x_{i2} + \ldots + \alpha_n\left[\frac{\lambda_n}{\lambda_1}\right]^k x_{in} \approx \alpha_1 x_{i1} + M\left[\frac{\lambda_2}{\lambda_1}\right]^k, \tag{9}$$

where M is constant.

Each component vanishes separately according to a function

$$f_{ij}(k) = \gamma_{ij} a_j^k, \tag{10}$$

where $a_j = (\lambda_j/\lambda_1)$ and the meaning of γ_{ij} follows from *eqn.*(9). Regarding (10) and (1) the rate of convergence is apparently determined by the second largest eigenvalue.

Now combining (9) and(10)

$$^{k}\tilde{u}_i = \alpha_1 x_{i1} + \sum_{j=2}^{n} \gamma_{ij} a_j^k \tag{11}$$

is obtained.

The next goal is to determine a single function

$$\Theta_i(t) = \xi_i e^{-\nu_i t}, \quad 0 \le t < \infty \tag{12}$$

which approximates satisfactorily the function[5]

$$F_i(k) = \sum_{j=2}^{n} f_{ij}(k). \tag{13}$$

[5]Function (13) having an integer argument is not a continuous function in contrast to (12) which has a real argument.

Substituting $f_{ij}(k)$ by

$$\varphi_{ij}(t) = \gamma_{ij} e^{-\beta_{ij} t}, \tag{14}$$

then for integer values of t the equality

$$\overset{k}{\tilde{u}}_i = \alpha_1 x_{i1} + \sum_{j=2}^{n} f_{ij}(k) = \alpha_1 x_{i1} + \sum_{j=2}^{n} \varphi_{ij}(t)\Big|_{t=k} \tag{15}$$

will hold. It can readily be shown that as an exact substitute for (13) no function of type (12) is expected to be found. To realize this, let us consider the difference function $\omega_i(t)$ and its derivatives $\omega_i^{(r)}(t)$:

$$\omega_i(t) = \Theta_i(t) - \sum_{j=2}^{n} \varphi_{ij}(t) = \xi_i e^{-\nu_i t} - \sum_{j=2}^{n} \gamma_{ij} e^{-\beta_{ij} t} =$$

$$= \sum_{j=2}^{n} \left[\frac{1}{n-1} \xi_i e^{-\nu_i t} - \gamma_{ij} e^{-\beta_{ij} t} \right], \tag{16}$$

$$\omega_i^{(r)}(t) = (-1)^r \xi_i \nu_i^r e^{-\nu_i t} - \sum_{j=2}^{n} (-1)^r \gamma_{ij} \beta_{ij}^r e^{-\beta_{ij} t} =$$

$$\sum_{j=2}^{n} (-1)^r \gamma_i^r \left\{ \frac{1}{n-1} \xi_i e^{-\nu_i t} - \gamma_{ij} \left[\frac{\beta_{ij}}{\nu_i} \right]^r e^{-\beta_{ij} t} \right\}. \tag{17}$$

From (16) and (17) follows that at a given point, in general, $\omega_i(t)$ and its derivatives can not simultaneously be zero unless one of the following conditions is

fulfilled[6]:

$$\frac{\beta_{ij}}{\nu_i} = 1, \quad j = 2,\ldots,n,$$

$$\beta_{i1} \neq 0, \quad \text{but } \beta_{ij} = 0 \text{ if } j \neq 1,$$

$$\gamma_{i1} \neq 0, \quad \text{but } \gamma_{ij} = 0 \text{ if } j \neq 1.$$

Now for judging the possibility of an approximation, let us examine the behaviour of (15) describing the changes of the components of the iterated vectors in the course of iteration. We represent first (15) in the form

$$\overset{k}{\tilde{u}}_i = \alpha_1 x_{i1} + \sum_{j=2}^{n} \gamma_{ij} e^{-\beta_{ij} t} = \alpha_1 x_{i1} + \left\{ \sum_{j=2}^{n} \gamma_{ij} e^{-z_{ij} t} \right\} e^{-\xi_i t}. \tag{18}$$

