2. **Тихонов А. Н.** О приближенных системах линейных алгебраических уравнений.— Журн. вычисл. математики и мат. физики, 1980, т. 20, № 6, с. 1373—1383.
3. **Булавский В. А.** Итеративный метод решения задач линейного программирования.— Докл. АН СССР, 1961, т. 137, № 2, с. 258—260.
4. **Еремин И. И.** Обобщение релаксационного метода Моцкина — Агмона.— Успехи мат. наук, 1965, т. 20, вып. 2, с. 183—187.
5. **Булавский В. А.** Методы релаксации для систем неравенств. — Новосибирск: НГУ, 1981.— 83 с.
6. **Motzkin T. S., Schoenberg J. J.** The relaxation method for linear inequalities.— Canad. J. Mathem. 1954, N 6, № 3, p. 394—404.
7. **Еремин И. И., Мазуров В. Д., Астафьев Н. Н.** Несобственные задачи линейного и выпуклого программирования.— М.: Наука, 1983.— 336 с.
8. **Булавский В. А.** Релаксация в задачах с неравенствами.— В кн.: Оптимизация. Новосибирск: ИМ СО АН СССР, 1979, вып. 23 (40), с. 32—40.
9. **Булавский В. А.** Квазилинейное программирование и векторная оптимизация.— Докл. АН СССР, 1981, т. 257, № 4, с. 788—791.
10. **Eaves B. C.** The linear complementarity problem.— Manag. Sci., 1971, v. 17, p. 612—634.

# ПРИМЕНЕНИЕ ОБОБЩЕННОЙ ТЕОРЕМЫ ШТУРМА ДЛЯ ВЫЧИСЛЕНИЯ СОБСТВЕННЫХ ЧИСЕЛ ОДНОГО КЛАССА МАТРИЦ

*В. А. БУЛАВСКИЙ, М. А. ЯКОВЛЕВА*

Статья является расширенным изложением и развитием ранее опубликованной работы авторов [1]. Вместо предложенного там термина «обобщенные трехдиагональные матрицы» здесь вводится название «обобщенные якобиевы матрицы», поскольку слово «трехдиагональные» оказалось чересчур наглядным и настоящих трех ненулевых диагоналей у интересующих нас матриц, строго говоря, нет. Основное содержание статьи можно охарактеризовать следующим образом. Рассматривается класс матриц, содержащий, в частности, трехдиагональные матрицы. Для этого класса формулируется и доказывается подходящим образом обобщенная теорема Штурма о числе корней многочлена. На ее основе описывается метод определения границ собственных чисел обобщенной якобиевой матрицы, полностью аналогичный известному методу для трехдиагональных матриц [2, 3]. Для данного метода приводятся условия и оценки, гарантирующие его безаварийную реализацию на ЭВМ и объявленную точность полученного результата.

## § 1. ОБОБЩЕННЫЕ ЯКОБИЕВЫ МАТРИЦЫ

Класс матриц, который вводится в рассмотрение, естественным образом обобщает понятие трехдиагональной матрицы. Фактически это понятие определяется конфигурацией элементов, значения которых не предполагаются нулевыми. Если дополнительно некоторые из них окажутся нулями, то матрица может распасться на независимые диагональные клетки той же структуры. Поэтому в определении удобно описывать не фактическое расположение ненулевых элементов, а допустимую их конфигурацию.

Рассмотрим отображение

$$i \to k(i), \quad i = 1, 2, \ldots, n-1, \tag{1}$$

где $i+1 \leqslant k(i) \leqslant n$ при всех $i$.

**Определение.** *Матрица $A$ порядка $n$ с элементами $a_{ij}$ называется обобщенной якобиевой матрицей со структурой* (1), *если у нее отличны от нуля разве лишь диагональные элементы, а также элементы $a_{i,\,k(i)}$ и $a_{k(i),\,i}$, $i = 1, 2, \ldots, n-1$.*

В дальнейшем для ненулевых элементов обобщенной якобиевой матрицы примем обозначения:

$$d_i = a_{ii}, \; a_i = a_{i,\,k(i)}, \; b_i = a_{k(i),\,i}.$$

Легко понять, что обычная трехдиагональная структура получается, если $k(i)=i+1$. Такая структура сохраняется лишь при одной перестановке строк и столбцов — если их перенумеровать в обратном порядке. В общем же случае возможных перестановок больше, и для наглядности изображения обобщенной якобиевой матрицы удобно путем перестановки строк и столбцов привести ее к некоторой стандартной форме. Чтобы сделать это, перейдем на язык графов.

