Х. Д. ИКРАМОВ

НЕСИММЕТРИЧНАЯ ПРОБЛЕМА СОБСТВЕННЫХ ЗНАЧЕНИЙ

ЧИСЛЕННЫЕ МЕТОДЫ

МОСКВА «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
1991

ББК 22.193
И42
УДК 519.61

Икрамов Х. Д. НЕСИММЕТРИЧНАЯ ПРОБЛЕМА СОБСТВЕННЫХ ЗНАЧЕНИЙ. Численные методы.— М.: Наука. Гл. ред. физ.-мат. лит., 1991.— 240 с.— ISBN 5-02-014462-2.

Посвящена важной задаче численной линейной алгебры — вычислению собственных значений и векторов несимметричных матриц. Основной текст книги представляет собой учебник по численным методам решения спектральных задач для несимметричных матриц; по уровню изложения он доступен студентам и выпускникам технических вузов. Дополнения к основному тексту рассчитаны на специалистов и дают обзор практически всей современной журнальной литературы в данной области.

Для студентов, аспирантов, научных сотрудников, специализирующихся в численном анализе и занимающихся решением спектральных задач на ЭВМ.

Табл. 5. Ил. 6. Библиогр. 217 наим.

Рецензент доктор физико-математических наук *А. А. Абрамов*

Научное издание

Икрамов Хаким Дододжанович

НЕСИММЕТРИЧНАЯ ПРОБЛЕМА СОБСТВЕННЫХ ЗНАЧЕНИЙ

Численные методы

Заведующий редакцией *Е. Ю. Ходан.* Редактор *Т. В. Шароватова*
Художественный редактор *Т. Н. Кольченко.* Технический редактор *С. Я. Шкляр*
Корректоры: *М. А. Смирнов, Н. Д. Дорохова*

ИБ № 32305

Сдано в набор 17.05.90. Подписано к печати 09.04.91 . Формат 60×90/16. Бумага тип. № 1. Гарнитура таймс. Печать офсетная. Усл. печ. л. 15,0. Усл. кр.-отт. 15,25 . Уч.-изд. л. 16,48 . Тираж 1800 экз. Заказ № 2275. Цена 5 р. 40 к.

Издательско-производственное и книготорговое объединение «Наука»
Главная редакция физико-математической литературы
117071 Москва В-71, Ленинский проспект, 15

Ордена Октябрьской революции и ордена Трудового Красного Знамени МПО «Первая Образцовая типография» Государственного комитета СССР по печати. 113054 Москва, Валовая, 28

И 1602110000-008 / 053(02)-91 10-92

ISBN 5-02-014462-2

ПРЕДИСЛОВИЕ

Под *полной проблемой собственных значений* в численной линейной алгебре понимают задачу вычисления *всех* собственных значений или собственных векторов матрицы. Если же нужны только одно или небольшая группа собственных значений (векторов), то говорят о *частичной проблеме.* В случае симметричной или эрмитовой матрицы употребляется термин *симметричная проблема собственных значений,* в противном случае — *несимметричная проблема.*

По-видимому, все эти словосочетания возникли как буквальный перевод соответствующих английских выражений «complete eigen problem», «partial eigenproblem» и т. п. Но если в английском языке «problem» означает попросту задачу, то русское слово «проблема» принадлежит совсем другому, более приподнятому стилевому ряду. Применение этого слова по отношению к обычной расчетной практике имеет, пожалуй, слегка комический оттенок.

В связи со сказанным мне припоминается один эпизод из собственной биографии. Молодым выпускником аспирантуры мехмата МГУ я отчитывался о своей диссертации на заседании Ученого совета. Как только впервые было произнесено «полная проблема собственных значений», член Совета профессор Е. М. Ландис встал и направился к ближайшему коллеге для выяснения, что это за проблема (дело происходило в не слишком заполненной аудитории, а ближайшим оказался академик П. С. Александров). То ли объяснение не удовлетворило профессора, то ли не понравился термин, но диссертант получил «черный шар», в чем винит не Е. М. Ландиса, а излишне витиеватую алгебраическую фразеологию.

И все же «проблема собственных значений» вынесена в название книги. Хоть это и непоследовательно, мотивы для такого решения были. Во-первых, употребление данного термина освящено традицией, идущей от классиков нашей научной области — В. Н. и Д. К. Фаддеевых. Во-вторых, самая знаменитая книга по вычислительной линейной алгебре называется «Алгебраическая проблема собственных значений» (Дж. Уилкинсон.— М.: Наука, 1970). В-третьих, сравнительно недавно в русском переводе издана книга американского алгебраиста Б. Парлетта «Симметричная проблема собственных значений» (М.: Мир, 1983). Настоящая книга задумана как дополнение к ней, посвященное спектральным задачам для несимметричных матриц.

Если исключить почти полное совпадение названий, эта книга совсем не похожа на книгу Парлетта. Последняя рассчитана на

читателя с изрядной математической культурой и читается местами с немалым напряжением. Так, во всяком случае, кажется автору, бывшему одним из ее переводчиков. Напротив, здесь в качестве основного читателя или читателя основного текста (и то и другое будет правильно!) предполагается инженер, физик или химик, вообще работник нематематической специальности, прослушавший типовой курс линейной алгебры, заканчивающийся теоремой о жордановой форме и ее разъяснениями (доказательство при этом не обязательно). Автор старался облегчить такому читателю овладение материалом книги. Поэтому, если доказательство какой-либо теоремы по сложности или громоздкости превышало средний уровень, оно переносилось в раздел дополнений (подобный раздел имеется почти в каждом параграфе книги) либо заменялось ссылкой на книгу или статью, где это доказательство можно найти. Зато на обсуждение смысла теорем, их мотивировок и следствий автор не жалел места.

Разделы дополнений и два параграфа, отмеченные звездочкой, содержат, кроме того, обзор самой свежей — на момент написания книги — журнальной литературы, а также работ, может быть, и не совсем новых, но незаслуженно незамеченных специалистами. Автор считает, что эти обзоры могут быть полезны для второй предполагаемой категории читателей — квалифицированных вычислителей-алгебраистов.

По мере сил автор стремился уменьшить число недостатков книги, но вряд ли вполне преуспел в этом. Всякому читателю, не пожалевшему своего времени, чтобы сообщить автору о замеченных недостатках и ошибках в книге, он будет глубоко признателен (впрочем, читателю, не затруднившемуся написать автору о ее достоинствах, он будет признателен еще больше!). Однако один недостаток известен автору уже сейчас. Именно пользоваться книгой как справочником (а не как учебником) нужно с осмотрительностью. Например, не стоит выхватывать из контекста формулировки отдельных теорем. В ряде случаев, для того чтобы «разгрузить» эти формулировки, часть условий переводилась в предшествующий текст или же оговаривалась в качестве общих условий параграфа.

Замысел этой книги был одобрен Екатериной Ивановной Стечкиной. Надеюсь, что его воплощение достойно ее светлой памяти.

Х. Д. Икрамов

СПИСОК ОБОЗНАЧЕНИЙ

$\equiv$	определяет величину, стоящую слева или справа от данного символа
$\mathbf{R}^n$ $(\mathbf{C}^n)$	арифметическое пространство n-мерных вещественных (комплексных) вектор-столбцов
$R^{k\times l}$ $(C^{k\times l})$	пространство вещественных (комплексных) $k \times l$-матриц
I, I_n	единичная матрица в $\mathbf{R}^{n\times n}$ $(\mathbf{C}^{n\times n})$
$A=[a_i^j]$	матрица с элементами a_{ij}
$A=[a_1 \mid \ldots \mid a_l]$	матрица со столбцами $a_1, \ldots, a_l$
n	как правило, обозначает порядок квадратной матрицы A
$\ker A$	ядро (нуль-пространство) матрицы A
$\operatorname{def} A$	дефект матрицы A
$\operatorname{Im} A$	образ (область значений) матрицы A
$\operatorname{rank} A$	ранг матрицы A
$\det A$	определитель квадратной матрицы A
$\operatorname{tr} A$	след квадратной матрицы A
$\operatorname{cond} A$	число обусловленности невырожденной матрицы A
$\lambda(A)$	собственное значение матрицы A
$\operatorname{ind}\lambda$	индекс собственного значения λ
$k(\lambda)$	число обусловленности собственного значения λ
$\sigma(A)$	спектр матрицы A
$\rho(A)$	спектральный радиус матрицы A
A^{T}	транспонированная матрица
A^*	сопряженная матрица
A^{-1}	обратная матрица
A^+	псевдообратная матрица Мура — Пенроуза
$A^{-\mathrm{T}}$	матрица, полученная транспонированием и обращением (выполняемыми в произвольном порядке)
A^{-*}	матрица, полученная сопряжением и обращением (выполняемыми в произвольном порядке)
$B\oplus C\oplus \ldots \oplus F$	прямая сумма матриц B, C, ..., F
$\otimes$	символ кронекерова произведения матриц
$\mathscr{D}_i(\alpha), \mathscr{N}_{ij}, \mathscr{L}_i$	элементарные матрицы
$\mathscr{P}$	матрица-перестановка
$\mathscr{P}_{ij}$	матрица-транспозиция
$\mathscr{H}$	матрица отражения
$\mathscr{R}_{ij}$	матрица вращения
$\mathscr{R}_{ij}$ или $\mathscr{U}_{ij}$	элементарная унитарная матрица
$\mathscr{S}(k, \nu)$	специальная симплектическая матрица
$\operatorname{diag}(\lambda_1, \ldots, \lambda_n)$	диагональная матрица с диагональными элементами $\lambda_1, \ldots, \lambda_n$
$J_m(\lambda)$	жорданова клетка порядка m для числа λ

$e_1, \ldots, e_n$	координатные векторы пространства R^n (C^n)
$\operatorname{span}(x_1, \ldots, x_m)$	линейная оболочка векторов $x_1, \ldots, x_m$
$\dim \mathscr{L}$	размерность подпространства $\mathscr{L}$
$\mathscr{M}^\perp$	ортогональное дополнение множества $\mathscr{M}$
$P_{\mathscr{L}}$	ортопроектор на подпространство $\mathscr{L}$
$S_{\mathscr{L}}$	единичная сфера подпространства $\mathscr{L}$
$\|\cdot\|_p$	гельдерова норма векторов с показателем p или соответствующая подчиненная матричная норма, в частности,
$\|\cdot\|_2$	евклидова длина векторов или спектральная матричная норма
$\|\cdot\|_E$	евклидова матричная норма
T_k	многочлен Чебышева первого рода степени k
$\deg \varphi$	степень многочлена φ

§ 1. Необходимые сведения из линейной алгебры

Основным объектом изучения в книге является множество собственных значений квадратной матрицы $A=[a_{ij}]$, вещественной или комплексной. Оно называется *спектром* этой матрицы и обозначается через $\sigma(A)$. Как правило, n обозначает порядок матрицы A; прямоугольные $k \times l$-матрицы используются в книге преимущественно как вспомогательные объекты.

Квадратная матрица A естественным образом интерпретируется как линейный оператор в арифметическом пространстве $\mathbf{R}^n$ или $\mathbf{C}^n$. Если, как это делается в книге, считать элементами пространства вектор-столбцы, то действие матрицы A на вектор x определяется с помощью операции матрично-векторного умножения: $y=Ax$. Возможность операторного взгляда на матрицу оправдывает применение таких терминов, как ядро (или нуль-пространство) матрицы, образ (область значений), инвариантное подпространство матрицы, сужение на подпространство и т. п.

В данном параграфе собраны необходимые сведения из стандартного, а также в некоторых случаях из более продвинутого курса линейной алгебры. Начнем с определений, относящихся к линейным подпространствам.

Линейные подпространства. *Размерность* подпространства $\mathscr{L}$, т. е. максимальное число линейно независимых векторов в нем, обозначается через $\dim \mathscr{L}$. Всякая максимальная линейно независимая система векторов $a_1, \ldots, a_l$ из подпространства $\mathscr{L}$ называется его *базисом*; при этом $l=\dim \mathscr{L}$. Матрица $A_{\mathscr{L}}$, составленная по столбцам из векторов $a_1, \ldots, a_l$:

$$A_{\mathscr{L}}=[a_1 | \ldots | a_l],$$

называется *базисной матрицей* подпространства $\mathscr{L}$. Исключая тривиальный случай $\mathscr{L}=\{0\}$, базисов и базисных матриц для подпространства $\mathscr{L}$ бесконечно много. Любые две базисные матрицы $A_{\mathscr{L}}$ и $B_{\mathscr{L}}$ данного подпространства связаны соотношением

$$B_{\mathscr{L}}=A_{\mathscr{L}}Q,$$

где Q — невырожденная матрица порядка l. Она называется *матрицей перехода* от одного базиса $\{a\} \equiv a_1, \ldots, a_l$ подпространства $\mathscr{L}$ к другому базису $\{b\}=b_1, \ldots, b_l$ и составлена по столбцам из координат векторов b_i в первом базисе.

Нередко линейные подпространства задают как линейные оболочки некоторых (ие обязательно линейно независимых) систем векторов.

При этом *линейной оболочкой* системы $c_1, \ldots, c_m$ называется множество всех линейных комбинаций $\alpha_1 c_1 + \ldots + \alpha_m c_m$, где коэффициенты в зависимости от контекста берутся из $\mathbf{R}$ или $\mathbf{C}$. Линейная оболочка обозначается символом $\operatorname{span}(c_1, \ldots, c_m)$, а векторы c_i называют ее *образующими*.

Если $\mathbf{R}^n$ рассматривается как евклидово пространство, то скалярное произведение векторов $x=(\xi_1, \ldots, \xi_n)^T$ и $y=(\eta_1, \ldots, \eta_n)^T$ обычно задается формулой

$$(x, y)=\xi_1\eta_1+\ldots+\xi_n\eta_n. \tag{1.1}$$

Точно так же в $\mathbf{C}^n$, трактуемом как унитарное пространство, скалярное произведение чаще всего определяют правилом

$$(x, y)=\xi_1\overline{\eta}_1+\ldots+\xi_n\overline{\eta}_n. \tag{1.2}$$

Векторы, ортогональные к фиксированному множеству $\mathcal{M}$ (т. е. ортогональные к каждому вектору из $\mathcal{M}$), образуют в совокупности подпространство, называемое *ортогональным дополнением* для $\mathcal{M}$ и обозначаемое через $\mathcal{M}^\perp$. Если, в частности, $\mathcal{M}$ — подпространство и $\dim \mathcal{M}=m$, то $\dim \mathcal{M}^\perp=n-m$.

Система векторов $x_1, \ldots, x_m$ *ортогональна*, если $(x_i, x_j)=0$ при $i\neq j$, и *ортонормирована*, если

$$(x_i, x_j)=\delta_{ij}\equiv\begin{cases}1, & i=j,\\ 0, & i\neq j.\end{cases}$$

Символ $\equiv$ здесь и в дальнейшем указывает на введение некоторой величины (за исключением тех случаев, когда он обозначает тождество), стоящей по одну его сторону; определение этой величины размещается с другой стороны.

Если система $x_1, \ldots, x_m$ линейно независима, но не ортогональна, то найдется система $y_1, \ldots, y_m$ такая, что

$$(x_i, y_j)=\delta_{ij}, \qquad i, j=1, \ldots, m. \tag{1.3}$$

Системы $x_1, \ldots, x_m$ и $y_1, \ldots, y_m$ называют *двойственными* или *биортогональными*. В общем случае для системы $x_1, \ldots, x_m$ существует бесконечно много биортогональных систем; так, при $m=1$ двойственным для вектора x будет любой вектор y, для которого $(x, y)=1$. Однако если $x_1, \ldots, x_n$ — базис пространства, то *биортогональный базис* $y_1, \ldots, y_n$ определен уже однозначно: всякий вектор y_i есть тот единственный вектор из одномерного подпространства $[\operatorname{span}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)]^\perp$, для которого $(x_i, y_i)=1$.

Возможен и полезен иной взгляд на пару биортогональных базисов $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$. Если построить из них по столбцам матрицы

$$X=[x_1 | \ldots | x_n], \qquad Y=[y_1 | \ldots | y_n], \tag{1.4}$$

то соотношения (1.3) приводят к матричному равенству

$$Y^T X=I \tag{1.5а}$$

в вещественном случае и к равенству

$$Y^* X=I \tag{1.5б}$$

в комплексном случае. Здесь I — принятое в книге обозначение *единичной матрицы*; если нужно явно указать ее порядок, то пишут I_n; Y^{τ} и Y^* — соответственно транспонированная и сопряженная матрицы: $Y^{\tau}=[y_{ji}]$, $Y^*=[\bar{y}_{ji}]$.

Подпространства $\mathscr{L}$ и $\mathscr{M}$ называют *дополнительными*, если всякий вектор z пространства можно (и притом однозначно) представить в виде

$$z=x+y, \quad x\in\mathscr{L}, \quad y\in\mathscr{M}. \tag{1.6}$$

Необходимые и достаточные условия дополнительности имеют вид $\mathscr{L}\cap\mathscr{M}=\{0\}$, $\dim\mathscr{L}=\dim\mathscr{M}^{\perp}$. Если $\mathscr{L}$ и $\mathscr{M}$ — дополнительные подпространства, то говорят, что пространство $\mathbf{R}^n$ (или $\mathbf{C}^n$) является их *прямой суммой*.

Оператор $P_{\mathscr{L},\mathscr{M}}$, который каждому вектору z ставит в соответствие компоненту x в разложении (1.6), называется *проектором* или, более подробно, *проектором на $\mathscr{L}$ параллельно $\mathscr{M}$*. Для проектора $P_{\mathscr{L},\mathscr{M}}$ подпространство $\mathscr{L}$ является образом, а $\mathscr{M}$ — ядром.

Фиксируя в $\mathscr{L}$ некоторый базис, мы однозначно определяем двойственный к нему базис в $\mathscr{M}^{\perp}$. Для соответствующих базисных матриц $A_{\mathscr{L}}$ и $B_{\mathscr{M}^{\perp}}$ выполняется условие

$$B^*_{\mathscr{M}^{\perp}}A_{\mathscr{L}}=I. \tag{1.7}$$

Отождествляя операторы с матрицами, можно дать явную формулу для проектора $P_{\mathscr{L},\mathscr{M}}$:

$$P_{\mathscr{L},\mathscr{M}}=A_{\mathscr{L}}B^*_{\mathscr{M}^{\perp}}. \tag{1.8а}$$

В частном случае, когда $\mathscr{M}=\mathscr{L}^{\perp}$, пишут $P_{\mathscr{L}}$ вместо $P_{\mathscr{L},\mathscr{M}}$ и говорят об *ортопроекторе*. Если $A_{\mathscr{L}}$ — базисная матрица с ортонормированными столбцами, то формула (1.8а) принимает вид

$$P_{\mathscr{L}}=A_{\mathscr{L}}A^*_{\mathscr{L}}. \tag{1.8б}$$

Проекторы, отвечающие неортогональным подпространствам $\mathscr{L}$ и $\mathscr{M}$, называют *косыми*. Любой проектор, косой или ортогональный, удовлетворяет соотношению

$$P^2=P, \tag{1.9}$$

которое при желании может быть принято за определение этого класса операторов (матриц).