The expression in the parentheses can be transcribed as follows:

$$\sum_{j=2}^{n} \gamma_{ij} e^{-z_{ij} t} = \sum_{j=2}^{n} \gamma_{ij} e^{-\delta_i(t) t}, \tag{19}$$

where

$$\delta_i(t) = -\frac{\ln\left[\sum_{j=2}^{n} \gamma_{ij} e^{-z_{ij} t} \Big/ \sum_{j=2}^{n} \gamma_{ij} \right]}{t} =$$

[6]All these cases lie outside the region of practical importance.

$$= -\frac{\ln \sum_{j=2}^{n} d_{ij} e^{-z_{ij} t}}{t}, \tag{20}$$

$$d_{ij} = \frac{\gamma_{ij}}{\sum_{l=2}^{n} \gamma_{il}} \tag{21}$$

and

$$\sum_{j=2}^{n} d_{ij} = 1. \tag{22}$$

The sum in the argument of the logarithm is increasing or decreasing strictly monotonously depending upon the signs of z_{ij}. In the first case ($z_{ij} < 0$) the range of the argument function is between 1 and ∞ if $0 \leq t < \infty$ and the relations[7]

$$\ln \left[\sum_{j=2}^{n} d_{ij} e^{-z_{ij} t} \right] < \ln \left[\sum_{j=2}^{n} d_{ij} e^{-z_{i,\min} t} \right] \equiv$$

$$\equiv \ln(e^{-z_{i,\min} t}) = |z_{i,\max}|^{t} \tag{23}$$

hold.

If the second case ($z_{ij} > 0$) occurs, the argument function ranges between 1 and 0 if $0 \leq t < \infty$ and the relations

$$\ln \left[\sum_{j=2}^{n} d_{ij} e^{-z_{ij} t} \right] < \ln \left[\sum_{j=2}^{n} d_{ij} e^{-z_{i,\min} t} \right] \equiv$$

[7] $z_{i,\min}$ and $z_{i,\max}$ denote the minimum and maximum elements in the set z_{ij}.

$$\equiv \ln(e^{-z_{i,\min}t}) = |z_{i,\min}|^t \tag{24}$$

will be satisfied.

For $z_{ij} < 0$

$$\lim_{t\to\infty} \delta_i(t) = -\frac{\ln e^{|z_{i,\max}|^t}}{t} = -|z_{i,\max}| \tag{25}$$

and for $z_{ij} > 0$

$$\lim_{t\to\infty} \delta_i(t) = -\frac{\ln e^{|z_{i,\min}|^t}}{t} = -|z_{i,\min}| \tag{26}$$

will be valid.

It seems that in any case $\delta_i(t)$ becomes constant in the course of iteration. By using these last results we get an approximation for the exact formula[8]:

$$k\tilde{u}_i = \alpha_1 x_{i1} + \sum_{j=2}^{n} \gamma_{ij} e^{-\delta_i(t)t}\, e^{-\zeta_i t} \approx \alpha_1 x_{i1} + \sum_{j=2}^{n} \gamma_{ij} e^{-(z_{i,\max/\min} + \zeta_i)t}. \tag{27}$$

We expect this approximating formula (27) to be satisfactory for practical purposes only if it is used after having reached the required accuracy.

The computations proceed as follows:

i) Because of the possibility of the presence of negative eigenvalues (in contrast

[8]The subscript max/min means either maximum or minimum in conformity with the former remarks.

to our assumption about the positive semidefinite nature of A), if the accuracy is below a predetermined limit, each second vectors start to be stored. For computational efficiency it is practical and enough to use but three vectors.

ii) The formula (27) will be used in the form

$$^{k}\tilde{u}_i \approx S_i(t) = \Theta_i(t) + K_i. \tag{28}$$

To have a descending sequence, if necessary, the i^{th} components of the vectors used are to be transformed. By supposing that the following equation system can be set up for each component of the trial vector separately, it will be solved for all the unknown parameters[9]:

$$\xi_i e^{-\nu_i k_1} + K_i = {}^{k_1}b_i, \tag{29}$$

$$\xi_i e^{-\nu_i k_2} + K_i = {}^{k_2}b_i, \tag{30}$$

$$\xi_i e^{-\nu_i k_3} + K_i = {}^{k_3}b_i. \tag{31}$$

k_1; $k_2 = k_1 + 2$; $k_3 = k_2 + 2$ are the serial numbers of iteration, ${}^{k_j}b_i$ are the i^{th} components of the iterated vectors, and ξ_i, ν_i and K_i are the unknown parameters to be determined. To get ν_i one has to solve a second order equation, however, ξ_i and K_i are in linear dependence.