Свяжем со структурой (1) граф, вершинами которого являются номера строк, а дугами упорядоченные пары $[i,\ k(i)]$, $i=1,\ 2,\ \ldots,\ n-1$. Поскольку $k(i)>i$, то такой граф не содержит циклов, а так как к тому же любая вершина $i$, кроме вершины $n$, является начальной ровно в одной дуге, то полученный граф служит связным деревом с корнем в вершине $n$. Для каждой вершины $k$ графа обозначим через $F(k)$ совокупность начальных вершин тех дуг, для которых вершина $k$ — конечная, т. е.

$$F(k)=\{i : k=k(i)\}. \tag{2}$$

Отметим, что для вершин, не имеющих входящих в них дуг, множества $F(k)$ пустые. Вершины из множества $F(k)$ непосредственно предшествуют вершине $k$. Если же для некоторой вершины $s$ имеется цепочка вершин $s_0,\ s_1,\ \ldots,\ s_m$, где $s_0=s$, $s_m=k$ и $s_{\nu-1}\in F(s_\nu)$, $\nu=1,\ 2,\ \ldots,\ m$, то вершину $s$ будем считать предшествующей вершине $k$. Совокупность, состоящую из вершины $k$ и всех предшествующих ей вершин, обозначим через $H(k)$. Нетрудно установить, что $H(s)\subset H(k)$ в том и только том случае, когда вершина $s$ предшествует вершине $k$. Если же ни одна из этих вершин не предшествует другой, то $H(s)\cap H(k)=\varnothing$. Построенный граф будем называть соответствующим структуре (1). Отметим, что для множеств $H(k)$ можно записать рекуррентное правило получения:

$$H(k)=\{k\}\cup\Big[\bigcup_{i\in F(k)} H(i)\Big]. \tag{3}$$

Для наглядного представления обобщенной якобиевой матрицы предположим, что при выбранной нумерации строк и столбцов вершины, попадающие в одно множество $H(s)$, имеют номера без пропусков из некоторого промежутка, или подматрица $A_s$, стоящая в пересечении строк и столбцов с номерами из $H(s)$, является квадратной диагональной клеткой. Возможность такой нумерации легко устанавливается рекурсивно. Заметим, что подматрица $A_s$ при любом $s$ сама обобщенная якобиевая матрица со структурой, индуцированной структурой (1). Теперь можем класс рассматриваемых матриц описать рекурсивно следующим образом.

Простейшими обобщенными якобиевыми матрицами являются матрицы порядка один, т. е. скаляры. Если уже построены обобщенные якобиевы матрицы

$$A_i,\ i\in F(k), \tag{4}$$

то для построения матрицы $A_k$ нужно из них составить клеточно-диагональную матрицу и окаймить ее строкой и столбцом, в которых кроме диагонального элемента $d_k$ отличны от нуля лишь элементы, номера которых совпадают с корневыми вершинами графов, соответствующих матрицам (4), т. е. элементы $a_i$ и $b_i$ при $i\in F(k)$. Пример такого рекурсивного построения дает следующая матрица:

$$\left[\begin{array}{ccc|c|ccc|c}
d_1 & \cdot & a_1 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & d_2 & a_2 & \cdot & \cdot & \cdot & \cdot & \cdot \\
b_1 & b_2 & d_3 & \cdot & \cdot & \cdot & \cdot & a_3 \\
\hline
\cdot & \cdot & \cdot & d_4 & \cdot & \cdot & \cdot & a_4 \\
\hline
\cdot & \cdot & \cdot & \cdot & d_5 & \cdot & a_5 & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & d_6 & a_6 & \cdot \\
\cdot & \cdot & \cdot & \cdot & b_5 & b_6 & d_7 & a_7 \\
\hline
\cdot & \cdot & b_3 & b_4 & \cdot & \cdot & b_7 & d_8
\end{array}\right]. \tag{5}$$

Возьмем случай симметричной обобщенной якобиевой матрицы, где положим $a_i = b_i = c_i$, $i = 1, 2, \ldots, n-1$. Можно рассмотреть и более общий случай, когда $a_i$ и $b_i$ различны, но $a_i \cdot b_i > 0$. Однако, как видно из доказываемой ниже теоремы, такие матрицы подобны симметричным матрицам той же структуры с внедиагональными элементами $c_i = \sqrt{a_i b_i}$, так что при вычислении собственных значений можно обойтись симметричным случаем.