Специальные классы матриц. Матрица A называется:

диагональной, если $a_{ij}=0$ при $i\neq j$;

трехдиагональной, если $a_{ij}=0$ при $|i-j|>1$;

ленточной, если для некоторого натурального числа m, называемого *полушириной ленты*, $a_{ij}=0$ при $|i-j|>m$. *Ширина ленты* — это число $2m+1$; в частности, трехдиагональная матрица — это ленточная матрица с шириной ленты 3;

правой (или *верхней*) *треугольной*, если $a_{ij}=0$ при $i>j$;

левой (или *нижней*) *треугольной*, если $a_{ij}=0$ при $i<j$;

правой (или *верхней*) *хессенберговой*, если $a_{ij}=0$ при $i>j+1$;

левой (или *нижней*) *хессенберговой*, если $a_{ij}=0$ при $j>i+1$.

Треугольная матрица с нулевыми диагональными элементами называется *строго треугольной*, а если все диагональные элементы равны единице, то *унитреугольной*. Словосочетания типа «верхняя треугольная» иногда заменяют более короткими *верхнетреугольная*, и т. п.

Специальное расположение спектра на комплексной плоскости характеризует следующие два класса матриц: *устойчивые* матрицы, все собственные значения которых принадлежат открытой левой полуплоскости, и *сходящиеся* матрицы, спектр которых заключен внутри единичного круга.

Определения ряда классов матриц связаны с операциями транспонирования и сопряжения. Квадратная матрица называется:

симметричной, если $A = A^{\mathrm{T}}$;
кососимметричной, если $A = -A^{\mathrm{T}}$;
ортогональной, если $A^{\mathrm{T}}A = AA^{\mathrm{T}} = I$;
эрмитовой, если $A = A^*$;
косоэрмитовой, если $A = -A^*$;
унитарной, если $A^*A = AA^* = I$;
нормальной, если $A^*A = AA^*$.

Вещественные симметричные, кососимметричные и ортогональные, а также комплексные эрмитовы, косоэрмитовы и унитарные матрицы являются нормальными. Матрицу, не являющуюся нормальной, условимся называть *анормальной*.

Множества ортогональных и унитарных матриц образуют группы по умножению, т. е. замкнуты относительно операций умножения и обращения.

Эрмитова матрица A *положительно* (*отрицательно*) *определена*, если $(Ax, x) > 0$ (соответственно $(Ax, x) < 0$) для любого $x \neq 0$ из $\mathbf{C}^n$. Если строгие неравенства заменить нестрогими, то получим *положительно* (*отрицательно*) *полуопределенные* матрицы.

Блочными называются матрицы

$$A = \begin{bmatrix} A_{11} & \dots & A_{1l} \\ \dots & \dots & \dots \\ A_{k1} & \dots & A_{kl} \end{bmatrix},$$

элементы A_{ij} которых в свою очередь являются матрицами. Если такая матрица рассматривается и как числовая, то ее обычные размеры следует отличать от блочных размеров $k \times l$. Как правило, мы будем иметь дело только с блочными матрицами, для которых $k = l$; в этом случае говорят о *блочном порядке* k.

Специальные классы квадратных блочных матриц выделяются, как и в скалярном случае, условиями равенства нулю тех или иных блоков. Например, если $A_{ij} = 0$ при $i \neq j$, получаем *блочно диагональную матрицу*

$$A = \begin{bmatrix} A_{11} & & 0 \\ & A_{22} & \\ & \ddots & \\ 0 & & A_{kk} \end{bmatrix}.$$

Матрицу такого вида часто называют *прямой суммой* матриц $A_{11}, \ldots, A_{kk}$ и используют запись $A = A_{11} \oplus \ldots \oplus A_{kk}$.

Подобным же образом можно ввести классы *блочно треугольных, блочно хессенберговых* и *блочно трехдиагональных* матриц.

Операции над блочными матрицами определяются таким же образом, как для матриц с числовыми элементами, при этом требуется согласование размеров соответствующих блоков матриц-операндов.

Некоторые специальные матрицы, часто используемые в вычислительной линейной алгебре, будут определены в § 2.

Спектральная теория. Число $\lambda \in \mathbf{C}$ и ненулевой вектор $x \in \mathbf{C}^n$ называются *собственным значением* и *собственным вектором* $n \times n$-матрицы A, если

$$Ax = \lambda x. \tag{1.10}$$

Чтобы подчеркнуть связь между λ и x, употребляют выражения типа «собственный вектор x относится к собственному значению λ», «собственное число λ отвечает собственному вектору x» и т. п. О совокупности λ, x будем говорить как о *собственной паре* матрицы A.

Собственные значения матрицы A суть корни ее *характеристического уравнения*

$$\det(\lambda I - A) = 0. \tag{1.11}$$

Его левая часть представляет собой многочлен от λ степени n — так называемый *характеристический многочлен* матрицы A. Поэтому всякая матрица порядка n имеет с учетом кратностей ровно n собственных значений. Нужно учитывать, что число *вещественных* собственных значений вещественной матрицы A может быть меньше n или даже равно нулю. Спектр вещественной матрицы обладает такой особенностью: наряду с каждым невещественным числом λ в него входит сопряженное число $\bar{\lambda}$.

Кратность собственного значения λ как корня характеристического многочлена называется его *алгебраической кратностью*. Собственные векторы, отвечающие числу λ, могут быть найдены как ненулевые решения однородной линейной системы

$$(\lambda I - A)x = 0. \tag{1.12}$$

Размерность подпространства решений системы (1.12) называется *геометрической кратностью* собственного значения λ. Геометрическая кратность никогда не превосходит алгебраическую.

Если алгебраическая кратность равна 1, собственное значение называют *простым*, в противном случае — *кратным*. Кратное собственное значение, для которого алгебраическая и геометрическая кратности совпадают, называется *полупростым* (этот необщепринятый термин заимствован в [24]). Если геометрическая кратность собственного значения больше 1, его называют *дефектным*.

Матрицы A и B называют *подобными*, если

$$B = P^{-1} A P \tag{1.13}$$

для некоторой невырожденной матрицы P. При этом про матрицу P говорят, что она *трансформирует* A в B. Само преобразование $A \to P^{-1}AP$ называют *подобием*, порождаемым матрицей P.

Из формулы (1.13) вытекает, что подобные матрицы A и B имеют одни и те же собственные значения, а из собственного вектора y матрицы B собственный вектор (для того же λ) матрицы A получается по формуле

$$x = Py.$$

Основная задача спектральной теории — это установление простейшей (*канонической*) формы, к которой может быть приведена матрица посредством подобий. Для большинства комплексных матриц такой формой будет диагональная. Назовем матрицу *простой* или *полупростой*, если все ее собственные значения соответственно простые или полупростые. Справедливо утверждение:

Всякая полупростая (и, в частности, всякая простая) матрица A может быть приведена к диагональному виду, т. е. найдется невырожденная матрица P такая, что

$$P^{-1}AP = \Lambda \tag{1.14}$$

и Λ — диагональная матрица. Диагональные элементы матрицы Λ — это собственные значения $\lambda_1, \ldots, \lambda_n$ матрицы A, что кратко можно выразить записью

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n). \tag{1.15}$$

Это утверждение объясняет другие, чаще используемые наименования полупростых матриц — *диагонализуемые* матрицы или матрицы *простой структуры*. Вместо «простая матрица» часто говорят «*матрица с простым спектром*».

Переписывая (1.14) в виде $AP = P\Lambda$ и приравнивая одноименные столбцы слева и справа, приходим к важному выводу:

Матрица P, которая трансформирует A к диагональному виду, составлена по столбцам из собственных векторов. Поэтому полупростым матрицам можно дать еще одно определение, а именно определить их как *матрицы, для которых в $\mathbf{C}^n$ существует базис из собственных векторов.*

Для неполупростой матрицы A приведение к диагональному виду невозможно, однако путем подобий можно получить так называемую *жорданову каноническую форму*, т. е. блочную матрицу вида

$$J = \begin{bmatrix} J_{m_1}(\lambda_1) & & & 0 \\ & J_{m_2}(\lambda_2) & & \\ & & \ddots & \\ 0 & & & J_{m_k}(\lambda_k) \end{bmatrix}, \tag{1.16}$$

где каждый из диагональных блоков $J_{m_i}(\lambda_i)$ есть *жорданова клетка порядка m_i, относящаяся к числу λ_i*:

$$J_{m_i}(\lambda_i)=\begin{bmatrix} \lambda_i & 1 & & & & \\ & \lambda_i & 1 & & 0 & \\ & & \lambda_i & \ddots & & \\ & & & \ddots & 1 & \\ & 0 & & & \lambda_i & 1 \\ & & & & & \lambda_i \end{bmatrix}. \tag{1.17}$$

Жорданова клетка первого порядка — это попросту диагональный элемент матрицы J. Числа $\lambda_1, \ldots, \lambda_k$ суть собственные значения матрицы A; некоторые из них могут совпадать, поэтому данному собственному значению в жордановой форме могут соответствовать несколько клеток разных порядков. Точное правило для числа клеток, отвечающих собственному значению λ: оно равно геометрической кратности λ, другими словами, *дефекту* (т. е. размерности ядра) матрицы $\lambda I - A$. Чтобы найти и порядки клеток (а не только их количество), нужны все различные числа последовательности дефектов

$$n_s = \operatorname{def}(\lambda I - A)^s, \qquad s = 1, 2, \ldots$$

Именно число входящих в J клеток порядка p для λ выражается формулой

$$S_p = 2n_p - n_{p+1} - n_{p-1}.$$

Последовательность $\{n_s\}$ монотонно неубывающая и, начиная с некоторого значения l, стабилизируется:

$$n_1 < \ldots < n_l = n_{l+1} = n_{l+2} = \ldots \tag{1.18}$$

Это число l называется *индексом* собственного значения λ и обозначается $\operatorname{ind}\lambda$. Оно указывает наибольший порядок жордановых клеток, отвечающих λ. Индекс полупростого λ равен 1.

Наибольший из индексов собственных значений будем называть *индексом матрицы*.

Матрицу, хотя бы одно собственное значение которой дефектно, называют *дефектной*. В противном случае матрица *недефектная*. Недефектные матрицы можно характеризовать тем свойством, что в их жордановой форме каждому собственному значению отвечает *ровно одна* жорданова клетка.

Очень важно понимать геометрическую картину, стоящую за задачей о приведении к канонической форме. В этой картине исходная матрица A интерпретируется, с одной стороны, как оператор $\mathscr{A}$, действующий в арифметическом пространстве, а с другой — как его представление в естественном базисе *координатных векторов* $e_1, \ldots, e_n$: $e_i = (0, \ldots, 0, \overset{i}{1}, 0, \ldots 0)^{\mathrm{T}}$. Напомним, что матрица A_e оператора $\mathscr{A}$ в базисе $\{e\}$ строится по столбцам из координат векторов $\mathscr{A}e_i$ ($i = 1, \ldots, n$) в этом базисе. Матрицы оператора $\mathscr{A}$ в базисах $\{e\}$ и $\{f\}$ подобны, причем трансформирующей матрицей служит матрица перехода от одного базиса к другому.

Подпространство $\mathscr{L}$ называется *инвариантным подпространством* оператора $\mathscr{A}$, если $\mathscr{A}\mathscr{L} \subset \mathscr{L}$. Рассматривая действие $\mathscr{A}$ только на векторы из $\mathscr{L}$, получаем оператор из $\mathscr{L}$ в $\mathscr{L}$; он называется *сужением* $\mathscr{A}$ на $\mathscr{L}$. Если взять в пространстве базис $\{f\}$, первые векторы которого образуют базис в $\mathscr{L}$ (пусть для определенности $\dim\mathscr{L}=l$), то матрица A_f имеет вид

$$A_f = \left[\begin{array}{ccc|c} b_{11} & \dots & b_{1l} & b_{1,l+1} \dots b_{1n} \\ \dots & \dots & \dots & \dots \\ b_{l1} & \dots & b_{ll} & b_{l,l+1} \dots b_{ln} \\ \hline & & & b_{l+1,l+1} \dots b_{l+1,n} \\ \multicolumn{3}{c|}{0} & \dots \\ & & & b_{n,l+1} \dots b_{nn} \end{array}\right]. \tag{1.19}$$

Таким образом, знание инвариантного подпространства позволяет упростить вид матрицы оператора. Канонические формы соответствуют наибольшему возможному упрощению. Диагональная форма отвечает ситуации, когда пространство можно представить в виде прямой суммы *собственных подпространств* $\mathscr{L}_\lambda$:

$$\mathbf{C}^n = \mathscr{L}_{\lambda_1} \oplus \dots \oplus \mathscr{L}_{\lambda_t}. \tag{1.20}$$

При этом $\mathscr{L}_\lambda$ определяется как множество всех решений системы (1.12), т. е., иными словами, как ядро матрицы $\lambda I - A$.

Для нормальной матрицы A подпространства $\mathscr{L}_{\lambda_i}$ попарно ортогональны. Эрмитовы (унитарные) матрицы характеризуются дополнительно тем, что все λ_i вещественны (равны по модулю 1).

Собственные подпространства, несомненно, являются наиболее удобными инвариантными подпространствами оператора (матрицы). Действие оператора на них сводится к равномерному растяжению с коэффициентом, равным соответствующему λ_i. Однако для недиагонализуемой матрицы A собственных подпространств недостаточно, чтобы породить $\mathbf{C}^n$. В этом случае вместо (1.20) получаем разложение

$$\mathbf{C}^n = K_{\lambda_1} \oplus \dots \oplus K_{\lambda_t}. \tag{1.21}$$

Корневое подпространство K_λ определяется как множество всех векторов x, для которых $(\lambda I - A)^p x = 0$ при некотором натуральном p. Векторы из K_λ называются *корневыми векторами*. Наименьшее число p в определении корневого вектора называется его *высотой*; в частности, собственные векторы — это корневые векторы высоты 1.

Если $l = \operatorname{ind}\lambda$, то из (1.18) следует, что

$$K_\lambda = \ker(\lambda I - A)^l. \tag{1.22}$$

С формулой разложения (1.21) тесно связана *теорема Гамильтона — Кэли*:

Всякая матрица A является корнем своего характеристического многочлена $\varphi(\lambda)$, т. е. $\varphi(A) = 0$.

Переход к базису пространства, получаемому объединением базисов корневых подпространств, дает *блочно диагональную* форму для матрицы оператора:

$$A_f = \begin{bmatrix} B_{11} & & & 0 \\ & B_{22} & & \\ & & \ddots & \\ 0 & & & B_{tt} \end{bmatrix}.$$

Дальнейшее упрощение, приводящее к жордановой форме, достигается выбором специальных базисов в корневых подпространствах. В совокупности они составляют *корневой базис* оператора (матрицы A). Корневой базис, отвечающий жордановой матрице (1.16), состоит из подсистем — так называемых *жордановых цепочек*, находящихся во взаимно однозначном соответствии с жордановыми клетками. Каждая цепочка начинается с собственного вектора; если λ_i — соответствующее собственное значение, то соседние векторы цепочки x_r и x_{r-1} связаны равенством

$$x_{r-1} = Ax_r - \lambda_i x_r.$$

Высота вектора x_{r-1} на единицу меньше высоты вектора x_r.

Если обратить порядок векторов в каждой цепочке, то новому базису соответствует другая разновидность жордановой формы, получаемая из (1.16) транспонированием.

Пусть P — невырожденная матрица, которая трансформирует A к жордановой форме (1.16):

$$P^{-1}AP = J. \tag{1.23}$$

Столбцы $p_1, \ldots, p_n$ матрицы P составляют корневой базис матрицы A, который мы будем (по причинам, которые сейчас выяснятся) называть *правым*. Переходя в обеих частях равенства (1.23) к сопряженным матрицам, получим

$$Q^{-1}A^*Q = J^*,$$

где положено $Q \equiv P^{-*}$. Таким образом, Q трансформирует сопряженную матрицу A^* тоже к жордановой форме, только к нижнетреугольной.

И в теории, и зачастую при практическом решении спектральных задач полезно учитывать наряду с правым корневым базисом $p_1, \ldots, p_n$ базис, образованный столбцами $q_1, \ldots, q_n$ матрицы Q. По отношению к матрице A этот последний называют *левым корневым базисом*. Названия «левый» и «правый» связаны с расположением матриц в (1.23); нужно иметь в виду, что P^{-1} с точностью до сопряжения совпадает с Q.

Из определения матрицы Q вытекает, что $Q^*P = I$. В соответствии с (1.5) это означает, что базисы $\{p\}$ и $\{q\}$ биортогональные. Тем самым с каждой комплексной матрицей ассоциированы два биортогональных корневых базиса.

Поскольку частями корневых базисов являются базисы отдельных корневых подпространств K_λ (*правых корневых подпространств*) и аналогичных подпространств K_λ^* для сопряженной матрицы (*левых корневых подпространств* матрицы A), то соотношения биортогональности приводят к важному заключению:

Правое корневое подпространство K_{λ_i} и левое корневое подпространство $K_{\lambda_j}^*$ ортогональны, если $\lambda_i \neq \lambda_j$.

Всякий правый (левый) корневой вектор можно включить в некоторый правый (левый) корневой базис. Поэтому два корневых вектора — правый и левый — ортогональны, если относятся к различным собственным значениям.

В случае простой матрицы направления собственных векторов (и правых, и левых) определены однозначно. После соответствующей нормировки, обеспечивающей условия $(p_i, q_i)=1$, правый и левый *собственные базисы* автоматически оказываются биортогональными.

С корневыми подпространствами связаны *спектральные проекторы*, играющие важную роль в теории. Спектральным проектором P_{λ_i}, ассоциированным с собственным значением λ_i, называется проектор на корневое подпространство K_{λ_i} параллельно прямой сумме прочих (правых) корневых подпространств. Так как в силу соотношений двойственности ортогональным дополнением к подпространству $\mathcal{M} = K_{\lambda_1} \oplus \ldots \oplus K_{\lambda_{i-1}} \oplus K_{\lambda_{i+1}} \oplus \ldots \oplus K_{\lambda_t}$ является левое корневое подпространство $K_{\lambda_i}^*$, то можно применить формулу (1.7). Нужно лишь выбрать в K_{λ_i} и $K_{\lambda_i}^*$ пару биортогональных базисов. Если, в частности, λ_i — простое собственное значение, p_i и q_i — соответствующие правый и левый собственные векторы, нормированные условием $(p_i, q_i)=1$, то

$$P_{\lambda_i} = p_i q_i^*. \tag{1.24}$$

Отметим, что в общем случае спектральные проекторы косые. Однако для класса нормальных матриц они являются ортопроекторами.