[9]It is justifiable to use an equation system, instead of using a least–square type procedure, only if we assume that (27) is exactly valid. Computational experience shows this procedure to be admissible without any significant deterioration concerning the accuracy of the results.

iii) The i^{th} component of the last vector will be replaced by the limit value[10] of (28):

$$\lim_{t\to\infty} S_i(t) = K_i. \tag{32}$$

iv) The original iterative scheme will be continued by repeating the same procedure, until the accuracy is above a predetermined limit.

If there are only positive eigenvaluee, to enchance the efficiency of the procedure three consecutive vectors are to be used in *eqns.* (29–31). In special cases, for increasing the separation of eigenvalues any usual method (*e.g.* raising the matrix to power, *etc.*) can be appealed to.

Examples

The results obtained in the course of testing the procedure are summarized in Table I. In the columns 1, 2 and 3 the dimensions of matrices (MD)[11], the ratios of the total number of iterations (NIR) in the accelerated procedure related to that in the basic procedure and the ratios of the total iteration times (TIR) formed similarly as NIR are displayed. For testing purposes real symmetrical matrices having pseudo random number elements were used.

All the testing examples –quite independently from the dimensions of the matrices used–show the savings in the total numbers and times of iterations mainly to be between 0.4 and 0.6, *i.e.* the acceleration reduces both quantities roughly to the halves of their original values, thus giving hope to expect similar gains at larger matrices. When using a backstorage device for the high dimension of the matrix to be diagonalized, the gain in

[10]If necessary, it has to be retransformed.

[11]The dimensions of matrices were confined to 55 regarding that –for practical reasons– the method was tested on a microcomputer.

computer time is expected to be even larger.

Table I

Data comparing results

MD	NIR	TIR
3	0.52	0.60
3	0.56	0.69
4	0.59	0.68
5	0.60	0.70
6	0.52	0.58
6	0.48	0.52
7	0.60	0.68
8	0.51	0.58
8	0.52	0.55
10	0.51	0.54
10	0.32	0.44
12	0.48	0.51
20	0.51	0.53
30	0.56	0.57
55	0.56	0.57

Acknowledgement

In writing up this paper thanks are due to Dr. Imre Bálint for the consent of using many results of his work, consultations and advice during the period working in my group between January 1985 and July 1987. The support of this work is acknowledged to the Hungarian Ministry of Higher Education (grant Nos. P–437/84. and 746/86).

References

[1] *Varga, R.S.*: Matrix Iterative Analysis. Prentice Hall, Englewood Cliffs, 1962.

[2] *Householder, A.S.*: Principles of Numerical Analysis. McGraw–Hill Book Co., New York, 1953.

[3] *Householder, A.S.*: The Theory of Matrices in Numerical Analysis. Blaisdell, New York, 1964.

[4] *Lánczos, C.*: Applied Analysis. Prentice Hall, Englewood Cliffs, 1961.

[5] *Axelsson, O.*: "Solution of Linear System of Equations: Iterative Methods" in Sparse Matrix Techniques (V.A. Barker, ed.), Springer, Berlin–Heidelberg–New York, 1977.

МЕТОД БЫСТРОЙ ИТЕРАТИВНОЙ ДИАГОНАЛИЗАЦИИ ДЛЯ БОЛЬШИХ МАТРИКСОВ

М.И. БАН

Описана ускоренная техника итеративной диагонализации матриксов, позволяющая сократить время необходимое для проведения расчетов метод особенно полезень, когда скорость конвергенции чрезвычайно мала и итерационный процесс не может быть завершен.