**Теорема 1.** *Если в обобщенной якобиевой матрице $a_i b_i > 0$ при всех $i = 1, 2, \ldots, n-1$, то существуют такие положительные числа $\lambda_i$, $i = 1, 2, \ldots, n-1$, что при умножении каждой строки и делении каждого столбца матрицы на соответствующее $\lambda_i$ получается симметричная матрица той же структуры.*

Доказательство. Проведем индукцию по порядку матрицы. Для матриц первого порядка утверждение тривиально. Если оно справедливо для матриц, порядок которых меньше $n$, то для матрицы $A$ порядка $n$ с обобщенной якобиевой структурой (1) можно поступить следующим образом. Рассмотрим соответствующий граф и для каждого $i \in F(n)$ положим $\lambda_i = \sqrt{b_i/a_i}$. На $\lambda_i$ умножим и разделим соответствующие строки и столбцы с номерами $i \in F(n)$. Таким образом добьемся симметрии для элементов последней строки и последнего столбца. Затем для уже измененных матриц $A_i$, $i \in F(n)$, имеющих порядок меньше $n$, определим нужные числа

$$\lambda_s, \quad s \in \bigcup_{i \in F(n)} [H(i) \setminus \{i\}],$$

и на них умножим и разделим соответствующие строки и столбцы. Поскольку при этом не затрагиваются ненулевые элементы последней строки и последнего столбца матрицы $A$, то вся матрица окажется симметричной. Теорема доказана.

Отметим, что при построенном в теореме преобразовании не меняются диагональные элементы и произведения симметрично расположенных внедиагональных элементов. Поэтому в полученной симметричной матрице произведение элементов с номерами $(i, k(i))$ и $(k(i), i)$ будет равно $a_i b_i$ и можно считать, что $c_i = \sqrt{a_i b_i}$.

## § 2. ОБОБЩЕННАЯ ТЕОРЕМА ШТУРМА

Рассмотрим снова граф, соответствующий структуре (1), и построим по нему множества (2). Каждой вершине $i$ сопоставим некоторый многочлен $g_i(t)$ и положим

$$G_k(t) = \prod_{i \in F(k)} g_i(t), \quad Q_k(t) = \frac{g_k(t)}{G_k(t)}. \tag{6}$$

Если $F(k) = \varnothing$, то, как обычно считаем, что $G_k(t) \equiv 1$. Обозначим через $S(t)$ число отрицательных чисел в совокупности

$$\{Q_k(t), \quad k = 1, 2, \ldots, n\}. \tag{7}$$

**Теорема 2.** *Пусть для любого $k = 1, 2, \ldots, n$ выполнены следующие условия:*

а) *никакие два из многочленов $g_i(t)$, $i \in F(k) \cup \{k\}$, не имеют общих корней;*

б) *с возрастанием $t$ отношение $Q_k(t)$ меняет знак с плюса на минус при переходе через корень многочлена $g_k(t)$ и, наоборот, с минуса на плюс при переходе через корень $G_k(t)$;*

в) *числа $\alpha$ и $\omega$ не являются корнями многочлена $g_k(t)$. Тогда число корней многочлена $g_n(t)$, лежащих в промежутке $(\alpha, \omega)$, равно разности $S(\omega) - S(\alpha)$.*

Доказательство. Если некоторая точка $t \in [\alpha, \omega]$ не является корнем ни одного из многочленов $g_i(t)$, то она не является корнем и многочленов $G_k(t)$, $k = 1, 2, \ldots, n$. Следовательно, в ее окрестности знаки членов совокупности (7) не меняются. Пусть $t$ — корень многочленов $g_i$ при $i \in I \subset \{1, 2, \ldots, n\}$. По условию а), во-первых, $k(i) \notin I$ при всех $i \in I \setminus \{n\}$ и, во-вторых, $k(s) \neq k(j)$ для любых двух различных номеров $s$ и $j$ из множества $I \setminus \{n\}$. Поэтому множество

$$J = \{k(i) : i \in I \setminus \{n\}\}$$

не пересекается с множеством $I$ и имеет столько же элементов, что и множество $I \setminus \{n\}$. С другой стороны, согласно условию б) величины $Q_k(t)$, $k \in J$, меняют знак с минуса на плюс, а величины $Q_i(t)$, $i \in I$, с плюса на минус. Остальные члены семейства (7) в точке $t$ имеют конечные и отличные от нуля значения, так что знака при переходе аргумента через точку $t$ не меняют. Таким образом, если $n \notin I$, то число отрицательных членов в семействе (7) не меняется. Если же $n \in I$, то число элементов в множестве $I$ на единицу больше, чем в множестве $J$, и в семействе (7) число отрицательных чисел увеличивается на единицу. Теорема доказана.

Данная теорема считает число корней многочлена $g_n(t)$ без учета их кратности и фактически приспособлена для случая простых корней. Если же возможны и кратные корни, воспользуемся следующей конструкцией.