Разложению $\mathbf{C}^n$ в прямую сумму корневых подпространств (см. (1.21)) отвечает следующее тождество для спектральных проекторов:

$$I = P_{\lambda_1} + P_{\lambda_2} + \ldots + P_{\lambda_t}. \tag{1.25}$$

Теорема Шура. Эта теорема играет важную роль в вычислительной линейной алгебре, особенно при решении спектральных задач малых и средних размеров. Она заключается в следующем:

Всякую комплексную матрицу A посредством унитарного подобия (т. е. подобия с унитарной трансформирующей матрицей) можно привести к треугольному виду

$$P^* A P = \Delta.$$

Треугольная матрица Δ, называемая *формой Шура* матрицы A, содержит на диагонали ее собственные значения $\lambda_1, \ldots, \lambda_n$.

Значение теоремы Шура состоит именно в том, что подобие является унитарным. Это позволяет вычислить форму Шура численно устойчивым образом в отличие, например, от жордановой формы.

В то же время спектральная информация доставляется почти столь же богатая: собственные значения даны непосредственно, а собственные векторы могут быть без труда найдены.

Форма Шура определена неоднозначно. Прежде всего она может быть выбрана верхней или нижней треугольной матрицей, поэтому правильней говорить о *верхней* и *нижней формах Шура*. Далее для любого заданного упорядочения собственных значений найдется верхняя (нижняя) форма Шура, на диагонали которой числа $\lambda_1, .., \lambda_n$ расположены в нужной последовательности. Неоднозначность, вообще говоря, не исчезает и при фиксированном порядке собственных чисел и проявляется в значениях внедиагональных элементов. Это не относится к *нормальным матрицам*, *формы Шура которых являются диагональными матрицами*.

Если A — вещественная матрица, то полезно было бы иметь некоторый аналог формы Шура, к которому можно перейти посредством вещественных ортогональных преобразований. Таким аналогом является *вещественная форма Шура*, представляющая собой блочно треугольную матрицу с диагональными блоками 1×1 и 2×2. Блоки первого типа суть вещественные собственные значения матрицы A, блоки второго типа отвечают парам сопряженных комплексных собственных чисел.

Функции от матрицы. Простейшими функциями от матрицы являются многочлены. Если $\varphi(t)=\alpha_0+\alpha_1 t+...+\alpha_m t^m$, то многочлен $\varphi(A)$ от матрицы A задается формулой

$$\varphi(A)=\alpha_0 I+\alpha_1 A+...+\alpha_m A^m.$$

Пусть теперь $f(\lambda)$ — произвольная функция комплексной переменной λ, представимая в виде суммы степенного ряда

$$f(\lambda)=\sum_{i=0}^{\infty}\alpha_i(\lambda-\lambda_0)^i. \tag{1.26}$$

Если круг сходимости ряда (1.26) содержит внутри себя спектр матрицы A, то матричный ряд $\sum_{i=0}^{\infty}\alpha_i(A-\lambda_0 I)^i$ сходится и его сумма принимается в качестве $f(A)$. Однако оказывается, что при фиксированной матрице A всякая построенная посредством степенного ряда функция $f(A)$ совпадает с некоторым многочленом от A, а именно с *интерполяционным многочленом Лагранжа — Сильвестра* $\varphi(A)$. Если $\lambda_1, ..., \lambda_t$ — все различные собственные значения матрицы A и $n_1, ..., n_t$ — их алгебраические кратности, то многочлен Лагранжа — Сильвестра $\varphi(t)$ однозначно определяется условиями

$$\varphi(\lambda_1)=f(\lambda_1),\quad \varphi'(\lambda_1)=f'(\lambda_1),\ ...,\ \varphi^{(n_1-1)}(\lambda_1)=f^{(n_1-1)}(\lambda_1),$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\varphi(\lambda_t)=f(\lambda_t),\quad \varphi'(\lambda_t)=f'(\lambda_t),\ ...,\ \varphi^{(n_t-1)}(\lambda_t)=f^{(n_t-1)}(\lambda_t),$$

т. е. значениями функции $f(\lambda)$ и ее производных на спектре матрицы A.

Итак, если $f(\lambda)$ — аналитическая функция, для которой $f(A)$ имеет смысл, то: а) $f(A)$ коммутирует с A; б) $f(P^{-1}AP)=P^{-1}f(A)P$;

в) собственными значениями матрицы $f(A)$ являются числа $f(\lambda_1), \ldots, f(\lambda_t)$; г) всякая матрица P, трансформирующая A к треугольному виду, делает то же с матрицей $f(A)$.

Существуют и некоторые функции от матрицы (например, квадратные корни), не сводящиеся к многочленам.

Кронекерово произведение. *Кронекеровым произведением* $A \otimes B$ $k \times k$-матрицы A и $l \times l$-матрицы B называется блочная матрица

$$C = \begin{bmatrix} a_{11}B \ldots a_{1k}B \\ \ldots \ldots \ldots \\ a_{k1}B \ldots a_{kk}B \end{bmatrix}.$$

Ее порядок равен kl.

Из определения видно, что кронекерово умножение некоммутативно. Однако матрица $B \otimes A$ может быть получена из $A \otimes B$ симметричной перестановкой строк и столбцов. Если $\lambda_1, \ldots, \lambda_k$ и $\mu_1, \ldots, \mu_l$ — собственные значения матриц A и B, то спектр обоих кронекеровых произведений состоит из всевозможных чисел вида $\lambda_i \mu_j$.

Наряду с кронекеровыми произведениями часто приходится рассматривать *кронекерову разность* A и B: $D = A \otimes I_l - I_k \otimes B$. Ее спектр составляют всевозможные разности $\lambda_i - \mu_j$.

Если матрицы A и B нормальные, то нормальными являются и матрицы C, D.

§ 2. Сведения из вычислительной линейной алгебры

Нормы векторов и матриц. *Нормой* в $\mathbf{R}^n$ или $\mathbf{C}^n$ (*векторной нормой*) называется всякая числовая функция от вектора со свойствами:

1) $\|x\| \geqslant 0$; $\|x\| = 0 \Leftrightarrow x = 0$;
2) $\|\alpha x\| = |\alpha| \, \|x\|$;
3) $\|x + y\| \leqslant \|x\| + \|y\|$.

Здесь x, y — произвольные векторы, α — произвольное число.

Наиболее употребительны следующие три нормы ($\xi_1, \ldots, \xi_n$ — компоненты вектора x):

$$\|x\|_1 = |\xi_1| + \ldots + |\xi_n|; \tag{2.1}$$

$$\|x\|_2 = (|\xi_1|^2 + \ldots + |\xi_n|^2)^{1/2}; \tag{2.2}$$

$$\|x\|_\infty = \max_{1 \leqslant i \leqslant n} |\xi_i|. \tag{2.3}$$

Это три представителя семейства *гельдеровых* норм

$$\|x\|_p = (|\xi_1|^p + \ldots + |\xi_n|^p)^{1/p}, \qquad 1 \leqslant p \leqslant \infty.$$

Норма $\|x\|_2$ есть обычная евклидова длина вектора x.

Векторная норма называется *абсолютной*, если зависит лишь от *модулей* компонент вектора. Это свойство эквивалентно *монотонности*, что означает: если для векторов x и $y = (\eta_1, \ldots, \eta_n)^T$ справедливы неравенства $|\xi_i| \leqslant |\eta_i|$ $(i = 1, \ldots, n)$, то $\|x\| \leqslant \|y\|$.

Единичной сферой пространства с некоторой фиксированной нормой называется множество векторов нормы единица.

Норма на пространстве $m \times n$-матриц, рассматриваемом как векторное пространство размерности mn, называется *обобщенной матричной нормой*. Если, кроме того, $\| AB \| \leqslant \| A \| \| B \|$ для любых двух матриц A и B, допускающих произведение, то матричная норма называется *мультипликативной*.

В случае квадратных матриц свойство мультипликативности называют также *кольцевым свойством* нормы.

При одновременном рассмотрении матриц и векторов важно, чтобы используемые для них нормы были *согласованы*, т. е. чтобы выполнялось неравенство

$$\| Ax \| \leqslant \| A \| \| x \| \tag{2.4}$$

для любой матрицы A и любого вектора x. Если векторная норма $\| \cdot \|$ выбрана, то согласованную с ней матричную норму можно получить по правилу

$$\| A \| = \max_{x \neq 0} \frac{\| Ax \|}{\| x \|}. \tag{2.5}$$

Норма этого типа называется *операторной* или же матричной нормой, *подчиненной* векторной норме $\| x \|$. Всякая подчиненная норма мультипликативна; подчиненная норма единичной матрицы равна 1.

Абсолютная векторная норма может быть дополнительно охарактеризована таким свойством подчиненной ей матричной нормы: для любой диагональной матрицы $D = \operatorname{diag}(d_1, \ldots, d_n)$ справедливо равенство

$$\| D \| = \max_{1 \leqslant i \leqslant n} | d_i |.$$

Иногда в определении (2.5) используют разные нормы для векторов x и $y = Ax$. В этом случае говорят, что матричная норма *подчинена паре векторных норм*. Такая матричная норма уже не обязана быть мультипликативной, и норма единичной матрицы может быть не равна единице.

Широко используются матричные нормы, подчиненные нормам (2.1) — (2.3). Они помечаются теми же знаками 1, 2, ∞. Первая и третья из них легко могут быть вычислены по элементам $m \times n$-матрицы A:

$$\| A \|_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{m} | a_{ij} |, \tag{2.6}$$

$$\| A \|_\infty = \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{n} | a_{ij} |. \tag{2.7}$$

Норма $\| A \|_2$ называется *спектральной*, что объясняется связью этой нормы с наибольшим собственным значением положительно полуопределенных матриц A^*A и AA^*:

$$\| A \|_2 = \left[\lambda_{\max}(A^*A)\right]^{1/2} = \left[\lambda_{\max}(AA^*)\right]^{1/2}. \tag{2.8}$$

В дальнейшем нам потребуется следующий факт:

Спектральная норма матрицы не меньше евклидовой длины любого ее столбца или строки.

Из норм, не являющихся подчиненными, чаще всего используется *евклидова норма* матриц

$$\| A \|_E = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} | a_{ij} |^2 \right)^{1/2}.$$

Она тоже представляет семейство матричных гельдеровых норм

$$\| A \|_{l_p} = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} | a_{ij} |^p \right)^{1/p}, \qquad p \geqslant 1,$$

так что $\| A \|_E = \| A \|_{l_2}$.

Спектральная и евклидова нормы принадлежат к числу *унитарно инвариантных* матричных норм, т. е. норм, не меняющихся при умножении данной матрицы — одностороннем или двустороннем — на произвольные унитарные матрицы. Между собой эти нормы связаны неравенствами

$$\| A \|_2 \leqslant \| A \|_E \leqslant \min\left(\sqrt{m}, \sqrt{n}\right) \| A \|_2.$$

Левое неравенство показывает, что евклидова норма матриц согласована с евклидовой нормой векторов.

Матричные разложения. *Треугольным разложением* квадратной матрицы A называется представление

$$A = LR, \tag{2.9}$$

где L и R — соответственно нижняя и верхняя треугольные матрицы. Треугольное разложение существует не для всякой матрицы. Если A не вырождена, то необходимым и достаточным условием существования треугольного разложения является отличие от нуля *ведущих главных миноров*

$$a_{11}, \quad \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \ \ldots$$

В случае существования треугольное разложение определено неединственным образом: диагональным элементам одного из треугольных множителей могут быть присвоены произвольные ненулевые значения, тогда все остальное находится уже однозначно. Чаще всего для выделения единственного разложения требуют, чтобы сомножитель L был унитреугольным.

Ортогонально-треугольным разложением (или, короче, QR-*разложением*) называется представление квадратной матрицы A в виде

$$A = QR, \tag{2.10}$$

где Q — ортогональная матрица, а R — верхнетреугольная матрица. QR-разложение существует для любой вещественной матрицы и может

быть построено с помощью вращений или отражений (см. ниже). Если A — комплексная невырожденная матрица, то для существования ортогонально-треугольного разложения необходимо и достаточно, чтобы были ненулевыми ведущие главные миноры матрицы $A^{\mathrm{T}}A$. Однако для комплексных матриц всегда существует *унитарно-треугольное* разложение, т. е. представление произведением унитарной и верхнетреугольной матриц.

Вытекающее из (1.23) представление

$$A = PJQ^*, \tag{2.11}$$

где J — жорданова форма для A и $Q^* = P^{-1}$, называют *спектральным разложением* матрицы A. Если A — нормальная матрица, то J нужно заменить диагональной матрицей Λ, а P можно выбрать унитарной. Тогда спектральное разложение принимает вид

$$A = P\Lambda P^*. \tag{2.12}$$

Это представление, использующее унитарные и диагональные сомножители, можно обобщить на случай анормальных матриц и, более того, на прямоугольные матрицы. Именно любую $m \times n$-матрицу A можно представить произведением

$$A = USV^*, \tag{2.13}$$

где U и V — унитарные матрицы соответственно порядка m и n, а S — «диагональная» $m \times n$-матрица (т. е. матрица, для которой $s_{ij} = 0$ при $i \neq j$) с неотрицательными диагональными элементами $s_1, \ldots, s_n$ (s_m). Без ограничения общности числа $s_1, \ldots, s_n$ можно считать упорядоченными, например, по убыванию:

$$s_1 \geqslant s_2 \geqslant \ldots \geqslant s_n \geqslant 0.$$

Представление (2.13) называется *сингулярным разложением* матрицы A, числа $s_1, \ldots, s_n$ — *сингулярными числами*, столбцы $u_1, \ldots, u_m$ и $v_1, \ldots, v_n$ матриц U и V — *левыми* и *правыми сингулярными векторами*. Два последних термина объясняются соотношениями

$$Av_i = s_i u_i, \quad A^* u_i = s_i v_i, \qquad i = 1, \ldots, \qquad p = \min(m, n).$$

Сингулярные числа и векторы имеют следующий смысл: $s_1^2, \ldots, s_n^2$ и $v_1, \ldots, v_n$ суть собственные значения и собственные векторы матрицы A^*A; $s_1^2, \ldots, s_m^2$ и $u_1, \ldots, u_m$ суть собственные значения и собственные векторы матрицы AA^*.

Количество ненулевых сингулярных чисел равно рангу r матрицы. Если $r < \min(m, n)$, то сингулярному разложению можно придать более экономную форму

$$A = \hat{U}\hat{S}\hat{V}^*, \tag{2.14}$$

где $\hat{S} = \mathrm{diag}(s_1, \ldots, s_r)$; $\hat{U}$ — $m \times r$-, а V — $n \times r$-матрица с ортонормированными столбцами. Нередко сингулярными числами называют именно ненулевые числа $s_1, \ldots, s_r$.

Сингулярные числа матрицы не меняются при умножении ее — слева, справа или с обеих сторон — на произвольные унитарные

матрицы. Унитарная инвариантность спектральной и евклидовой матричных норм объясняется как раз их связью с сингулярными числами:

$$\| A \|_2 = \max_{1 \leqslant i \leqslant r} s_i, \qquad \| A \|_E = \left(s_1^2 + \ldots + s_r^2\right)^{1/2}. \tag{2.15}$$

Вспомогательные матрицы. В этом разделе перечисляются матрицы, используемые в вычислительной линейной алгебре для выполнения стандартных операций, таких как перестановки строк и столбцов заданной матрицы или аннулирование в ней указанных элементов.

Матрицы-перестановки. Так называются простейшие ортогональные матрицы, характеризуемые следующим свойством: в каждой строке и каждом столбце матрицы этого типа отличен от нуля ровно один элемент; его численное значение равно 1. Результатом умножения матрицы-перестановки $\mathscr{P}$ на вектор x будет соответствующая перестановка компонент вектора. Перестановку строк матрицы A, после которой расположение строк — в исходной нумерации — задается списком (другими словами, обычной *перестановкой из n чисел*) $\sigma_1, \ldots, \sigma_n$, можно описать как умножение A слева на следующую матрицу-перестановку $\mathscr{P} = [p_{ij}]$:

$$p_{ij} = \begin{cases} 1, & \text{если } j = \sigma_i, \\ 0 & \text{в противном случае.} \end{cases}$$

Если мы хотим в том же порядке $\sigma_1, \ldots, \sigma_n$ поставить *столбцы* квадратной матрицы A, то следует вычислить произведение $A\mathscr{P}^{\mathrm{T}}$. Простейшей перестановкой множества $\{1, \ldots, n\}$ является транспозиция чисел k и l. Отвечающую ей матрицу-перестановку будем обозначать через $\mathscr{P}_{kl}$; она отличается от единичной матрицы только четырьмя элементами: $p_{kk} = p_{ll} = 0$, $p_{kl} = p_{lk} = 1$. Умножение на матрицу $\mathscr{P}_{kl}$ вызывает перестановку k-й и l-й строк или k-го и l-го столбцов.

Элементарные матрицы. К этому классу относят матрицы, описывающие элементарные преобразования строк или столбцов заданной матрицы: их перестановки, умножение на ненулевые числа, прибавление к одной строке (или столбцу) кратного другой строки (столбца). О матрицах-перестановках рассказано в подразд. «Матрицы-перестановки». Умножение i-й строки матрицы A на число α равносильно вычислению произведения $\mathscr{D}_i(\alpha) A$, где диагональная матрица $\mathscr{D}_i(\alpha)$ отличается от единичной матрицы i-м диагональным элементом, равным α. Точно так же матрица $A\mathscr{D}_i(\alpha)$ получается из A умножением i-го столбца на число α.

Прибавление к i-й строке матрицы A ее j-й строки, умноженной на число n_{ij}, можно описать как результат матричного умножения $\mathscr{N}_{ij} A$, где $\mathscr{N}_{ij}$ — следующая элементарная матрица (для определенности считаем, что $i > j$):

$$\mathcal{N}_{ij} = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & \vdots & \ddots & & & \\ & & n_{ij} & \cdots & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix}.$$

(Необозначенные здесь внедиагональные элементы равны нулю.)

При вычислении произведения $A\mathcal{N}_{ij}$, напротив, к j-му столбцу добавляется i-й столбец, умноженный на n_{ij}; остальные столбцы не изменяются.