Для каждого $\varepsilon > 0$ рассмотрим семейство многочленов

$$g_i(t, \varepsilon), \quad i = 1, 2, \ldots, n, \tag{8}$$

и построим по ним функции

$$G_k(t, \varepsilon) = \prod_{i \in F(k)} g_i(t, \varepsilon), \quad Q_k(t, \varepsilon) = \frac{g_k(t, \varepsilon)}{G_k(t, \varepsilon)}. \tag{9}$$

**Теорема 3.** *Предположим, что выполнено условие* в) *теоремы* 2 *и при всех* $\varepsilon > 0$ *многочлены* (8) *удовлетворяют следующим условиям:*

1) *для всех* $i$ *коэффициенты многочленов* $g_i(t, \varepsilon)$ *и* $g_i(t)$ *отличаются не более чем на* $\varepsilon$;

2) *корни многочлена* $g_n(t, \varepsilon)$ *вещественные и простые;*

3) *функции* (8) *и* (9) *удовлетворяют условиям* а) *и* б) *теоремы* 2.

*Тогда разность* $S(\omega) - S(\alpha)$ *совпадает с суммарной кратностью корней многочлена* $g_n(t)$, *лежащих в промежутке* $(\alpha, \omega)$.

Доказательство. Так как значение многочлена непрерывно зависит от коэффициентов, то при достаточно малых $\varepsilon$ многочлены (8) не обращаются в ноль в точках $\alpha$ и $\omega$, знаки величин $Q_k(t, \varepsilon)$ в этих точках такие же, как и у величин $Q_k(t)$. Следовательно, по теореме 2 при достаточно малых $\varepsilon$ разность $S(\omega) - S(\alpha)$ совпадает с числом корней многочлена $g_n(t, \varepsilon)$ в промежутке $(\alpha, \omega)$. Ввиду непрерывной зависимости корней многочлена от его коэффициентов и условия 2) нужное утверждение получается переходом к пределу при $\varepsilon \to 0$. Теорема доказана.

Рассмотрим теперь симметричную обобщенную якобиеву матрицу $A$ порядка $n$ со структурой (1), диагональными элементами $d_k$, $k = 1, 2, \ldots, n$, и внедиагональными элементами $c_i$, $i = 1, 2, \ldots, n-1$. Множества $F(k)$ и $H(k)$ снова построим по формулам (2) и (3) и рассмотрим подматрицы $A_k$, стоящие в пересечении строк и столбцов с номерами из $H(k)$. Семейство многочленов $g_k(t)$ определим формулами

$$g_k(t) = \det[A_k - tI], \tag{10}$$

где $I$ — единичная матрица. Учитывая рекуррентную формулу (3) для множеств $H(k)$ и описанный в первом параграфе способ рекурсивного построения обобщенных якобиевых матриц, раскроем определитель (10) по элементам $k$-го столбца и $k$-й строки. Получим

$$g_k(t) = (d_k - t) G_k(t) - \sum_{i \in F(k)} c_i^2 G_i(t) \cdot \frac{G_k(t)}{g_i(t)}. \tag{11}$$

Заметим, что формула остается справедливой и для случая пустого $F(k)$, так как в этом случае матрица $A_k$ совпадает с диагональным элементом $d_k$, $G_k(t) \equiv 1$, а пустая сумма равна нулю. Таким образом, для концевых вершин графа, соответствующего структуре (1), $g_k(t) = d_k - t$. Разделив левую и правую части равенства (11) на $G_k(t)$, находим

$$Q_k(t) = (d_k - t) - \sum_{i \in F(k)} \frac{c_i^2}{Q_i(t)}, \quad k = 1, 2, \ldots, n. \tag{12}$$

**Теорема 4.** *Пусть $c_i \neq 0$, $i = 1, 2, \ldots, n-1$, числа $\alpha$ и $\omega$ не являются корнями многочленов* (10), *а $S(t)$ при каждом $t$ — число отрицательных чисел в совокупности* (7). *Тогда разность $S(\omega) - S(\alpha)$ совпадает с числом корней многочлена*

$$g_n(t) = \det(A - tI), \tag{13}$$

*лежащих в промежутке $(\alpha, \omega)$, если корни считать с учетом их кратности.*

Доказательство. Сначала рассмотрим случай, когда при любом $k$ никакие два многочлена $g_s(t)$ и $g_j(t)$ для $s, j \in F(k)$, $s \neq j$, не имеют общих корней. Покажем, что при этом дополнительном предположении выполнены условия а) и б) теоремы 2. Для проверки условия а) остается лишь установить, что многочлены $g_k(t)$ и $g_i(t)$ при $i \in F(k)$ также не имеют общих корней. Будем рассуждать от противного. Если это не так, то можно выбрать наименьший номер $k$, для которого $g_k(t) = g_j(t) = 0$ при некоторых $j \in F(k)$ и $t$. Поскольку тогда $G_k(t) = 0$ и $g_i(t) \neq 0$ для $i \in F(k) \setminus \{j\}$, то согласно (11)

$$0 = c_j^2 G_j(t) \cdot \Big[ \prod_{s \in F(k) \setminus \{j\}} g_s(t) \Big].$$

В этом произведении $c_j^2$ и выражение в квадратных скобках отличны от нуля. Следовательно, $G_j(t) = 0$ и найдется номер $\mu \in F(j)$, для которого $g_\mu(t) = 0$. Но $k = k(j) > j$, что противоречит минимальности выбранного $k$. Таким образом, условие а) теоремы 2 выполнено.

Покажем, что в любой точке непрерывности, т. е. за исключением корней многочленов $g_i(t)$, $i \in F(k)$, отношения $Q_k(t)$ убывают. Проверку данного утверждения можно провести индуктивно. Так как $Q_1(t) = d_1 - t$, то $Q_1'(t) = -1 < 0$. Предположим, что утверждение справедливо для номеров, меньших $k$. Поскольку из (12) находим

$$Q_k'(t) = -1 + \sum_{i \in F(k)} \frac{c_i^2 Q_i'(t)}{[Q_i(t)]^2} \tag{14}$$

и $Q_i'(t) \leqslant 0$ для $i < k$ (для $i \in F(k)$, в частности), то $Q_k'(t) \leqslant -1$. То обстоятельство, что формула (14) применима лишь вне корней и полюсов отношений $Q_i(t)$, $i \in F(k)$, не имеет значения, так как таких исключительных точек конечное число и неравенство $Q_k(t) \leqslant -1$ будет выполняться в них по непрерывности.

Теперь можем проверить выполнение условия б) теоремы 2. Действительно, как было показано выше, корни многочлена $G_k(t)$ не совпадают с корнями многочлена $g_k(t)$ и поэтому являются полюсами отношения $Q_k(t)$. Ввиду того, что это отношение убывает в окрестности полюса, при переходе через полюс оно меняет знак с минуса на плюс. Наоборот, корень многочлена $g_k(t)$ является одновременно и корнем отношения $Q_k(t)$, ввиду его убывания отношение меняет знак с плюса на минус. Заметим, что у $g_k(t)$ и $Q_k(t)$ совпадают также кратности корней, и поскольку $Q_k'(t) \leqslant -1$, то все корни многочленов (10), и в частности многочлена $g_n(t)$, простые.

Покажем теперь, что сколь угодно малым изменением диагональных элементов матрицы $A$ можно добиться выполнения сделанного в начале доказательства дополнительного предположения. Если некоторое число

$\delta_i$ добавить ко всем диагональным элементам $d_\nu$ при $\nu \in H(i)$, то ко всем матрицам $A_s$, $s \in H(i)$, добавится скалярная матрица $\delta_i I$, т. е. все собственные числа указанных матриц сдвинутся на $\delta_i$, а их взаимное расположение не изменится. Если же $s < k = k(i)$ и $s \notin H(i)$, то $H(s) \cap \cap H(i) = \varnothing$ и собственные числа матрицы $A_s$ вообще не меняются.

Будем добиваться взаимной простоты многочленов

$$g_i(t), \quad i \in F(k), \tag{15}$$

поочередно в порядке возрастания номера $k$. Допустим, что для $j < k$ многочлены $g_s(t)$, $s \in F(j)$, взаимно простые. Если многочлены (15) не взаимно просты, то можно подобрать сколь угодно малые изменения $\delta_i$, $i \in F(k)$, диагональных элементов $d_\nu$, $\nu \in H(i)$, после которых многочлены (15) уже будут взаимно просты. Взаимная простота многочленов $g_s(t)$, $s \in F(j)$ при $j < k$, как отмечалось выше, не нарушится. Таким образом, действительно можно добиться взаимной простоты всех групп (15) сколь угодно малым изменением диагональных элементов матрицы $A$.

Для завершения доказательства остается сослаться на теорему 3. По любому $\varepsilon > 0$ можно так изменить диагональные элементы, что группы многочленов (15) будут взаимно простые и коэффициенты характеристических многочленов матриц $A_k$, $k = 1, 2, \ldots, n$, будут отличаться от коэффициентов исходных многочленов (10) меньше чем на $\varepsilon$. Если новые характеристические многочлены принять в качестве семейства (8), то будут выполнены условия теоремы 3. Теорема доказана.

В заключение заметим, что $g_k(t) \to +\infty$, $k = 1, 2, \ldots, n$, при $t \to -\infty$. Поэтому $S(t) = 0$ при больших отрицательных $t$. Следовательно, число отрицательных чисел в совокупности (7) равно числу собственных чисел матрицы $A$, лежащих левее $t$, с учетом их кратности.