Из матриц $\mathcal{N}_{ij}$ формируются матрицы, также называемые элементарными, которые описывают более сложные операции над строками и столбцами. Так, прибавление к строкам 2, 3, ..., n строки 1, умноженной соответственно на числа $l_{21}, l_{31}, \ldots, l_{n1}$, можно представить посредством цепочки (левых) умножений на матрицы $\mathcal{N}_{21}, \mathcal{N}_{31}, \ldots, \mathcal{N}_{n1}$, а можно изобразить компактней — как результат единственного умножения на матрицу вида

$$\mathcal{L}_1 = \begin{bmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & 0 & 1 & & \\ \cdots & \cdots & \cdots & \cdots & \\ l_{n1} & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Полезно заметить, что: 1) матрица $\mathcal{L}_1^{-1}$ отличается от $\mathcal{L}_1$ только знаками поддиагональных элементов; 2) чтобы получить из матрицы A матрицу $A\mathcal{L}_1$, нужно к 1-му столбцу матрицы A добавить все остальные, умноженные соответственно на числа $l_{21}, l_{31}, \ldots, l_{n1}$; столбцы 2, 3, ..., n-й при этом не меняются.

Вращения и элементарные унитарные матрицы. Матрицы вращений (или просто *вращения*) — это вещественные матрицы вида

$$\mathcal{R}_{ij} = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & c & \cdots & -s & & \\ & & & \ddots & & & \\ & & s & \cdots & c & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \begin{matrix} \\ \\ i \\ \\ j \\ \\ \\ \end{matrix}, \quad c^2 + s^2 = 1. \tag{2.16}$$

(Необозначенные здесь внедиагональные элементы равны нулю.) Вращением такая матрица называется потому, что описывает поворот плоскости координатных векторов e_i, e_j на угол φ, однозначно определяемый соотношениями $\cos\varphi = c$, $\sin\varphi = s$, $0 \leqslant \varphi < 2\pi$.

Обычно вращения используются для аннулирования элементов преобразуемой матрицы A. Именно подходящим выбором угла φ можно добиться обращения в нуль любого элемента i-й либо j-й строки матрицы $\tilde{A}=\mathscr{R}_{ij}A$. Действительно, в матрице $\tilde{A}$ изменятся по сравнению с A только элементы этих двух строк, причем

$$\tilde{a}_{ik}=ca_{ik}-sa_{jk}, \quad \tilde{a}_{jk}=sa_{ik}+ca_{jk}, \qquad k=1, \ldots, n.$$

Пусть для определенности нужно аннулировать элемент $\tilde{a}_{jl}$. Тогда можно положить

$$c=a_{il}/(a_{il}^2+a_{jl}^2)^{1/2}, \qquad s=-a_{jl}/(a_{il}^2+a_{jl}^2)^{1/2}, \tag{2.17}$$

и требование $\tilde{a}_{jl}=0$ будет удовлетворено; при этом $\tilde{a}_{il}=(a_{il}^2+a_{jl}^2)^{1/2}$.

В комплексном случае для той же цели применяются элементарные унитарные матрицы. Они отличаются от вращений только видом своего *опорного квадрата*, т. е. четырьмя элементами, стоящими на пересечении строк и столбцов с номерами i и j:

$$\begin{bmatrix} ce^{i\varphi_1} & -se^{i\varphi_2} \\ se^{i\varphi_3} & ce^{i\varphi_4} \end{bmatrix}. \tag{2.18}$$

Здесь по-прежнему c и s — вещественные числа, причем $c^2+s^2=1$; кроме того, должно выполняться равенство $\varphi_1-\varphi_2=\varphi_3-\varphi_4$, обеспечивающее унитарность матрицы. Мы сохраним для элементарных унитарных матриц обозначение $\mathscr{R}_{ij}$.

Если A — комплексная матрица и снова требуется аннулировать элемент $\tilde{a}_{jl}$, то можно вместо (2.17) взять значения

$$c=|a_{il}|/(|a_{il}|^2+|a_{jl}|^2)^{1/2}, \qquad s=|a_{jl}|/(|a_{il}|^2+|a_{jl}|^2)^{1/2}$$

и положить, кроме того, $\varphi_1=-\arg a_{il}$, $\varphi_2=\pi-\arg a_{jl}$, $\varphi_3=-\varphi_2$, $\varphi_4=-\varphi_1$. Снова $\tilde{a}_{jl}=0$, $\tilde{a}_{il}=(|a_{il}|^2+|a_{jl}|^2)^{1/2}$.

Отражения. Матрицами Хаусхолдера или *отражениями* называются матрицы вида

$$\mathscr{H}=I-2\beta^2 uu^*. \tag{2.19}$$

Здесь u — ненулевой вектор-столбец, β — скаляр и $\beta^{-1}=\|u\|_2$. Матрица $\mathscr{H}$ эрмитовая и унитарная; поэтому $\mathscr{H}^{-1}=\mathscr{H}^*=\mathscr{H}$. Название этого класса матриц связано с тем, что вещественная $n\times n$-матрица (2.19) задает в $\mathbf{R}^n$ ортогональное отражение относительно гиперплоскости, проходящей через начало координат и имеющей нормальный вектор u.

Как и вращения, отражения используются для аннулирования элементов преобразуемой матрицы. Однако в отличие от вращений посредством одного отражения можно обратить в нуль целую группу элементов некоторой строки или столбца. Рассмотрим, например, следующую задачу: построить отражение $\mathscr{H}$ так, чтобы в l-м столбце матрицы $\tilde{A}=\mathscr{H}A$ элементы с $(m+1)$-го по n-й стали нулями, а первые $m-1$ элементов не изменились. Для упрощения записей предположим,

что матрица A вещественная. Тогда решение задачи дают две матрицы $\mathscr{H}$, порождаемые векторами вида

$$\{u\}_i=\begin{cases}0, & i=1, \ldots, m-1,\\ a_{ml}\mp S, & i=m,\\ a_{il}, & i=m+1, \ldots, n,\end{cases} \tag{2.20}$$

где $S=(a_{ml}^2+a_{m+1,l}^2+\ldots+a_{nl}^2)^{1/2}$. Полагая $K^2=(2\beta^2)^{-1}$, имеем

$$K^2=S(S\mp a_{ml})=\mp\{u\}_m S. \tag{2.21}$$

Из соображений численной устойчивости рекомендуется выбор знака перед S, увеличивающий абсолютное значение компоненты $\{u\}_m$. После преобразования значение элемента $\tilde{a}_{ml}$ равно S или $-S$, причем знак противоположен знаку, выбранному в (2.20).

Аналогичные выкладки можно провести в комплексном случае.

Построенная матрица $\mathscr{H}$ фактически осуществляет отражение в подпространстве размерности $n-m+1$, определяемом последними компонентами вектор-столбцов. Ее можно представить в блочном виде:

$$\mathscr{H}=\begin{bmatrix}I_{m-1} & 0\\ 0 & \mathscr{H}_{n-m+1}\end{bmatrix}, \tag{2.22}$$

где $\mathscr{H}_{n-m+1}$ — отражение порядка $n-m+1$. Число $n-m+1$ естественно называть *реальной размерностью отражения*. Компоненту $\{u\}_m$ называют *главной*.

Произведение $\tilde{A}=\mathscr{H}A$ вычисляют по столбцам, находя произведения $\mathscr{H}a$, где в качестве a последовательно берут 1-й, ..., $(l-1)$-й, ..., $(l+1)$-й, ..., n-й столбцы матрицы A. Произведение $\mathscr{H}a$ в соответствии с формулой

$$\mathscr{H}a=a-\frac{1}{K^2}u(u^*a)$$

вычисляется так: сначала определяем число $\delta=u^*a$, затем — число $\gamma=\delta/K^2$, наконец — вектор $a-\gamma u$. Естественно, что с нулевыми компонентами вектора u и одноименными компонентами вектора a никаких действий производить не нужно. Столбец l матрицы $\tilde{A}$ вычисления не требует: наперед известно, что последние $n-m$ его элементов должны быть нулями, а элемент $\tilde{a}_{ml}$ равен S или $-S$. Суммарное число операций умножения при вычислении произведения $\mathscr{H}A$ составляет приблизительно $2n(n-m+1)$. Сопоставим это с n^3 операциями, необходимыми при перемножении двух $n\times n$-матриц общего вида.

Обусловленность вычислительной задачи. Под *обусловленностью вычислительной задачи* понимают чувствительность ее решения к малому изменению входных данных задачи. Для того чтобы характеризовать обусловленность количественно, вводят ту или иную *меру*, или *число обусловленности*. Задачи с малыми значениями числа обусловленности называют *хорошо обусловленными*, задачи с большими числами обусловленности — *плохо обусловленными*.

Числа обусловленности для спектральных задач будут введены в гл. 2. Здесь мы проиллюстрируем понятие числа обусловленности на примере важной задачи решения линейной системы

$$Ax=b \tag{2.23}$$

с квадратной невырожденной матрицей A. Рассмотрим наряду с (2.23) *возмущенную систему*

$$(A+F)x=b+g \tag{2.24}$$

с малыми *возмущениями* — матрицей F и вектором g. Фиксируем некоторую операторную норму для матриц, и пусть $\|A^{-1}F\|<1$. Тогда матрица $\tilde{A}=A+F$ не вырождена, а для решений x и $\tilde{x}$ систем (2.23) и (2.24) выполняется неравенство

$$\frac{\|\tilde{x}-x\|}{\|x\|}\leqslant\frac{\|A\|\|A^{-1}\|}{1-\|A^{-1}F\|}\left(\frac{\|F\|}{\|A\|}+\frac{\|g\|}{\|b\|}\right). \tag{2.25}$$

Доказательство этой оценки можно найти в [8]. Практическое значение она имеет лишь при более сильном, чем до сих пор, ограничении на величину возмущения F, а именно $\|A^{-1}F\|\ll 1$. Достаточно, например, чтобы $\|A^{-1}F\|<0.1$. В этом случае можно игнорировать знаменатель первого сомножителя правой части. Тогда оценка (2.25) показывает, что относительная ошибка вектора $\tilde{x}$, рассматриваемого как приближенное решение системы (2.23), ограничена произведением числа $\|A\|\|A^{-1}\|$ на сумму относительных ошибок матрицы и правой части системы. Тем самым число $\|A\|\|A^{-1}\|$ играет роль максимального возможного коэффициента усиления (относительной) ошибки от входной информации к решению. Его и можно принять в качестве числа обусловленности для задачи решения системы $Ax=b$. Поскольку от вектора b оно не зависит, чаще говорят о числе обусловленности матрицы A; его обозначают $\operatorname{cond} A$. Если нужно явно указать, какой нормой порождается число обусловленности, его символ помечается соответствующим знаком. Так, $\operatorname{cond}_2 A=\|A\|_2\|A^{-1}\|_2$ есть *спектральное число обусловленности.*

Так как $1=\|I\|=\|AA^{-1}\|\leqslant\|A\|\|A^{-1}\|$, то число обусловленности не может быть меньше 1. Поэтому хорошо обусловленными следует считать матрицы (или линейные системы), числа обусловленности которых близки к 1. В случае спектральной нормы наилучшими в смысле обусловленности нужно считать унитарные матрицы и кратные им матрицы: для таких (и только таких) матриц спектральное число обусловленности равно 1.

Иногда для оценки обусловленности матрицы используют неоператорные нормы, например евклидову норму.

Понятие числа обусловленности можно распространить на прямоугольные матрицы; для них вместо обратной матрицы используют *псевдообратную матрицу Мура — Пенроуза* (см. [8])

$$\operatorname{cond} A=\|A\|\|A^{+}\|. \tag{2.26}$$

Спектральное число обусловленности можно выразить через сингулярные числа матрицы A. Если s_1 — наибольшее ненулевое сингулярное число, а s_r наименьшее, то $\operatorname{cond}_2 A=s_1/s_r$.

Устойчивость численного алгоритма. Это понятие мы опять-таки проиллюстрируем на примере решения линейной системы. Однако то же понимание устойчивости распространяется на другие классы алгебраических задач.

Метод $\mathcal{M}$ решения линейных систем называют *численно устойчивым*, если для любой системы (2.23) или, во всяком случае, для любой достаточно хорошо обусловленной системы вычисленное методом $\mathcal{M}$ решение $\tilde{x}$ можно представить как *точное* решение возмущенной системы (2.24), причем для матрицы F и вектора g, называемых *эквивалентными возмущениями метода*, справедливы оценки типа

$$\| F \| \leqslant f(n) \| A \| \beta^{-t},$$
$$\| g \| \leqslant h(n) \| b \| \beta^{-t}. \qquad (2.27)$$

Здесь β — основание машинной арифметики, t — число разрядов машинного слова, отводимых для представления мантиссы, $f(n)$ и $h(n)$ — некоторые медленно растущие функции от n типа степенных Cn^k с небольшим показателем k.

Существуют определения устойчивости более общие, чем приведенное. Поэтому ради точности определенную выше устойчивость следовало бы называть *устойчивостью по Уилкинсону*. Для целей настоящей книги этого определения вполне достаточно, и мы будем говорить просто об устойчивости.

Разное. В этом разделе собраны определения, которым не нашлось естественного места под другими рубриками.

Наибольший среди модулей собственных значений матрицы A называется ее *спектральным радиусом* и обозначается через $\rho(A)$. Всякое собственное значение с модулем $\rho(A)$ называют *старшим* или *доминирующим собственным значением* матрицы A. Старшим или доминирующим будем называть и соответствующий собственный вектор. Для невырожденной матрицы A используется и термин «младшее собственное значение», подразумевающий собственное значение с наименьшим модулем.

Матрица A *разложима*, если приводится к блочно-треугольному виду посредством симметричной перестановки строк и столбцов, т. е. найдется матрица-перестановка $\mathcal{P}$ такая, что

$$\mathcal{P}^{\mathrm{T}} A \mathcal{P} = \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix}. \qquad (2.28)$$

В противном случае A — *неразложимая* матрица. Заметим, что, хотя в (2.28) изображена верхняя блочно-треугольная матрица, разложимую матрицу A при желании можно привести и к нижнему блочно-треугольному виду.

Хессенбергову матрицу H (для определенности — верхнюю) мы будем называть неразложимой, если все элементы ее *кодиагонали* (2,1), (3,2), ..., $(n, n-1)$ не равны нулю. Это толкование неразложимости согласовано с предыдущим, если наддиагональные элементы в H (или хотя бы только элемент h_{1n}) ненулевые.

Пусть $\Gamma(A)$ — (ориентированный) граф матрицы A, т. е. граф с n узлами, помеченными числами 1, ..., n, в котором ребро $\{i, j\}$ присутствует тогда и только тогда, когда $a_{ij} \neq 0$. Разложимость матрицы A означает, что множество узлов ее графа можно разбить на непустые непересекающиеся подмножества C_1 и C_2 такие, что никакой узел из C_2 не связан ребром с узлом из C_1 (хотя ребра, идущие в обратном направлении — из C_1 в C_2, могут присутствовать). Неразложимость матрицы равносильна *сильной связности* графа $\Gamma(A)$. Последнее означает, что из каждого узла графа в *любой* другой ведет ориентированный путь (а из того в свою очередь — в первый).

Граф $\Gamma(A)$ называется *слабосвязным*, если из каждого его узла идет ориентированный путь в *некоторый* другой узел, а из того в свою очередь — в первый. Матрица A со слабосвязным графом $\Gamma(A)$ называется *слабонеразложимой*.

Матрица вида

$$C_f = \begin{bmatrix} -a_{n-1} & -a_{n-2} & \dots & -a_1 & -a_0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \tag{2.29}$$

называется *сопровождающей матрицей* многочлена $f(\lambda) = \lambda^{n-1} + a_{n-1}\lambda^{n-1} + \dots + a_1\lambda + a_0$ по той причине, что ее характеристический многочлен совпадает с $f(\lambda)$.

ГЛАВА 2. ТЕОРЕМЫ ЛОКАЛИЗАЦИИ И ТЕОРЕМЫ О ВОЗМУЩЕНИЯХ

Теоремой локализации называют утверждение, указывающее область комплексной плоскости, содержащую все собственные значения данной матрицы или хотя бы одно. Иногда такого рода теоремы называют также *теоремами включения* в отличие от *теорем исключения*, определяющих области, где, напротив, собственных значений нет. Разумеется, всякий результат об исключении есть теорема включения «наоборот», поскольку область, дополнительная к найденной, локализует спектр матрицы.

Теоремы о возмущениях описывают изменения собственных значений, собственных векторов и инвариантных подпространств под действием малых возмущений элементов матрицы. Многие теоремы включения могут рассматриваться и как теоремы о возмущениях собственных значений. Различие здесь в том, что сама по себе теорема локализации не накладывает ограничений на величину возмущений в матрице; она дает содержательные результаты и при «больших» возмущениях. В последнем случае область локализации является довольно грубой и используется главным образом для того, чтобы составить представление о характере распределения матричного спектра на комплексной плоскости. В теореме локализации, применяемой подобным образом, ценятся простота вида описываемой области и несложность вычислений, приводящих к этому описанию. Конечно, простота не должна достигаться ценой утери содержания: ведь и утверждение «спектр принадлежит комплексной плоскости» тоже теорема локализации, хотя мало полезная. Пример идеального сочетания простоты и содержательности дает самая известная из теорем локализации — теорема Гершгорина. Это классическое по форме утверждение было опубликовано [13], можно сказать, совсем недавно — в 1931 г. — ленинградским математиком С. А. Гершгориным. Точнее надо говорить не о теореме, а о двух теоремах Гершгорина; вторая теорема является уточнением первой (см. § 3). Мало какой результат в истории линейной алгебры вызывал столь обильную последующую литературу, как эти две теоремы. В частности, почти все теоремы локализации и теоремы о возмущениях собственных значений, рассматриваемые в § 3—5, возникли на этой основе.

§ 3. Теоремы Гершгорина и их обобщения

Чтобы иметь критерий содержательности локализационного результата, начнем со следующего простого утверждения.

Теорема 3.1. *Пусть* $\|\cdot\|$ — *произвольная мультипликативная матричная норма. Для всякого собственного значения* λ *матрицы*

А справедлива оценка

$$|\lambda| \leqslant \| A \|. \tag{3.1}$$

Доказательство. Пусть x — собственный вектор матрицы A, относящийся к числу λ. Для выбранной мультипликативной матричной нормы найдется согласованная с ней векторная норма (это утверждение разъясняется в п. 1 дополнений к § 3); примем для последней то же обозначение $\| \cdot \|$. Тогда из равенства $Ax = \lambda x$ следует

$$|\lambda| \| x \| = \| \lambda x \| = \| Ax \| \leqslant \| A \| \| x \|,$$

откуда после сокращения на ненулевое число $\| x \|$ получаем требуемое неравенство (3.1).

Замечание 3.2. С помощью спектрального радиуса $\rho(A)$ матрицы A неравенству (3.1) можно придать вид

$$\rho(A) \leqslant \| A \|. \tag{3.2}$$

Замечание 3.3. Область локализации, указываемая теоремой 3.1, это круг

$$|z| \leqslant \| A \|. \tag{3.3}$$

Пример 3.4. Применение к матрице

$$A = \begin{bmatrix} 1 & -2 & 3 & 4 \\ 2 & 1 & -1 & 0 \\ 1 & -2 & 0 & 1 \\ 1 & 1 & 2 & -1 \end{bmatrix} \tag{3.4}$$

оценки (3.3) с нормой $\| \cdot \|_\infty$ дает область локализации $|z| \leqslant 10$; выбор другой нормы — $\| A \|_1$ — позволяет уменьшить радиус области: $|z| \leqslant 6$.