## § 3. ВЫЧИСЛЕНИЕ ГРАНИЦ СОБСТВЕННЫХ ЧИСЕЛ

Принципиальная схема алгоритма для определения границ собственных чисел матрицы $A$, основанная на использовании теоремы Штурма, хорошо известна. Если имеются границы $\alpha_0$ и $\omega_0$ спектра матрицы или интересующей нас его части, то дальнейшее уточнение границ отдельных собственных чисел можно осуществить, вычисляя величины (7) для некоторых значений $t$ по рекуррентным формулам (12) и определяя среди них число $S(t)$ отрицательных. На каждом частичном промежутке $(\alpha, \omega)$, полученном в процессе дробления, согласно теореме 4, расположено $S(\omega) - S(\alpha)$ корней характеристического многочлена. Начальные границы спектра можно вычислить на основании теоремы Гершгорина. Если положить

$$a_k = \sum_{i \in F(k)} |c_i|,$$

то получим

$$\alpha_0 = \min_k \{d_k - |c_k| - a_k\} \tag{16}$$

$$\omega_0 = \max_k \{d_k + |c_k| + a_k\}. \tag{17}$$

Здесь и в дальнейшем для единообразия записи положим $c_n = 0$. Отметим также, что для пустых множеств $F(k)$ получается $a_k = 0$. Для последующих оценок введем в рассмотрение число

$$H = \max\{|\alpha_0|, |\omega_0|\} = \max_k \{|d_k| + |c_k| + a_k\}. \tag{18}$$

Все точки $t$, для которых потребуется проводить вычисления, по абсолютной величине не превосходят $H$.

При реальном счете на ЭВМ с плавающей запятой возникают те же проблемы гарантированного ответа и безаварийной работы алгоритма, что и для случая обычных якобиевых матриц. Во-первых, нужно видо-

изменить формулы (12), чтобы обеспечить безаварийное вычисление величин (7); во-вторых, погрешности вычисления и погрешности, внесенные изменением формул (12), преобразовать в эквивалентные возмущения ненулевых элементов матрицы $A$. Наконец, нужно выяснить гарантированные границы для $r$-го и $(r+1)$-го собственных чисел точной матрицы, если в вычисленной совокупности величин (7) оказалось $r$ отрицательных.

Примем стандартное предположение о том, что заданы положительные числа $\varepsilon_1$ и $\varepsilon_2$, характеризующие процесс приближенных вычислений в следующем смысле:

а) число $1/\varepsilon_2$ еще не выходит за разрядную сетку машины, а число $\varepsilon_2/2$ при запоминании еще не заменяется нулем;

б) погрешности, возникающие при сложении, вычитании, умножении и делении чисел $\alpha$ и $\beta$, не превосходят соответственно

$$\varepsilon_1|\alpha+\beta|+\frac{\varepsilon_2}{2},\quad \varepsilon_1|\alpha-\beta|+\frac{\varepsilon_2}{2},$$

$$\varepsilon_1|\alpha\beta|+\frac{\varepsilon_2}{2},\qquad \varepsilon_1\left|\frac{\alpha}{\beta}\right|+\frac{\varepsilon_2}{2}.$$

Отметим, что число $\varepsilon_1$ равно наименьшей добавке к единице, которая может быть реализована в машине.

Заменим формулы (12) для вычисления $Q_i(t)$ на формулы

$$S_i=(d_i-t)-\sum_{j\in F(i)}\left[\frac{c_j}{Q_j(t)}\right]c_j, \tag{19}$$

$$\beta_i=[2\varepsilon_2\cdot|c_i|]\cdot a_{k(i)}+\frac{\varepsilon_2}{2}, \tag{20}$$

$$Q_i(t)=\begin{cases} S_i, & \text{если } |S_i|\geqslant\beta_i,\\ \beta_i, & \text{если } \beta_i>S_i>0,\\ -\beta_i, & \text{если } 0\geqslant S_i>-\beta_i. \end{cases} \tag{21}$$

Скобки в формулах (19) и (20) указывают порядок умножения.

Приведенные формулы очевидным образом обеспечивают неравенство $|Q_i(t)|\geqslant\frac{\varepsilon_2}{2}$, а так как при $j\in F(i)$ оказывается $k(j)=i$, то $|Q_j(t)|\geqslant 2\varepsilon_2|c_j|\cdot a_i$ и, следовательно,

$$\sum_{j\in F(i)}\frac{c_j^2}{|Q_j(t)|}\leqslant\frac{1}{2\varepsilon_2}\sum_{j\in F(i)}\frac{|c_j|}{a_i}=\frac{1}{2\varepsilon_2}. \tag{22}$$

Чтобы при вычислении $S_i$ и $\beta_i$ не выйти за пределы допустимого диапазона чисел, следует наложить ограничения на исходные данные. Будем предполагать, что абсолютные величины всех чисел $c_k$, $d_k$ и порядок матрицы $n$ не превосходит величины $1/2\sqrt{\varepsilon_2}$. Кроме того, во всех оценках будем исходить из неравенства $\varepsilon_2\leqslant 1/400$. Тогда

$$a_k\leqslant(n-1)\frac{1}{2\sqrt{\varepsilon_2}}\leqslant\frac{1}{4\varepsilon_2},$$

$$|t|\leqslant H=\max\{|\alpha_0|,\ |\omega_0|\}\leqslant\frac{1}{\sqrt{\varepsilon_2}}+\frac{1}{4\varepsilon_2}<\frac{1}{3\varepsilon_2}, \tag{23}$$

$$\beta_i\leqslant\left[2\varepsilon_2\cdot\frac{1}{2\sqrt{\varepsilon_2}}\right]\cdot\frac{1}{4\varepsilon_2}+\frac{\varepsilon_2}{2}<\frac{1}{3\sqrt{\varepsilon_2}}.$$

Поскольку $a_{k(j)}\geqslant|c_j|$, то $\beta_j\geqslant 2\varepsilon_2|c_j|^2+\frac{\varepsilon_2}{2}$, отсюда

$$\frac{|c_j|}{|Q_j(t)|}\leqslant\frac{|c_j|}{2\varepsilon_2|c_j|^2+\varepsilon_2/2}\leqslant\frac{1}{\varepsilon_2}\cdot\frac{1}{(2|c_j|+1/2|c_j|)}\leqslant\frac{1}{2\varepsilon_2}.$$

Наконец, учитывая (22) и (23), получим

$$|S_i| \leqslant \frac{1}{2\sqrt{\varepsilon_2}} + \frac{1}{3\varepsilon_2} + \frac{1}{2\varepsilon_2} \leqslant \frac{11}{12} \cdot \frac{1}{\varepsilon_2}.$$

Сделанные оценки показывают, что счет по формулам (19)—(21) при указанных выше границах для исходных данных будет идти безаварийно.

Проведем теперь обратный анализ погрешностей вычислений. При этом будем считать, что фактически вычисленные значения $Q_i(t)$, $i = 1, 2, \ldots, n$, являются точными для некоторой возмущенной матрицы той же структуры (1), но с элементами

$$d_i + \delta_i, \quad c_j + \gamma_j, \quad j \in F(i). \tag{24}$$

Во всех оценках будем пренебрегать слагаемыми с множителями $\varepsilon_1^2$ и $\varepsilon_1 \cdot \varepsilon_2$, если имеются аналогичные слагаемые с множителями $\varepsilon_1$ и $\varepsilon_2$ соответственно.

При вычислении отношения $c_j/|Q_j(t)|$ совершаем погрешность, оцениваемую величиной

$$\varepsilon_1 \frac{|c_j|}{|Q_j(t)|} + \frac{\varepsilon_2}{2}.$$

При последующем умножении на $c_j$ эта ошибка также умножится на $c_j$. Кроме того, к ней добавится еще ошибка умножения. Таким образом, слагаемое $c_j^2/Q_j(t)$ будет вычислено с ошибкой, не превосходящей величину

$$2\varepsilon_1 \frac{c_j^2}{|Q_j(t)|} + (1 + |c_j|)\frac{\varepsilon_2}{2}, \quad j \in F(i).$$

При вычислении правой суммы в (19) сама сумма и промежуточные результаты не превзойдут числа

$$\sum_{j \in F(i)} \frac{c_j^2}{|Q_j(t)|}.$$

Если через $|F(i)|$ обозначить число элементов в множестве $F(i)$, то при сложении накопится погрешность в пределах оценки

$$\left[\varepsilon_1 \sum_{j \in F(i)} \frac{c_j^2}{|Q_j(t)|} + \frac{\varepsilon_2}{2}\right] \cdot (|F(i)| - 1).$$

При определении разности $d_i - t$ погрешность не превзойдет $\varepsilon_1(|d_i| + |t|) + \varepsilon_2/2$. При последнем вычитании в (19) получим еще ошибку, оцениваемую величиной

$$\varepsilon_1 \left[|d_j| + |t| + \sum_{j \in F(i)} \frac{c_j^2}{|Q_j(t)|}\right] + \frac{\varepsilon_2}{2}.$$

Наконец, если вместо $S_i$ в (21) в качестве $Q_i(t)$ возьмем $\beta_i$ или $-\beta_i$, то дополнительная ошибка не превзойдет $\beta_i$. Суммируя все погрешности, найдем, что ошибка при вычислении $Q_i(t)$ оценивается выражением

$$\varepsilon_1 [|F(i)| + 2] \cdot \sum_{j \in F(i)} \frac{c_j^2}{|Q_j(t)|} + 2\varepsilon_1 \cdot (|d_i| + |t|) +$$

$$+ \frac{\varepsilon_2}{2} [2 \cdot |F(i)| + 2 + 4 \cdot |c_i| \cdot a_{k(i)} + a_i].$$