Пример 3.5. Для всех собственных значений трехдиагональной матрицы

$$A = \begin{bmatrix} a_1 & b_2 & & & & \\ c_2 & a_2 & b_3 & & 0 & \\ & c_3 & a_3 & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & & a_{n-1} & b_n \\ & & & & c_n & a_n \end{bmatrix}$$

справедлива оценка

$$|\lambda| \leqslant \max_i \{ |a_i| + |b_{i+1}| + |c_i| \}, \qquad c_1 = b_{n+1} = 0.$$

Это следует из теоремы 3.1, если положить $\| A \| = \| A \|_\infty$.

Далее в тексте будут использоваться обозначения

$$R_i(A) = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \qquad C_i(A) = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ji}|, \qquad i = 1, \ \ldots, \ n. \tag{3.5}$$

Отступая от формальной правильности, будем называть эти числа соответственно *строчными* и *столбцевыми суммами*.

Теорема 3.6 (первая теорема Гершгорина). *Все собственные значения $n\times n$-матрицы A принадлежат области $G(A)$, представляющей собой объединение n кругов*

$$G_i(A)=\{z\,|\,|z-a_{ii}|\leqslant R_i(A)\},\qquad i=1,\ldots,n, \tag{3.6}$$

так называемых кругов Гершгорина матрицы A).

Доказательство. Пусть λ — произвольное собственное значение матрицы A; соответствующий собственный вектор обозначим через x. Пусть ξ_i — компонента вектора x, имеющая максимальный модуль. Если таких компонент несколько, то ξ_i — любая из них. Из равенства $Ax=\lambda x$ следует, в частности, что

$$(\lambda-a_{ii})\xi_i=\sum_{\substack{j=1\\ j\neq i}}^{n}a_{ij}\xi_j.$$

Поэтому

$$|\lambda-a_{ii}|\leqslant\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\frac{|\xi_j|}{|\xi_i|}\leqslant\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|=R_i(A). \tag{3.7}$$

Это и означает, что $\lambda\in G(A)$.

Замечание 3.7. Доказательство теоремы заодно показывает, какому именно кругу Гершгорина принадлежит собственное значение λ: $\lambda\in G_i(A)$, где i — номер компоненты с максимальным модулем соответствующего собственного вектора. При наличии нескольких максимальных компонент λ попадает в каждый из одноименных кругов Гершгорина. Правда, информации о подобных номерах i в практических ситуациях почти никогда не имеется.

Пример 3.8. Матрица (3.4) имеет следующие круги Гершгорина:

$$|z-1|\leqslant 9,\quad |z-1|\leqslant 3,\quad |z|\leqslant 4,\quad |z+1|\leqslant 4.$$

Область Гершгорина $G(A)$ совпадает с первым кругом $G_1(A)$: $|z-1|\leqslant 9$. Он вложен в круг $|z|\leqslant 10$, который дает теорема 3.1 при выборе нормы $\|\cdot\|_\infty$; отношение площадей этих двух кругов составляет около 0.8.

Как известно, матрицы A и A^{T} имеют одни и те же собственные значения. Поэтому, применяя теорему 3.6 к матрице A^{T}, получаем «столбцевой вариант» теоремы Гершгорина:

Все собственные значения $n\times n$-матрицы A принадлежат области $G(A^{\mathrm{T}})=\bigcup_{i=1}^{n}G_i(A^{\mathrm{T}})$, *где*

$$G_i(A^{\mathrm{T}})=\{z\,|\,|z-a_{ii}|\leqslant C_i(A)\},\qquad i=1,\ldots,n. \tag{3.8}$$

---

*) Говоря об области, мы отступаем от канонического определения. Область в подразумеваемом нами смысле есть просто более краткий синоним термина «замкнутое множество» и, вообще говоря, не является связной.

Это замечание применимо ко многим теоремам локализации. Замена A на A^{T} позволяет из строчной версии вывести столбцевую, и наоборот.

Пример 3.9. Для матрицы (3.4) столбцевая теорема Гершгорина дает круги $|z-1|\leqslant 4$, $|z-1|\leqslant 5$, $|z|\leqslant 6$, $|z+1|\leqslant 5$ и $G(A^{\mathrm{T}})$ совпадает с кругом $|z|\leqslant 6$. Ту же область мы получили в примере 3.4, беря в оценке (3.3) норму $\|\cdot\|_1$.

Пример 3.10. Предположим, что все диагональные элементы комплексной $n\times n$-матрицы A не равны нулю, причем

$$-\mathrm{Re}\,a_{ii} > R_i(A), \qquad i=1,\ldots,n. \tag{3.9}$$

В таком случае все круги Гершгорина принадлежат левой полуплоскости комплексной плоскости, и это же верно в отношении собственных значений. Другими словами, неравенства (3.9) суть достаточные условия устойчивости матрицы A.

Ценность теоремы Гершгорина как локализационного результата значительно увеличивает следующее утверждение.

Теорема 3.11 (вторая теорема Гершгорина). *Предположим, что область Гершгорина $G(A)$ распадается на попарно непересекающиеся связные подобласти C_1, $C_2,\ldots,C_k$ — так называемые связные компоненты области $G(A)$. Тогда каждой из связных компонент принадлежит столько собственных значений матрицы A, сколько кругов Гершгорина составляют эту компоненту.*

Для правильного применения этой теоремы нужно учитывать, что и собственные значения, и круги Гершгорина могут быть кратными. Так, в примере 3.9 круг $|z-1|\leqslant 5$ имеет кратность 2. При определении числа кругов, составляющих данную связную компоненту C_m, каждый круг засчитывается столько раз, какова его кратность; это же относится к подсчету числа собственных значений, попадающих в C_m.

Мы не приводим доказательства второй теоремы Гершгорина, поскольку его можно найти во многих книгах по теории матриц и вычислительным методам линейной алгебры (см., например, [4, 29, 45, 49]). Отметим только, что оно опирается на непрерывную зависимость собственных значений матрицы от ее элементов.

Пример 3.12. Для матрицы

$$A=\begin{bmatrix} 1.23 & 0.03 & 0.04 \\ 0.03 & 2.17 & 0.01 \\ 0.02 & 0.04 & 3.06 \end{bmatrix}$$

область Гершгорина образована кругами

$$|z-1.23|\leqslant 0.07, \quad |z-2.17|\leqslant 0.04, \quad |z-3.06|\leqslant 0.06. \tag{3.10}$$

Все три круга попарно не пересекаются, поэтому из второй теоремы Гершгорина вытекает такой вывод: диагональные элементы матрицы A можно рассматривать как приближения к ее собственным значениям; погрешности этих приближений не превышают соответственно 0.07, 0.04 и 0.06.

В приведенном рассуждении теоремы Гершгорина использованы, по существу, в духе теории возмущений: матрица A рассматривается

как малое возмущение диагональной матрицы $D=\operatorname{diag}(1.23,\ 2.17,\ 3.06)$ и оцениваются возмущения, претерпеваемые собственными значениями при переходе от D к A.

В действительности диагональные элементы матрицы A являются значительно более точными приближениями к ее собственным значениям, чем это следует из неравенств (3.10). Обосновать это можно с помощью тех же теорем Гершгорина, применяемых следующим образом. Умножим элементы первой строки матрицы A на положительное число α, а затем разделим на то же число элементы первого столбца. Такое преобразование матрицы представляет собой простейший вид подобия $A\to B=PAP^{-1}$; в роли трансформирующей матрицы P выступает $\operatorname{diag}(\alpha,\ 1,\ 1)$.

Заметим, что диагональные элементы матрицы B по-прежнему имеют значения a_{ii}. Подберем число α так, чтобы как можно больше уменьшить радиус первого круга Гершгорина, сохраняя его изолированность от двух других кругов. Значение α, для которого последнее условие впервые нарушается, определяется равенством

$$0.07\alpha+0.03/\alpha+0.01=0.94;$$

меньший из двух корней этого уравнения равен

$$\alpha_0=\frac{6}{93+\sqrt{93^2-84}},$$

откуда $\alpha_0>6/(93+93)\approx 0.032$.

Итак, после преобразования с матрицей $P=\operatorname{diag}(0.033,\ 1,\ 1)$ первый круг Гершгорина все еще не пересекается с двумя остальными. Согласно теореме 3.11, он содержит ровно одно собственное значение B, т. е. (матрицы A и B подобны!) ровно одно собственное значение матрицы A. Радиус же этого круга равен $0.07\times 0.033\approx 0.002$; таким образом, диагональный элемент $a_{11}=1.23$ отличается от некоторого собственного значения не более чем на 0.0021. Сходные оценки погрешности можно получить для диагональных элементов a_{22} и a_{33}.

Рассуждение, использованное нами в данном примере, носит на самом деле общий характер. Пусть имеем матрицу A с попарно различными диагональными элементами $a_{11},\ldots,a_{nn}$ и «малыми» внедиагональными элементами $a_{ij}(i\neq j)$. Последнее означает, что

$$\varepsilon=\max_{i\neq j}|a_{ij}|\ll d=\min_{k\neq l}|a_{kk}-a_{ll}|.$$

Тогда, рассуждая, как в примере, мы сможем показать, что при некоторой нумерации собственных значений $\lambda_1,\ldots,\lambda_n$ матрицы A справедливы неравенства

$$|\lambda_i-a_{ii}|\leqslant 2n(n-1)\varepsilon^2/d,\qquad i=1,\ldots,n. \tag{3.11}$$

Более подробно об этом, как и о случае, когда не все диагональные элементы a_{ii} различны, можно прочесть в § 2.14 книги [42].

Мы дали теореме 3.6 самостоятельное доказательство. Между тем она была получена Гершгориным как простое следствие критерия невырожденности квадратной матрицы, обычно называемого теоремой

Адамара. Как свидетельствует история вопроса, более правильное название — «теорема Леви — Деспланка» (это название и принято в книге [32, с. 192]).

Теорема 3.13 (теорема Леви — Деспланка). *Матрица A не вырождена, если*

$$|a_{ii}| > R_i(A), \qquad i = 1, \ldots, n. \tag{3.12}$$

Свойство (3.12) называют *диагональным преобладанием*. Иногда приходится пользоваться более развернутым выражением «диагональное преобладание по строкам», поскольку невырожденность матрицы A обеспечивают и неравенства

$$|a_{ii}| > C_i(A), \qquad i = 1, \ldots, n, \tag{3.13}$$

т. е. диагональное преобладание по столбцам.

Теорема 3.6 получается из теоремы Леви — Деспланка посредством следующего рассуждения, типичного для связи теорем о невырожденности матрицы с теоремами локализации. Пусть λ — собственное значение матрицы A. Тогда матрица $B = A - \lambda I$ вырождена и хотя бы одно из условий (3.12) должно быть для нее нарушено. Найдется, следовательно, номер i такой, что $|a_{ii} - \lambda| \leqslant R_i(A)$, а это и есть теорема Гершгорина.

Сходным образом можно извлекать результаты о локализации собственных значений из других условий невырожденности. Наоборот, если известно правило, сопоставляющее каждой матрице A область $\mathcal{D}(A)$, содержащую весь спектр $\sigma(A)$, то достаточное условие невырожденности можно сформулировать как требование $0 \notin \mathcal{D}(A)$.

Прибавим к сказанному, что всякое конкретное условие невырожденности и всякую теорему локализации можно «размножить» в бесконечном числе экземпляров, записывая их применительно не к самой матрице A, а к матрице AP или $P^{-1}AP$, где P — фиксированная невырожденная матрица.

Пример 3.14. Пусть

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \qquad P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Применяя к матрице AP условия диагонального преобладания (3.12), получаем следующий достаточный критерий невырожденности 2×2-матриц:

$$|a_{12}| > |a_{11}|, \qquad |a_{21}| > |a_{22}|. \tag{3.14}$$

Из неравенств (3.14), рассуждая, как и выше, можем получить следующую теорему типа Гершгорина: оба собственных значения произвольной 2×2-матрицы A принадлежат области $\mathcal{D}(A) = L_1(A) \cup L_2(A)$, где

$$L_1(A) = \{z \,|\, |a_{11} - z| \geqslant |a_{12}|\}, \qquad L_2(A) = \{z \,|\, |a_{22} - z| \geqslant |a_{21}|\}.$$

Каждое из множеств $L_i(A)$ есть «почти»-дополнение к одноименному кругу Гершгорина $G_i(A)$; общая граница входит и в $L_i(A)$, и в $G_i(A)$.

Любопытно сопоставить полученный локализационный результат с обычной теоремой Гершгорина в ситуации, когда круги $G_1(A)$ и $G_2(A)$ пересекаются (рис. 1). В этом случае собственные значения λ_1 и λ_2, с одной стороны, принадлежат объединению $G_1(A)\cup G_2(A)$; с другой стороны, находясь в $\mathscr{D}(A)$, λ_1 и λ_2 не могут попасть во внутренность пересечения $G_1(A)\cap G_2(A)$. Таким образом, область Гершгорина можно сузить до области, указанной на рис. 1 штриховкой.

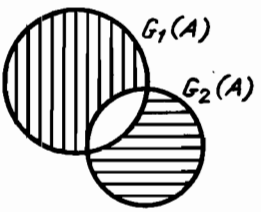

Рис. 1

Обобщая рассуждения примера 3.14, устанавливаем следующую теорему локализации [32, III, п. 2.2.9]. Пусть σ — перестановка чисел $1,\dots,n$; если $i\neq\sigma(i)$, полагаем

$$S_i(A)=2|a_{i\sigma(i)}|-R_i(A).$$

Тогда спектр $n\times n$-матрицы A заключен в объединении n областей

$$|z-a_{ii}|\leqslant R_i(A), \qquad i=\sigma(i),$$

$$|z-a_{ii}|\geqslant S_i(A), \qquad i\neq\sigma(i).$$

Матрица-перестановка P, используемая при доказательстве этого утверждения, определяется формулами $p_{ij}=\delta_{\sigma(i)j}$.

И теорема Леви — Десплanка, и более поздняя теорема Гершгорина породили массу литературы, посвященной уточнениям и обобщениям этих двух теорем. Мы рассмотрим сейчас усиление первой теоремы, принадлежащее Таусски [193]. Из него, как обычно, будет выведено следствие, относящееся к локализации собственных значений. В оставшейся части параграфа, а также в § 3—5 изучаются обобщения теорем Гершгорина в различных направлениях.

Теорема 3.15 (теорема Таусски). *Пусть для неразложимой $n\times n$-матрицы A выполнены ослабленные условия диагонального преобладания*

$$|a_{ii}|\geqslant R_i(A), \qquad i=1,\dots,n, \tag{3.15}$$

причем хотя бы в одном случае знак неравенства строгий. Тогда матрица A не вырождена.

Доказательство. Предположим, что A вырождена и x — некоторый ненулевой вектор из ее ядра. Можно считать, что среди компонент вектора x наибольший модуль имеют первые компоненты: $|\xi_1|=|\xi_2|=\dots=|\xi_k|$ $(k\geqslant 1)$. Если бы это было не так, то вместо равенства $Ax=0$ можно было бы рассматривать равенство $(PAP^{\mathrm{T}})(Px)=0$, где P — матрица-перестановка, обеспечивающая нужный порядок компонент вектора. Матрица $B=PAP^{\mathrm{T}}$ неразложима, и условия (3.15) для нее тоже выполнены.

Покажем, что в матрице A равны нулю все элементы a_{ij}, для которых $1\leqslant i\leqslant k$, $k<j\leqslant n$. Это будет означать, что A имеет вид

$$A=\begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix},$$

где A_{11} — матрица порядка k. Возможность такого представления противоречит условию неразложимости.

В самом деле, из $Ax=0$ выводим

$$a_{ii}\xi_i = -\sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}\xi_j,$$

откуда для $i=1,\ldots,k$ имеем

$$|a_{ii}| = \left|\sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}\frac{\xi_j}{\xi_i}\right| \leqslant \sum_{\substack{j=1\\ j\neq i}}^{k} |a_{ij}| + \sum_{j=k+1}^{n} |a_{ij}|\frac{|\xi_j|}{|\xi_i|}.$$

Так как $|\xi_j|<|\xi_i|$ при $j>k$, то

$$|a_{ii}| < \sum_{\substack{j=1\\ j\neq i}}^{k} |a_{ij}| + \sum_{j=k+1}^{n} |a_{ij}| = R_i(A),$$

если хотя бы один элемент $a_{ij}\,(k+1\leqslant j\leqslant n)$ отличен от нуля. Но это не согласуется с (3.15).

Заметим, что равенство $k=n$ невозможно, иначе $|a_{ii}|=R_i(A)$ для всех $i=1,\ldots,n$ вопреки третьему условию теоремы.

Следствие 3.16. *Каждое собственное значение неразложимой $n\times n$-матрицы A является либо внутренней точкой области Гершгорина $G(A)$, либо общей граничной точкой всех n кругов Гершгорина* (см. [32, п. 2.2.6]).

Доказательство. Если λ — собственное значение матрицы A, то для матрицы $A-\lambda I$ хотя бы одно из условий теоремы 3.15 должно быть нарушено. Это значит, что либо $|a_{ii}-\lambda|<R_i(A)$ при некотором i, т. е. λ находится внутри i-го круга Гершгорина, либо все условия (3.15) выполняются как равенства, т. е. λ принадлежит границе каждого круга.

Уточнения теорем Гершгорина появляются до сих пор, о чем свидетельствуют хотя бы недавние работы советских математиков Пупкова [38] и Соловьева [41]. В первой из них рассматривается ситуация, когда k-й круг Гершгорина изолирован от остальных, причем

$$T_k(A) = \max_{j\neq k} |a_{jk}| < R_k(A). \tag{3.16}$$

Из последнего условия вытекает, в частности, что строчная сумма $R_k(A)$ строго положительна.

Теорема 3.17 (В. А. Пупков). *В круге*

$$|z-a_{kk}| \leqslant T_k(A) \tag{3.17}$$

содержится ровно одно собственное значение матрицы A.

Доказательство. При любом $i\neq k$ справедливо неравенство

$$|a_{kk}-a_{ii}| > R_k(A)+R_i(A).$$

Пусть $1>\gamma>0$, и пусть $\mathscr{D}_k(\gamma)$ — элементарная диагональная матрица (см. § 2). Для матрицы $B=\mathscr{D}_k(\gamma)A\mathscr{D}_k^{-1}(\gamma)$ радиус k-го круга Гершгорина равен $\gamma R_k(A)$, а радиус i-го круга — числу $R_i(A)+|a_{ik}|(\gamma^{-1}-1)$.