Если ввести обозначение $R = \max\limits_k \{|F(k)|\}$ и учесть, что $|t| \leqslant H$, $|c_i| \leqslant H$, $a_{k(i)} \leqslant H$ и $a_i \leqslant H$, где $H$ определено по формуле (18), то получим, что возмущения $\delta_i$ и $\gamma_j$ в (24), эквивалентные ошибке при вычис-

лении $Q_i(t)$, могут быть выбраны так, что

$$|\delta_i| \leqslant 2\varepsilon_1(|d_i| + H) + \frac{\varepsilon_2}{2}(2R + 2 + 4H^2 + H),$$

$$|\gamma_j| \leqslant \varepsilon_1 \frac{(R+2)}{2}|c_j|, \quad j \in F(i).$$

Таким образом, вычисленную совокупность величин (7) можно считать точной для матрицы $A+B$, где $B$ — симметричная обобщенная якобиева матрица структуры (1) с элементами $\delta_i$ и $\gamma_i$ вместо $d_i$ и $c_i$. Спектр матрицы возмущений $B$ оценивается так же, как и для матрицы $A$, т. е. по теореме Гершгорина он лежит в промежутке $[-h, h]$, где

$$h = \max_i \left\{|\delta_i| + |\gamma_i| + \sum_{j \in F(i)} |\gamma_j|\right\}.$$

Используя полученные для $\delta_i$ и $\gamma_j$ оценки, найдем

$$|\delta_i| + |\gamma_i| + \sum_{j \in F(i)} |\gamma_j| \leqslant \varepsilon_1 \left\{\frac{(R+2)}{2}\left(|c_i| + \sum_{j \in F(i)} |c_j|\right) + 2|d_i|\right\} +$$

$$+ 2\varepsilon_1 H + \frac{\varepsilon_2}{2}(2R + 2 + H + 4H^2).$$

Поскольку $2 \leqslant \frac{R+3}{2}$, то

$$h \leqslant \varepsilon_1 \cdot \frac{R+7}{2} H + \frac{\varepsilon_2}{2}(2R + 2 + H + 4H^2) = \Delta.$$

Пусть теперь $S(t) = r$. Это значит, что у матрицы $A + B + \Delta I$ левее $t + \Delta$ с учетом их кратности лежит ровно $r$ собственных значений. Так как матрица $B + \Delta I$ положительно полуопределенная, то у матрицы $A$ левее $t + \Delta$ лежит не менее $r$ собственных значений. Аналогично у матрицы $A + B - \Delta I$ левее $t - \Delta$ лежит $r$ собственных значений. Поскольку матрица $B - \Delta I$ отрицательно полуопределенная, то у матрицы $A$ левее $t - \Delta$ лежит не более $r$ собственных значений. Наконец, если вычисления проведены для $t = \alpha$ и $t = \omega$, $\omega > \alpha$, и установлено, что $S(\omega) - S(\alpha) = r$, то можно утверждать, что в промежутке $[\alpha - \Delta, \omega + \Delta]$ находится не менее $r$ собственных чисел матрицы $A$, а в промежутке $(\alpha + \Delta, \omega - \Delta)$ — не более $r$ собственных чисел.

В заключение отметим, что при получении величин $\alpha_0$ и $\omega_0$ по формулам (16) и (17) тоже следует оценить неточность их вычисления и надлежащим образом раздвинуть эти границы.

### ЛИТЕРАТУРА


1. **Булавский В. А., Яковлева М. А.** Обобщенные трехдиагональные матрицы и теорема Штурма.— Докл. АН СССР, 1984, т. 275, № 2, с. 277—280.
2. **Воеводин В. В.** Вычислительные основы линейной алгебры.— М.: Наука, 1977.— 304 с.
3. **Годунов С. К.** Решение систем линейных уравнений.— Новосибирск: Наука, Сиб. отд-ние, 1980.— 177 с.


## ВАРИАНТ АЛГОРИТМА ОРТОГОНАЛЬНОГО ПРИВЕДЕНИЯ МАТРИЦЫ К ДВУХДИАГОНАЛЬНОМУ ВИДУ


*А. Г. АНТОНОВ*


В работе представлен вариант приведения прямоугольной матрицы $A$ размерности $M \times N$ к двухдиагональному виду с помощью ортогональных преобразований отражения. Такое приведение предложено Хаусхолдером в работе [3] и подробно описано, в частности, в [1] и [2].