Полагая

$$\gamma = T_k(A)/R_k(A), \tag{3.18}$$

имеем

$$R_i(B)+R_k(B)=R_i(A)+|a_{ik}|(\gamma^{-1}-1)+\gamma R_k(A)\leqslant$$
$$\leqslant R_i(A)+R_k(A)<|a_{kk}-a_{ii}|.$$

Среднее неравенство проверяется так:

$$R_k(A)-\left[|a_{ik}|(\gamma^{-1}-1)+\gamma R_k(A)\right]=$$
$$=R_k(A)(1-\gamma)-|a_{ik}|(1-\gamma)/\gamma=(1-\gamma)(R_k(A)-|a_{ik}|\gamma^{-1})\geqslant$$
$$\geqslant(1-\gamma)(R_k(A)-T_k(A)\gamma^{-1})=0.$$

Итак, k-й круг Гершгорина по-прежнему не пересекается с прочими кругами, откуда и следует утверждение теоремы.

Пусть λ — единственное собственное значение матрицы A, находящееся в круге (3.17). Согласно замечанию 3.7, k-я компонента собственного вектора x, относящегося к λ, должна иметь максимальный модуль. В условиях теоремы 3.17 можно сказать большее.

Теорема 3.18 (В. А. Пупков). *Для компонент* $\xi_1, \ldots, \xi_n$ *собственного вектора* x *выполняются соотношения*

$$|\xi_i|<\gamma|\xi_k|, \qquad i\neq k. \tag{3.19}$$

Значение γ *определяется формулой* (3.18).

Доказательство. Собственному значению λ матрицы $B=\mathscr{D}_k(\gamma)A\mathscr{D}_k^{-1}(\gamma)$ соответствует собственный вектор $y=\mathscr{D}_k(\gamma)x$. Для γ из формулы (3.18) k-й круг Гершгорина остается изолированным, а потому $|\eta_k|>|\eta_i|$ при $i\neq k$. Так как $\eta_k=\gamma\xi_k$ и $\eta_i=\xi_i$ $(i\neq k)$, это дает соотношения (3.19).

Еще более интересна структура собственного вектора y транспонированной матрицы A^{T}, который соответствует тому же значению λ.

Теорема 3.19 (В. А. Пупков). *Если* k*-й круг Гершгорина матрицы* A *изолирован, то для компонент вектора* y*, отвечающего собственному значению* $\lambda\in G_k(A)$*, выполняется неравенство*

$$|\eta_k|>\sum_{\substack{i=1\\ i\neq k}}^{n}|\eta_i|.$$

Итак, k-я компонента вектора y преобладает в строгом смысле над всеми остальными. Доказательство теоремы 3.19 дано в п. 13 дополнений к § 3.

Теорема 3.20 (В. А. Пупков). *Если у обеих матриц* A *и* A^{T} k*-й круг Гершгорина изолирован, то в круге*

$$|z-a_{kk}|\leqslant\beta,$$

где

$$\beta=\min\{\max_{j\neq k}|a_{jk}|,\ \max_{j\neq k}|a_{kj}|\},$$

содержится ровно одно собственное значение λ *матрицы* A.

Доказательство. Пусть x и y — собственные векторы матриц A и A^{T}, относящиеся к собственному значению $\lambda \in G_k(A)$. Согласно теореме 3.19,

$$|\xi_k| > \sum_{i \neq k} |\xi_i|, \quad |\eta_k| > \sum_{i \neq k} |\eta_i|.$$

Из соотношений $Ax = \lambda x$ и $A^{\mathrm{T}} y = \lambda y$ имеем

$$(\lambda - a_{kk})\xi_k = \sum_{i \neq k} a_{ki}\xi_i, \quad (\lambda - a_{kk})\eta_k = \sum_{i \neq k} a_{ik}\eta_i.$$

Отсюда

$$|\lambda - a_{kk}| \leqslant \sum_{i \neq k} |a_{ki}| \frac{|\xi_i|}{|\xi_k|} \leqslant \max_{j \neq k} |a_{kj}| \sum_{i \neq k} \frac{|\xi_i|}{|\xi_k|} \leqslant \max_{j \neq k} |a_{kj}|.$$

Аналогично показываем, что

$$|\lambda - a_{kk}| \leqslant \max_{j \neq k} |a_{jk}|.$$

Пример 3.21 (В. А. Пупков). Для матрицы

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 15 & 1 \\ 3 & 1 & 13 \end{bmatrix}$$

круг $G_1(A)$ изолирован, $R_1(A) = 5$, $T_1(A) = 4$, так что условие (3.16) выполнено при $k = 1$; формула (3.18) дает $\gamma = 0.8$. Согласно теореме 3.17, в круге $|z - 1| \leqslant 4$ содержится ровно одно собственное значение λ матрицы A. Однако теорема 3.20 уточняет эту оценку (ведь и круг $G_1(A^{\mathrm{T}})$ изолирован!): на самом деле верно, что $|\lambda - 1| \leqslant 3$. Для компонент собственного вектора x, относящегося к числу λ, выполняются неравенства $|\xi_1| > |\xi_2| + |\xi_3|$ (теорема 3.19), $|\xi_2| < 0.8|\xi_1|$, $|\xi_3| < 0.8|\xi_1|$ (теорема 3.18).

Поговорим теперь еще об одном направлении обобщений теоремы Леви — Десплanка и теории Гершгорина. Пусть $n \times n$-матрица A представлена в блочном виде

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1s} \\ A_{21} & A_{22} & \dots & A_{2s} \\ \dots & \dots & \dots & \dots \\ A_{s1} & A_{s2} & \dots & A_{ss} \end{bmatrix}, \tag{3.20}$$

где A_{ii} $(i = 1, \dots, s)$ — квадратные невырожденные подматрицы. Фиксируем произвольную мультипликативную матричную норму $\|\cdot\|$. Тогда справедлив следующий блочный вариант теоремы Леви — Десплanка.

Теорема 3.22. *Пусть для $i = 1, \dots, s$ выполняются неравенства*

$$\|A_{ii}^{-1}\|^{-1} > \sum_{\substack{j=1 \\ j \neq i}}^{s} \|A_{ij}\|. \tag{3.21}$$

В таком случае матрица A не вырождена.

Доказательство теоремы 3.22 можно найти в книге [12, гл. XIV, § 3]. Заметим, что при $s=n$, т. е. для матрицы A со скалярными элементами, неравенства (3.21) переходят в условия диагонального преобладания (3.12).

Из теоремы 3.22 можем получить теорему локализации, пользуясь уже хорошо нам известным способом рассуждений: для матрицы $B=A-\lambda I$ хотя бы одно из неравенств (3.21) должно быть нарушено. Эта блочная версия теоремы Гершгорина впервые была указаиа в [106].

Теорема 3.23 (теорема Фейнгольда — Варги). *Все собственные значения блочной матрицы* (3.20) *принадлежат объединению s областей*

$$\|(A_{ii}-zI)^{-1}\|^{-1} \leqslant \sum_{\substack{j=1 \\ j\neq i}}^{s} \|A_{ij}\|, \qquad i=1, 2, \ldots, s. \tag{3.22}$$

При этом, если z является собственным значением подматрицы A_{ii}, левую часть соотношения (3.22) *следует считать равной нулю* (так что неравенство заведомо выполнено).

Пример 3.24 (см. [12, с. 416—417]). Для матрицы

$$A=\begin{bmatrix} 0 & 4 & 1 & -1 \\ 4 & 0 & -1 & 1 \\ 1 & -1 & -1 & 15 \\ -1 & 1 & 15 & -1 \end{bmatrix}. \tag{3.23}$$

наибольший круг Гершгорина $|z+1|\leqslant 17$ имеет кратность 2 и содержит в себе другой, также двукратный круг Гершгорина $|z|\leqslant 6$. Поскольку речь идет о симметричной матрице, все собственные значения которой заведомо вещественны, то вместо кругов на комплексной плоскости можно рассматривать интервалы вещественной оси. Итак, теорема Гершгорина дает интервал локализации $-18\leqslant x\leqslant 16$.

Применим теперь теорему Фейнгольда — Варги, разбивая матрицу (3.23) на четыре 2×2-блока и выбирая в качестве матричной нормы $\|\cdot\|_\infty$. Тогда $\|A_{12}\|_\infty=\|A_{21}\|_\infty=2$,

$$(A_{11}-xI)^{-1}=\frac{1}{16-x^2}\begin{bmatrix} x & 4 \\ 4 & x \end{bmatrix},$$

$$\|(A_{11}-xI)^{-1}\|_\infty^{-1}=\frac{|x^2-16|}{|x|+4}=||x|-4|.$$

Аналогично

$$\|(A_{22}-xI)^{-1}\|_\infty^{-1}=||x+1|-15|.$$

Условия (3.22) принимают вид

$$||x|-4|\leqslant 2, \qquad ||x+1|-15|\leqslant 2$$

и приводят к системе из четырех интервалов:

$$\begin{aligned} -18\leqslant x\leqslant -16, \qquad & -6\leqslant x\leqslant -2, \\ 2\leqslant x\leqslant 6, \qquad & 12\leqslant x\leqslant 16. \end{aligned} \tag{3.24}$$

Объединение этих интервалов вложено в область Гершгорина, и их суммарная длина равна 14 по сравнению с длиной 34 гершгоринского интервала. Улучшение, таким образом, налицо, но и вычислительной работы потребовалось куда больше, чем для обычной теоремы Гершгорина.

Область, описываемая неравенствами (3.22), содержит в себе спектр матрицы A, какова бы ни была мультипликативная норма $\|\cdot\|$. Мы можем воспользоваться этим обстоятельством для еще большего сужения области локализации собственных значений матрицы (3.23). Беря при прежнем блочном разбиении матричную норму $\|\cdot\|_1$, приходим в соответствии с теоремой Фейнгольда — Варги к системе интервалов

$$-20 \leqslant x \leqslant -12, \quad -5 \leqslant x \leqslant -3, \quad 3 \leqslant x \leqslant 5, \quad 10 \leqslant x \leqslant 18. \tag{3.25}$$

Пересечение областей (3.24) и (3.25):

$$-18 \leqslant x \leqslant -16, \quad -5 \leqslant x \leqslant -3, \quad 3 \leqslant x \leqslant 5, \quad 12 \leqslant x \leqslant 16$$

также является областью локализации для матрицы A. Суммарная длина интервалов теперь уменьшилась до 10.

ДОПОЛНЕНИЯ К § 3.

1. Каждому вектору $x \in \mathbf{C}^n$ сопоставим матрицу X, первым столбцом которой является этот вектор; все прочие столбцы нулевые. Положим

$$\|x\| \equiv \|X\|, \tag{3.26}$$

где правая часть задается выбранной матричной нормой, а само равенство определяет векторную норму. Согласованность обеих норм легко следует из кольцевого свойства матричной нормы и соотношения

$$Ax \Leftrightarrow AX.$$

В случае евклидовой матричной нормы этот прием приводит к евклидовой же норме векторов; для подчиненных матричных норм он возвращает к порождающим векторным нормам. Так, в случае $\|X\|_\infty$ получаем $\|x\|_\infty$.

Вместо (3.26) можно пользоваться формулой

$$\|x\| \equiv \|xy^*\|,$$

где y — фиксированный ненулевой вектор.

Неравенство (3.1) выполняется не только для мультипликативных матричных норм, но и для более широкого класса норм, допускающих согласованные с ними векторные нормы. Описание этого класса матричных норм, а также обсуждение связи свойства согласованности со свойством *спектрального преобладания*, выражаемым соотношением (3.2), можно найти в § 5.7 книги [49].

2. В примере 3.4 выбор вместо $\|\cdot\|_\infty$ нормы $\|\cdot\|_1$ позволил существенно сократить диаметр области локализации. В связи с этим уместен вопрос: возможен ли выбор мультипликативной матричной нормы, для которой правая часть оценки (3.2) минимальна при любой матрице A? Ответ на этот вопрос отрицателен. Именно любые две подчиненные нормы матриц $M_1(A)$ и $M_2(A)$ не сравнимы, т. е. найдутся матрицы A_1 и A_2 такие, что $M_1(A_1) > M_2(A_1)$, но $M_1(A_2) < M_2(A_2)$ (см. [5, 49]).

3. Оценки спектрального радиуса, укладывающиеся в неравенство (3.2) при выборе той или иной конкретной нормы, были известны задолго до

того, как в алгебраической литературе установилось представление о нормах матриц. Так, неравенство

$$\rho(A)\leqslant\min\{\max_i\sum_{j=1}^n|a_{ij}|,\ \max_i\sum_{j=1}^n|a_{ji}|\} \tag{3.27}$$

доказано Брауэром [65] еще в 1946 г. (см. историю этой оценки в [32, с. 190—191]); при этом Брауэр не был знаком с теоремами Гершгорина, из которых (3.27) следует тривиально.

4. Хотя пересечение $G(A)\cap G(A^{\mathrm{T}})$ является областью локализации собственных значений матрицы A, неверной оказывается следующая комбинация строчной и столбцевой версий теоремы Гершгорина: все собственные значения матрицы A принадлежат объединению кругов

$$|z-a_{ii}|\leqslant\min\{R_i(A),\ C_i(A)\},\quad i=1,\ ...,\ n. \tag{3.28}$$

Например, для матрицы

$$\begin{bmatrix}0 & 0.1\\ -40 & 5\end{bmatrix}$$

оба собственных значения $\lambda_1=1$ и $\lambda_2=4$ не принадлежат системе кругов (3.28): $|z|\leqslant 0.1$; $|z-5|\leqslant 0.1$.

5. Правильное сочетание строчной и столбцевой теорем Гершгорина указано Островским [148]: областью локализации всех собственных значений матрицы A будет объединение n кругов

$$|z-a_{ii}|\leqslant[R_i(A)]^{\alpha}[C_i(A)]^{1-\alpha},\qquad i=1,\ ...,\ n. \tag{3.29}$$

Здесь α — произвольное число из $[0,1]$; концевым значениям α соответствуют: $\alpha=0$ — столбцевая, $\alpha=1$ — строчная версия теоремы Гершгорина. В литературе на русском языке доказательство теоремы Островского можно найти в книгах [32, с. 197—198; 49, § 6.4].

6. Если в неравенствах Островского используются произведения одноименных строчных и столбцевых сумм, то в другом обобщении теорем Гершгорина, принадлежащем Брауэру [65], рассматриваются парные произведения только строчных (или только столбцевых) сумм. Именно все собственные значения $n\times n$-матрицы A заключены в объединении $n(n-1)/2$ *овалов Кассини*

$$|z-a_{ii}||z-a_{jj}|\leqslant R_i(A)R_j(A),\qquad i,j=1,\ ...,\ n,\ i\neq j. \tag{3.30}$$

Аналогично формулируется столбцевой вариант теоремы Брауэра. Доказательство см. в [32, с. 195—196; 49, § 6.4].

7. Комбинируя теоремы Островского и Брауэра, приходим к области локализации, образуемой овалами Кассини вида (см. [32, с. 199])

$$|z-a_{ii}||z-a_{jj}|\leqslant[R_i(A)R_j(A)]^{\alpha}[C_i(A)C_j(A)]^{1-\alpha},$$

$$i,j=1,\ ...,\ n,\ i\neq j;\ 0\leqslant\alpha\leqslant 1.$$

8. Кажется вполне естественным предположить, что и теорема Гершгорина, и теорема Брауэра являются частными случаями следующего общего утверждения: для каждого натурального числа k $(0<k<n)$ спектр $n\times n$-матрицы A принадлежит объединению C_n^k областей

$$\prod_{\nu=1}^{k}|z-a_{i_\nu i_\nu}|\leqslant\prod_{\nu=1}^{k}R_{i_\nu}(A). \tag{3.31}$$

Однако при $k>2$ это утверждение неверно. Контрпримером к нему (автор —

М. Ньюмен) может служить матрица

$$A = \left[\begin{array}{ccc|cc} 1 & & 0 & & \\ & 1 & & & 0 \\ & & \ddots & & \\ 0 & & & 1 & \\ & & & & \\ \hline & & & 1 & 1 \\ & 0 & & 1 & 1 \end{array}\right].$$

Только две из ее строчных сумм отличны от нуля. Поэтому при $k>2$ все области (3.31) вырождаются в точку $z=1$. Между тем матрица A помимо 1 имеет собственные значения 0 и 2.

9. Корректное обобщение теоремы Брауэра получено в недавней работе Brualdi [66]: пусть A — слабо неразложимая (см. § 2) $n\times n$-матрица; $\mathscr{C}(A)$ — множество нетривиальных (т. е. содержащих хотя бы две различные вершины) простых циклов в графе этой матрицы. В таком случае каждое собственное значение матрицы A находится в области

$$\bigcup_{\gamma\in\mathscr{C}(A)} \{z\in C \mid \prod_{v_i\in\gamma} |z-a_{ii}| \leqslant \prod_{v_i\in\gamma} R_i(A)\}.$$

Эту запись надо понимать так: γ — простой цикл из $\mathscr{C}(A)$; $v_{i_1}, \ldots, v_{i_{k+1}}$ — вершины, составляющие цикл γ; при этом $v_{i_1}=v_{i_{k+1}}$.

10. Оригинальное обобщение теоремы Леви — Десплanка предложено В. Н. Соловьевым [41]. Обозначим $Q_j^{(r)}(A)$ сумму $r-1$ наибольших среди модулей внедиагональных элементов j-го столбца матрицы A; таким образом,

$$Q_j^{(2)}(A)=\max_{i\neq j}|a_{ij}|=T_j(A)$$

(см. (3.16)). Положим $P_j^{(r)}(A)=\min(Q_j^{(r)}(A), R_j(A))$. Матрица A будет невырождена, если для нее при некотором фиксированном натуральном r $(2\leqslant r\leqslant n)$ выполнены условия:

а) $|a_{jj}|>P_j^{(r)}(A)$ $(j=1, \ldots, n)$;

б) сумма модулей диагональных элементов в любых r различных строках больше суммы модулей всех внедиагональных элементов тех же строк.

Теорема Леви — Десплanка соответствует случаю $r=n$ этого утверждения. Из него обычным образом получается теорема локализации: каждое собственное значение матрицы A принадлежит либо одному из кругов

$$|z-a_{jj}|\leqslant P_j^{(r)}(A), \quad j=1, \ldots, n,$$

либо усреднению некоторых r кругов Гершгорина из системы $G_1(A), \ldots, G_n(A)$. Усредненнем кругов $G_{i_1}(A), \ldots, G_{i_r}(A)$ называется множество точек комплексной плоскости, сумма расстояний от которых до точек $a_{i_1i_1}, \ldots, a_{i_ri_r}$ не превосходит числа $R_{i_1}(A)+\ldots+R_{i_r}(A)$.

При $r=n$ эта последняя теорема Соловьева [41] дает теорему Гершгорина. Если k-й круг Гершгорина изолирован от прочих и справедливо неравенство (3.16), то из указанного результата можно другим путем вывести теорему 3.17.

11. Совсем в другом направлении идет обобщение теории Гершгорина, предпринятое Бауэром [59]. Пусть $S=\{y_1, \ldots, y_n\}$ — произвольный базис пространства $\mathbf{C}^n$. Положим

$$G_i^{(S)}(A)=\{(Ax, y_i) \mid x\in\mathbf{C}^n, (x, y_i)=1,$$

$$|(x, y_j)|\leqslant 1, \ j=1, \ldots, n, \ j\neq i\}, \ i=1, \ldots, n.$$

Тогда, как и в теореме Гершгорина, спектр матрицы A содержится в объединении областей $G_i^{(s)}(A)$.

Теорема Гершгорина получается из этого результата Брауэра при выборе стандартного базиса $y_i=e_i$ $(i=1, ..., n)$. В самом деле, рассмотрим, например, область $G_1^{(s)}(A)$. Если $x=(\xi_1, \xi_2, ..., \xi_n)^T$, то условия $(x, y_1)=1$, $|(x, y_j)|\leqslant 1$ $(j>1)$ означают, что $\xi_1=1$, $|\xi_j|\leqslant 1$ $(j>1)$ и

$$z=(Ax, e_1)=a_{11}+a_{12}\xi_2+...+a_{1n}\xi_n,$$

$$z-a_{11}=a_{12}\xi_2+...+a_{1n}\xi_n.$$

При произвольных ξ_j, по модулю не превосходящих 1, это просто другое описание гершгоринского круга $G_1(A)$.

Для бауэровского обобщения, вообще говоря, не имеет места аналог второй теоремы Гершгорина. Справедливо лишь следующее более слабое утверждение (Дойча—Зенгера; см. [93]):

Если область $G_k^{(s)}(A)$ не пересекается с выпуклой оболочкой прочих областей $G_j^{(s)}(A)$, то $G_k^{(s)}(A)$ содержит ровно одно собственное значение матрицы A.

Доказательство Дойча—Зенгера использует теорему Брауэра о неподвижной точке и не позволяет получить вторую теорему Гершгорина в частном случае $S=\{e_1, ..., e_n\}$. Доказательство, допускающее вывод второй теоремы Гершгорина, дано в недавней публикации [215].

12. Замечание 3.7 приводит к очень простому обоснованию следующего утверждения (см. [8, с. 194]): если λ — собственное значение матрицы A, имеющее геометрическую кратность m, то λ принадлежит по крайней мере m кругам Гершгорина. Действительно, пусть $y_1, ..., y_m$ — линейно независимые собственные векторы, относящиеся к λ, и пусть Y есть $n\times m$-матрица, составленная по столбцам из этих векторов. Выполняя элементарные операции над столбцами матрицы Y (своего рода столбцовый вариант метода Гаусса с выбором главного элемента *по столбцу*), можно добиться, чтобы максимальные по модулю элементы всех m столбцов преобразованной матрицы $\tilde{Y}$ стояли в *различных* m строках $i_1, ..., i_m$. Но $\tilde{y}_1, ..., \tilde{y}_m$ — столбцы матрицы $\tilde{Y}$ — также суть собственные векторы матрицы A, отвечающие числу λ. Согласно замечанию 3.7, λ должно содержаться по меньшей мере в кругах Гершгорина с номерами $i_1, ..., i_m$.

13. Приведем доказательство теоремы 3.19, принадлежащее Пункову [38]. Предположим, что, напротив,

$$|\eta_k|\leqslant \sum_{i\neq k} |\eta_i|. \tag{3.32}$$

Положив $D=\operatorname{diag}(a_{11}, ..., a_{nn})$, рассмотрим семейство матриц $A(t)=tA+(1-t)D$; в нем $A(0)=D$, $A(1)=A$. Поскольку у каждой матрицы $A(t)$, где $0\leqslant t\leqslant 1$, k-й круг Гершгорина остается изолированным, то собственное значение λ, непрерывно изменяясь при t, уменьшающемся от 1 до 0, при $t=0$ принимает значение a_{kk}. Если для всех t присвоить величину 1 какой-либо определенной компоненте соответствующего собственного (по отношению к матрице $A^T(t)$) вектора $y(t)$ (который в силу простоты собственного значения становится в этом случае однозначно определенным), то все его компоненты $\eta_i(t)$ также будут непрерывными функциями от t. При $t=0$ имеем $y(0)=e_k$; для этого вектора нужный вид преобладания выполнен. Вместе с (3.32) это означает, что при некотором значении t_0, $0<t_0<1$, должно осуществляться равенство

$$|\eta_k(t_0)|=\sum_{\substack{i=1\\ i\neq k}}^{n} |\eta_i(t_0)|. \tag{3.33}$$

Из равенства

$$(A^{\tau}(t_0)-a_{kk}I)y(t_0)=(\lambda(t_0)-a_{kk})y(t_0) \tag{3.34}$$

выводим

$$\sum_{\substack{i=1\\ i\neq k}}^{n}\left|\sum_{\substack{j=1\\ j\neq i}}^{n} t_0 a_{ji}\eta_j(t_0)+(a_{ii}-a_{kk})\eta_i(t_0)\right|=\left|\sum_{\substack{j=1\\ j\neq k}}^{n} t_0 a_{jk}\eta_j(t_0)\right|. \tag{3.35}$$

Это замысловатое соотношение получено таким образом. Для вектора $z=(\lambda(t_0)-a_{kk})y(t_0)$, пропорционального вектору $y(t_0)$, выполняется равенство, аналогичное (3.33). Левая часть в (3.35) есть $\sum\limits_{i\neq k}|\zeta_i|$, причем при выписывании компонент ζ_i использована левая часть в (3.34); правая же часть в (3.35)—это $|\zeta_k|$, где выражение для $\zeta_k=(\lambda(t_0)-a_{kk})\eta_k(t_0)$ опирается на тождества

$$(\lambda(t_0)-a_{kk})\eta_k(t_0)=\sum_{\substack{j=1\\ j\neq k}}^{n} a_{jk}(t_0)\eta_j(t_0)=\sum_{\substack{j=1\\ j\neq k}}^{n} t_0 a_{jk}\eta_j(t_0).$$

Наряду с равенством (3.35) имеет место оценка

$$\sum_{\substack{i=1\\ i\neq k}}^{n}\left|\sum_{\substack{j=1\\ j\neq i}}^{n} t_0 a_{ji}\eta_j(t_0)+(a_{ii}-a_{kk})\eta_i(t_0)\right|-\left|\sum_{\substack{j=1\\ j\neq k}}^{n} t_0 a_{jk}\eta_j(t_0)\right|\geqslant$$

$$\geqslant\sum_{\substack{i=1\\ i\neq k}}^{n}|a_{ii}-a_{kk}||\eta_i(t_0)|-t_0\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ji}||\eta_j(t_0)|=$$

$$=\sum_{\substack{i=1\\ i\neq k}}^{n}|a_{ii}-a_{kk}||\eta_i(t_0)|-t_0\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}||\eta_i(t_0)|=$$

$$=\sum_{\substack{i=1\\ i\neq k}}^{n}(|a_{ii}-a_{kk}|-t_0R_i(A))|\eta_i(t_0)|-t_0R_k(A)|\eta_k(t_0)|=$$

$$=\sum_{\substack{i=1\\ i\neq k}}^{n}(|a_{ii}-a_{kk}|-t_0R_i(A)-t_0R_k(A))|\eta_i(t_0)|>0.$$

В предпоследнем переходе использовалось равенство (3.33); заключение же о положительности вытекает из того, что содержимое каждых круглых скобок больше нуля ввиду изолированности k-го круга Гершгорина. Полученная оценка противоречит (3.35); стало быть, соотношение (3.33) невозможно, и вид преобладания для векторов e_k и y одинаков.

14. Предположим, что блочная матрица (3.20) имеет «малые» (в смысле значений норм) внедиагональные блоки A_{ij}, а спектры диагональных блоков попарно не пересекаются. В ряде работ рассматривались способы перенесения на этот случай оценки (3.11). Качественное обсуждение вопроса дано в [10, с. 116]. Приведем одну из возможных количественных формулировок [165].

Положим

$$\|A\|_{1,\infty}=\max_{1\leqslant i\leqslant s}\left(\sum_{j=1}^{s}\|A_{ij}\|_1\right).$$

Представим A в виде суммы $D+K$, где $D=A_{11}\oplus A_{22}\oplus\ldots\oplus A_{ss}$; блоки $K_{11}, \ldots, K_{ss}$ нулевые. *Разделенность спектров* различных диагональных блоков A_{ii} будем характеризовать величиной

$$\sigma=\max_{i\neq j}\{\|(A_{jj}^{\mathrm{T}}\otimes I_{n_i}-I_{n_j}\otimes A_{ii})^{-1}\|_{1,\infty}\}. \tag{3.36}$$

Здесь n_i, n_j — порядки блоков A_{ii} и A_{jj}. Число σ есть, в сущности, суммарная характеристика обусловленности ляпуновских операторов $A_{ii}X-XA_{jj}$, ассоциированных с парами диагональных блоков. Несмотря иа свою громоздкость, (3.36) есть точное обобщение соответствующей скаляриой величины; действительно, если все n_i равны 1, то $\sigma^{-1}=\min_{i\neq j}|a_{ii}-a_{jj}|$.

Теорема (Д. Пауэрс). *Если* $\sigma\|K\|_{1,\infty}<(\sqrt{3}-1)/2$, *то существует матрица* $B=B_1\oplus B_2\oplus\ldots\oplus B_s$, *подобная матрице* A *и такая, что*

$$\|B_{ii}-A_{ii}\|_1\leqslant\frac{\sigma\|K\|_{1,\infty}^2}{1-2\sigma\|K\|_{1,\infty}}.$$

При определеиных условиях (см. § 4) теорема Пауэрса означает, что собственные значения диагональных блоков A_{ii} приближают собственные значения матрицы A с точностью $O(\varepsilon^2)$, где $\varepsilon=\|K\|$.

15. В работе Коварика и Олески [134] дано следующее любопытиое описаиие блочных «кругов» Гершгорина

$$\hat{G}_\alpha(A)=\{z\,|\,\|(A_{\alpha\alpha}-zI_\alpha)^{-1}\|^{-1}\leqslant\hat{R}_\alpha(A)\},$$

$$\hat{R}_\alpha(A)\equiv\sum_{\substack{\beta=1\\ \beta\neq\alpha}}^{s}\|A_{\alpha\beta}\|$$

(напомним, что в случае, если z — собственное значение блока $A_{\alpha\alpha}$, левую часть неравенства следует считать нулем): «круг» $\hat{G}_\alpha(A)$ совпадает с множеством собственных значений всех матриц вида $A_{\alpha\alpha}+X$, где $\|X\|\leqslant\hat{R}_\alpha(A)$. Это описание аналогично характеристике обычного гершгоринского круга $G_i(A)$ как множества чисел вида $a_{ii}+x$, где $|x|\leqslant R_i(A)$. Нужно отметить, что описание области $\hat{G}_\alpha(A)$ предполагает (как и теорема Фейнгольда — Варги в авторской формулировке), что для всех блоков используется подчиненная матричная норма одного и того же типа.

Результат Коварика — Олески обобщается и уточняется в [179] с помощью аппарата векториальных норм (т. е. матричных норм со значениями в положительном конусе арифметического пространства).

16. Мы видели, что для конкретных матриц область локализации, указываемую теоремой Гершгорина, можно сузить посредством... теоремы Гершгорина. Для этого, согласно примеру 3.12, вместо исходной матрицы A теорему нужно применить к матрице PAP^{-1}, где P — подходящим образом подобранная диагональная матрица с положительной диагональю.

Нередко лучшие по сравнению с гершгоринскими оценки собственных значений дают другие теоремы локализации, использующие ту же информацию о матрице, а именно значения диагональных и модули внедиагональных элементов. К числу теорем с такой информацией относятся теоремы Бауэра [65], Бруалди [66], Соловьева [41] и др. И все же, как установил Варга [202], в классе теорем этого рода теорема Гершгорина, применяемая сразу ко всем диагонально подобным матрицам, обладает

следующим свойством оптимальности: пусть z — произвольная граничная точка области

$$\bigcap_P G(PAP^{-1}), \quad P=\operatorname{diag}(p_1, \ldots, p_n), \quad p_i>0, \ \forall i;$$

тогда найдется $n\times n$-матрица B такая, что: 1) $b_{ii}=a_{ii}$ $(i=1, \ldots, n)$; 2) $|b_{ij}|=|a_{ij}|$ $\forall i, j$; 3) z — собственное значение матрицы B.

§ 4. Теорема Бауэра — Файка и ее обобщения

Наиболее полезным среди многочисленных обобщений теоремы Гершгорина является, пожалуй, обсуждаемая в настоящем параграфе теорема Бауэра — Файка. Она уже настолько вошла в алгебраический фольклор, что в ряде учебников (см., например, [8, § 89; 29, теорема 7.4.2]) утеряла в своем названии имена авторов.

Теорема Бауэра — Файка оценивает расстояния между собственными значениями двух матриц, одну из которых мы условимся считать возмущенной по отношению к другой: $B=A+F$. Не требуется, чтобы матрица F была малой, и в этом смысле теорема Бауэра — Файка есть теорема локализации. Однако чаще всего теорема используется, чтобы исследовать влияние малых возмущений в элементах матрицы A на ее спектр.

Близость матриц A и B измеряется в теореме посредством матричных норм, подчиненных абсолютным векторным нормам. Напомним (см. § 2), что всякая такая норма обладает следующим свойством: если $D=\operatorname{diag}(d_1, \ldots, d_n)$, то

$$\|D\|=\max_i |d_i|.$$

Теорема 4.1 (первая теорема Бауэра — Файка). *Пусть A — диагонализуемая матрица и $P^{-1}AP=\Lambda=\operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Для всякого собственного значения μ_i возмущенной матрицы $B=A+F$ найдется собственное значение λ_{σ_i} матрицы A такое, что*

$$|\mu_i-\lambda_{\sigma_i}|\leqslant\|P^{-1}FP\|. \tag{4.1}$$

Доказательство. Если μ_i является одновременно собственным значением матрицы A, то полагаем $\lambda_{\sigma_i}=\mu_i$, и неравенство (4.1) выполнено. В противном случае матрица $\Lambda-\mu_i I$ не вырождена. В то же время вырождена матрица

$$B-\mu_i I=A-\mu_i I+F=P\Lambda P^{-1}-\mu_i I+F=P\left(\Lambda-\mu_i I+P^{-1}FP\right)P^{-1}=$$
$$=P\left(\Lambda-\mu_i I\right)\left[I+\left(\Lambda-\mu_i I\right)^{-1}P^{-1}FP\right]P^{-1}.$$

Матрица в квадратных скобках вырождена, а потому

$$\|\left(\Lambda-\mu_i I\right)^{-1}P^{-1}FP\|\geqslant 1. \tag{4.2}$$

Действительно, матрица вида $I+G$, где $\|G\|<1$, всегда не вырождена. Из (4.2) выводим

$$1\leqslant\|P^{-1}FP\|\,\|\left(\Lambda-\mu_i I\right)^{-1}\|=\|P^{-1}FP\|/\min_{\lambda_\sigma}|\lambda_\sigma-\mu_i|.$$

Это и есть требуемое неравенство (4.1).

Замечание 4.2. Теорема Бауэра — Файка содержит в себе теорему Гершгорина, правда, в несколько загрубленном варианте. Будем рассматривать матрицу A как возмущение ее собственной диагональной части — матрицы $D=\operatorname{diag}(a_{11}, \ldots, a_{nn})$. Тогда матрицу P можно считать единичной, и, если в качестве матричной нормы взять $\|\cdot\|_\infty$, из теоремы Бауэра — Файка следует, что всякое собственное значение λ матрицы A принадлежит некоторому кругу вида

$$|z-a_{ii}| \leqslant \max_{1\leqslant\sigma\leqslant n} R_\sigma(A).$$

Замечание 4.3. Если говорить сразу обо всех собственных значениях матриц A и B, то теорема Бауэра — Файка допускает двоякую интерпретацию. С одной стороны, совокупность всех кругов

$$|z-\lambda_\sigma| \leqslant \| P^{-1}FP\|, \qquad \sigma=1, \ldots, n, \tag{4.3}$$

с центрами в собственных значениях λ_σ матрицы A заключает в себе *все* собственные значения возмущенной матрицы B. С другой стороны, *каждый* круг

$$|z-\mu_i| \leqslant \| P^{-1}FP\|, \quad i=1, \ldots, n,$$

содержит хотя бы одно собственное значение матрицы A; в то же время нельзя поручиться, что в объединение этих кругов попадет весь спектр.

Справедлив следующий аналог второй теоремы Гершгорина.

Теорема 4.4 (вторая теорема Бауэра — Файка). *Каждая связная компонента области, получаемой объединением кругов* (4.3), *содержит столько собственных значений возмущенной матрицы B, сколько кругов составляют эту компоненту.*

Доказательство теоремы 4.4 проводится по той же схеме, что и доказательство второй теоремы Гершгорина.

Замечание 4.5. Теорема Бауэра — Файка*) обычно используется в ослабленной формулировке: неравенство (4.1) заменяется на

$$|\mu_i-\lambda_{\sigma_i}| \leqslant \| P\| \| P^{-1}\| \| F\| = \operatorname{cond} P \| F\|. \tag{4.4}$$

Замечание 4.6. При малых F, учитывая теорему 4.4, получаем из (4.4) очень важный качественный вывод:

Возмущения собственных значений диагонализуемой матрицы A, отвечающие малым возмущениям ее элементов, ограничены величиной, пропорциональной норме матрицы-возмущения. В качестве коэффициента пропорциональности выступает число обусловленности матрицы P, столбцами которой являются собственные векторы матрицы A.

Даже для матрицы A с простым спектром матрица собственных векторов определена неоднозначно. Еще больше произвол в выборе матрицы P, если у A имеются кратные собственные значения. Теоремы Бауэра — Файка выполняются при любом выборе, но ясно,

*) Далее при упоминании теоремы Бауэра — Файка имеется в виду теорема 4.1.

что для более точной оценки возмущений спектра мы заинтересованы в матрице P с наименьшим возможным числом обусловленности.

Остановимся в связи со сказанным на случае нормальной матрицы A. Среди матриц P, диагонализующих матрицу A, есть унитарная (хотя, разумеется, есть и неунитарные). Беря эту унитарную матрицу и спектральную норму, выводим из теоремы Бауэра — Файка такое следствие:

При возмущении нормальной матрицы A матрицей F собственные значения возмущенной матрицы $B=A+F$ заключены в объединении кругов

$$|z-\lambda_i|\leqslant\|F\|_2, \qquad i=1, \ldots, n. \tag{4.5}$$

Поскольку ни для какой матрицы P число обусловленности не может быть меньше 1, мы вправе сказать, что собственные значения нормальных матриц наиболее устойчивы к возмущениям матричных элементов.

Пример 4.7. Пользуясь теоремой Бауэра — Файка, построим область локализации для собственных значений матрицы

$$B=\begin{bmatrix} 2.001 & 1.499 & 0.001 \\ 0.499 & 1.001 & -0.001 \\ -0.001 & 0.001 & 0.999 \end{bmatrix}.$$

Будем рассматривать матрицу B как возмущенную по отношению к матрице

$$A=\begin{bmatrix} 2 & 1.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Собственные значения матрицы A — это числа $\lambda_1=2.5$, $\lambda_2=0.5$, $\lambda_3=1$. В качестве матрицы P можно взять матрицу

$$P=\begin{bmatrix} 3 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Выбирая норму $\|\cdot\|_\infty$, находим $\|P\|_\infty=4$, $\|P^{-1}\|_\infty=1$, $\|B-A\|_\infty=0.003$, и круги Бауэра — Файка имеют вид

$$|z-2.5|\leqslant 0.012, \quad |z-1|\leqslant 0.012, \quad |z-0.5|\leqslant 0.012.$$

Условие теоремы Бауэра — Файка о диагонализуемости матрицы A существенно. Дело в том, что кратное собственное значение, которому в жордановой форме матрицы отвечает жорданова клетка порядка $k\geqslant 2$, гораздо чувствительней к малым возмущениям элементов матрицы, чем «линейно изменяющиеся» собственные значения из оценки (4.4).

Теорема о возмущениях собственных значений, справедливая для *любой* матрицы, сформулирована в 1957 г. Островским [149]. Будем считать исходную $n\times n$-матрицу A нормированной так, чтобы $|a_{ij}|\leqslant 1\ \forall i, j$. Матрицу-возмущение F удобно представить в виде εH, где ε — малый положительный параметр и $|h_{ij}|\leqslant 1\ \forall i, j$.

Теорема 4.8 (теорема Островского). *Для каждого собственного значения μ_i возмущенной матрицы $B=A+\varepsilon H$ найдется собственное значение λ_{σ_i} матрицы A, такое, что*

$$|\mu_i-\lambda_{\sigma_i}|\leqslant(n+2)(n^2\varepsilon)^{1/n}. \tag{4.6}$$

Между собственными значениями обеих матриц можно установить взаимно однозначное соответствие $\mu_i\leftrightarrow\lambda_{\tau_i}$ $(i=1, \ldots, n)$, при котором

$$|\mu_i-\lambda_{\tau_i}|\leqslant 2(n+1)^2(n^2\varepsilon)^{1/n}. \tag{4.7}$$

Мы не даем доказательства теоремы Островского, потому что ниже будет приведен (с доказательством) более точный результат о возмущениях собственных значений матриц общего вида. Взаимосвязь между соотношениями типа (4.6) и (4.7) обсуждается в п. 1 дополнений к § 4.

Главное отличие оценки (4.6) от оценки (4.4) состоит в том, что (если говорить в одинаковых терминах) параметр ε имеет показатель $1/n$ вместо 1. При больших n и малых ε разница между ε и $\varepsilon^{1/n}$ огромная. Так, при $\varepsilon=10^{-n}$ будем иметь $\varepsilon^{1/n}=0.1$ — число, которое не назовешь очень малым.

Присутствие множителя $\varepsilon^{1/n}$ неизбежно в любой общей оценке возмущений собственных значений. Это утверждение основывается на исследовании возмущений жордановой клетки порядка n. Если в позицию $(n, 1)$ матрицы

$$J_n(\alpha)=\begin{bmatrix} \alpha & 1 & & & \\ & \alpha & 1 & & \\ & & \ddots & \ddots & \\ & & & \alpha & 1 \\ & & & & \alpha \end{bmatrix}$$

внести возмущение, равное $(-1)^n\varepsilon$ $(\varepsilon>0)$, то характеристическое уравнение возмущенной матрицы примет вид

$$(\mu-a)^n=\varepsilon,$$

т. е. $\mu_i=a+\omega^i\varepsilon^{1/n}$ $(i=1, \ldots, n)$, где ω — какой-либо первообразный корень из 1.

Если максимальный порядок m жордановых клеток в жордановой форме матрицы A (в § 1 он был назван индексом матрицы) существенно меньше ее порядка n, то для возмущений собственных значений можно дать значительно более оптимистическую, чем в теореме Островского, оценку.

Теорема 4.9 (см. [132]). *Пусть $P^{-1}AP=J$ и индекс матрицы A равен m. Для всякого собственного значения μ_i возмущенной матрицы $B=A+F$ найдется собственное значение λ_{σ_i} матрицы A такое, что*

$$|\mu_i-\lambda_{\sigma_i}|^m/(1+|\mu_i-\lambda_{\sigma_i}|)^{m-1}\leqslant\|P^{-1}FP\|_2. \tag{4.8}$$

Если обозначить через $\omega(m, \varepsilon)$ *неотрицательный корень уравнения* $\xi^m = \varepsilon(1+\xi)^{m-1}$, *то можно установить взаимно однозначное соответствие между собственными значениями матриц* A *и* B: $\mu_i \leftrightarrow \lambda_{\tau_i}$ $(i=1, \ldots, n)$, *при котором*

$$|\mu_i - \lambda_{\tau_i}| \leqslant (2n-1)\,\omega(m, \|P^{-1}FP\|_2). \tag{4.9}$$

Доказательство. Первая часть доказательства дословно (с заменой Λ на J) совпадает с доказательством теоремы Бауэра—Файка. Если μ_i не является одновременно собственным значением матрицы A, то приходим (см. (4.2)) к неравенству

$$1/\|(J-\mu_i I)^{-1}\|_2 \leqslant \|P^{-1}FP\|_2. \tag{4.10}$$

В оставшейся части доказательства мы дадим для левой части оценку снизу, зависящую только от λ_σ, μ_i и m. Спектральная норма блочно-диагональной матрицы $(J-\mu_i I)^{-1}$ равна максимальной из спектральных норм диагональных блоков. Поэтому достаточно произвести оценку для случая, когда J—жорданов блок порядка s, отвечающий числу λ.

Ввиду связи спектральной нормы с сингулярными числами будем говорить о минимальном сингулярном числе матрицы $J_s(\lambda-\mu_i)$ или о минимальном собственном значении трехдиагональной матрицы

$$T = J_s(a)\,J_s^*(a) = \begin{bmatrix} 1+|a|^2 & \bar{a} & & & & \\ a & 1+|a|^2 & \bar{a} & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1+|a|^2 & \bar{a} & \\ & & & a & |a|^2 & \end{bmatrix}, \quad a = \lambda - \mu_i.$$

Собственные значения матрицы T будем считать упорядоченными по возрастанию: $0 < \rho_1 \leqslant \rho_2 \leqslant \ldots \leqslant \rho_s$. Тогда

$$\rho_1 = \frac{\det T}{\rho_2 \cdots \rho_s} = \frac{|\det J_s(a)|^2}{\rho_2 \cdots \rho_s} = \frac{|a|^{2s}}{\rho_2 \cdots \rho_s} \geqslant \frac{|a|^{2s}}{(1+|a|)^{2s-2}}.$$

Последний переход использует верхнюю оценку собственных значений (см. пример 3.5)

$$\rho_t \leqslant 1+|a|^2+2|a| = (1+|a|)^2.$$

Минимальное сингулярное число матрицы $J_s(a)$, будучи корнем из ρ_1, удовлетворяет неравенству

$$1/\|[J_s(a)]^{-1}\|_2 \geqslant |a|^s/(1+|a|)^{s-1}. \tag{4.11}$$

С возрастанием s (при фиксированном a) правая часть этой оценки убывает. Если λ_{σ_i}—собственное значение жорданова блока матрицы J, на котором реализуется норма в левой части неравенства (4.10), то, согласно (4.11), для максимального возможного порядка m должно быть

$$|\lambda_{\sigma_i} - \mu_i|^m/(1+|\lambda_{\sigma_i} - \mu_i|)^{m-1} \leqslant \|P^{-1}FP\|_2.$$

По поводу второго утверждения теоремы см. п. 1 дополнений к § 4.

Замечание 4.10. При $m=1$, т. е. для диагонализуемой матрицы A, (4.8) превращается в оценку Бауэра—Файка.

ДОПОЛНЕНИЯ К § 4

1. Пусть A и B—комплексные $n\times n$-матрицы со спектрами соответственно $\sigma(A)=\{\lambda_1, \ldots, \lambda_n\}$ и $\sigma(B)=\{\mu_1, \ldots, \mu_n\}$. Для *расстояния между спектрами* матриц A и B можно использовать различные определения. Вот некоторые варианты:

$$s_A(B)=\max_i \min_j |\lambda_j-\mu_i|, \tag{4.12}$$

$$h_A(B)=\max_{0\leqslant t\leqslant 1} s_A((1-t)A+tB), \tag{4.13}$$

$$v(A, B)=\min_{\pi} \max_i |\mu_i-\lambda_{\pi(i)}|. \tag{4.14}$$

В последнем случае минимум берется по всем перестановкам π из чисел $1,\ldots, n$. Мера $v(A, B)$ в отличие от первых двух симметрична относительно матриц A, B. В теоремах Бауэра—Файка, Островского и 4.9 мы фактически имеем дело с различными оценками указанных мер.

Изучению соотношений различных расстояний между спектрами посвящен ряд работ. Так, в [101] показано, что

$$v(A, B)\leqslant(2n-1)h_A(B), \tag{4.15}$$

$$v(A, B)\leqslant a_n \max\{h_A(B), h_B(A)\}, \tag{4.16}$$

где

$$a_n=\begin{cases} n, & n \text{ четное,} \\ n-1, & n \text{ нечетное.}\end{cases}$$

При этом константы $2n-1$ и a_n в оценках (4.15), (4.16) не улучшаемы.

Однако для наших теорем полезней было бы соотношение между мерами $s_A(B)$ и $v(A, B)$. Такое соотношение дано в [128].

Теорема (E. Jiang). *Предположим, что для каждой матрицы-возмущения E найдется число $\eta(E)$, непрерывно зависящее от E и такое, что всякому собственному значению μ_i возмущенной матрицы $A+E$ соответствует собственное значение λ_{σ_i}, для которого*

$$|\mu_i-\lambda_{\sigma_i}|\leqslant\eta(E).$$

Если $\eta(tE)\leqslant\eta(E)$ для всех $0\leqslant t\leqslant 1$, то существует такое упорядочение собственных значений матрицы A: $\lambda_{\tau_1}, \lambda_{\tau_2}, \ldots, \lambda_{\tau_n}$, что

$$|\mu_i-\lambda_{\tau_i}|\leqslant(2k-1)\eta(E), \qquad i=1, \ldots, n,$$

где k—число различных среди собственных значений λ. Во всяком случае,

$$|\mu_i-\lambda_{\tau_i}|\leqslant(2n-1)\eta(E).$$

Эта теорема объясняет переход от (4.8) к (4.9) и показывает заодно, что во втором неравенстве Островского (4.7) коэффициент $2(n+1)^2$ можно заменить на несколько меньший коэффициент $(n+2)(2n-1)$.

Оценке Островского (4.6) можно придать вид, в котором правая часть, по существу, не имеет коэффициента, растущего вместе с n. Если для

измерений взять спектральную норму и положить $M=\max\{\|A\|_2,\ \|B\|_2\}$, то [61]

$$S_A(B)\leqslant n^{1/n}(2M)^{1-1/n}\,\|A-B\|_2^{1/n}.$$

Эльзнер [102] еще улучшил это неравенство:

$$S_A(B)\leqslant(\|A\|_2+\|B\|_2)^{1-1/n}\,\|A-B\|_2^{1/n}.$$

Эта последняя оценка точная, и в [102] дано описание всех матричных пар A, B, для которых она достигается.

2. Очевидная близость формулировок теорем Гершгорина и Бауэра—Файка вызвала ряд попыток объединить их (и другие результаты о локализации) в рамках единой схемы. Рассмотрим, например, схему, предложенную в [125].

Пусть X, Y, Z—матрицы размеров соответственно $n\times p$, $n\times n$ и $p\times p$, причем $p\leqslant n$. Предположим, что в $\mathbf{C}^n$ и $\mathbf{C}^p$ выбраны векторные нормы соответственно $\|\cdot\|'$ и $\|\cdot\|''$ и определяемая ими подчиненная норма (см. § 2) используется для $n\times p$-матриц. Если ζ—собственное значение матрицы Z, не являющееся в то же время собственным значением матрицы Y, то

$$\min_{\|u\|''=1}\|Xu\|'=\min_{u\neq 0}\frac{\|Xu\|'}{\|u\|''}\leqslant\|(Y-\zeta I)^{-1}(YX-XZ)\|. \tag{4.17}$$

В самом деле, пусть v—собственный вектор матрицы Z, относящийся к числу ζ. Можно считать v нормированным: $\|v\|''=1$. Тогда

$$\|Xv\|'=\|(Y-\zeta I)^{-1}(Y-\zeta I)Xv\|'=\|(Y-\zeta I)^{-1}(YXv-X(\zeta v))\|'=$$
$$=\|(Y-\zeta I)^{-1}(YX-XZ)v\|'\leqslant\|(Y-\zeta I)^{-1}(YX-XZ)\|.$$

Выведем теперь из (4.17) теоремы Гершгорина и Бауэра—Файка. Положим $p=n$, $X=I_n$, $Z=A$, $Y=\operatorname{diag}(a_{11},\ \ldots,\ a_{nn})$, $\|\cdot\|'=\|\cdot\|''=\|\cdot\|_\infty$. В этом случае $\|Xu\|_\infty=\|u\|_\infty=1$, $(Y-\zeta I)^{-1}=\operatorname{diag}(1/(a_{11}-\zeta),\ \ldots,\ 1/(a_{nn}-\zeta))$,

$$Y-Z=\begin{bmatrix}0 & -a_{12} & \ldots & -a_{1n}\\ -a_{21} & 0 & \ldots & -a_{2n}\\ \ldots & \ldots & \ldots & \ldots\\ -a_{n1} & -a_{n2} & \ldots & 0\end{bmatrix}.$$

Согласно (4.17),

$$1\leqslant\|(Y-\zeta I)^{-1}(Y-Z)\|_\infty=\max_i R_i(A)/|a_{ii}-\zeta|.$$

Стало быть, найдется i_0 такое, что

$$|a_{i_0i_0}-\zeta|\leqslant R_{i_0}(A).$$

Это и есть теорема Гершгорина.

Пусть снова $n=p$ и $X=I$. В качестве $\|\cdot\|'$ и $\|\cdot\|''$ возьмем какую-нибудь (одну и ту же) абсолютную векторную норму. Полагая $Y=A$, $Z=B$ и считая A диагонализуемой матрицей: $P^{-1}AP=\Lambda=\operatorname{diag}(\lambda_1,\ \ldots,\ \lambda_n)$, получаем из (4.17)

$$1\leqslant\|P(\Lambda-\zeta I)^{-1}P^{-1}(A-B)\|\leqslant\|P\|\,\|(\Lambda-\zeta I)^{-1}\|\,\|P^{-1}\|\,\|A-B\|=$$
$$=\frac{1}{\min\limits_i|\lambda_i-\zeta|}\operatorname{cond}P\,\|A-B\|,$$

т. е. $\min\limits_i|\lambda_i-\zeta|\leqslant\operatorname{cond}P\,\|A-B\|$—одна из формулировок теоремы Бауэра—Файка.

3. В [178] указано следующее обобщение теоремы Бауэра — Файка и теоремы 4.9: для всякого собственного значения μ_i возмущенной матрицы $B=A+F$ найдется собственное значение λ_{σ_i} матрицы A такое, что

$$|\mu_i-\lambda_{\sigma_i}|\leqslant\max\left(r_i\,\|P^{-1}FP\|,\ (r_i\,\|P^{-1}FP\|)^{1/m_i}\right).$$

Здесь P — матрица, трансформирующая A к жордановой форме; m_i — индекс собственного значения λ_{σ_i}; $r_i=[m_i(m_i+1)/2]^{1/2}$.

4. В своей последней работе Уилкинсон [211] обсуждает блочное обобщение теоремы Бауэра — Файка, высказанное в устной форме (но не опубликованное) Фейнгольдом. Оно имеет следующий вид. Если X — невырожденная матрица, трансформирующая A к блочно-диагональному виду $A_{11}\oplus A_{22}\oplus\ldots\oplus A_{ss}$, то собственные значения μ_i возмущенной матрицы $B=A+F$ принадлежат объединению областей

$$\frac{1}{\|(A_{ii}-zI)^{-1}\|}\leqslant\operatorname{cond}X\cdot\|F\|,\qquad i=1,\ \ldots,\ s \tag{4.18}$$

(ср. с (4.4)). Левую часть в (4.18) следует считать нулем, если z совпадает с собственным значением блока A_{ii}. От нормы $\|\cdot\|$ требуется, чтобы в случае блочно диагональной матрицы она равнялась наибольшей из норм диагональных блоков.

§ 5. Чувствительность собственных значений и мера анормальности матрицы

Как было отмечено в предыдущем параграфе (см. неравенства (4.5)), собственные значения нормальных матриц наиболее устойчивы к возмущениям матричных элементов. На это свойство устойчивости не влияет возможная кратность собственных значений. Между тем в любой окрестности нормальной матрицы A с кратным λ имеется матрица с недиагональной жордановой формой. Так, при любом ε жорданову клетку

$$J_{n,\varepsilon}(1)=\begin{bmatrix} 1 & \varepsilon & & & \\ & 1 & \varepsilon & & \\ & & \ddots & \ddots & \\ & & & 1 & \varepsilon \\ & & & & 1 \end{bmatrix} \tag{5.1}$$

можно рассматривать как ε-возмущение нормальной матрицы, а именно единичной матрицы I_n. Для собственных значений недиагонализуемых матриц характерен (см. теорему 4.9) совсем другой закон реагирования на малые возмущения элементов. И это кажется противоестественным, если принять во внимание непрерывную зависимость собственных значений от элементов матрицы.

Выход из обрисованного парадокса состоит в том, чтобы соизмерять степень отклонения матрицы A от множества нормальных матриц — *меру анормальности* матрицы A — и допустимый уровень возмущений в ее элементах. Например, обсуждая теорему Островского 4.8, мы отметили, что внесение возмущения, равного $(-1)^n\varepsilon$ $(\varepsilon>0)$, в позицию $(n,\ 1)$ жордановой клетки $J_n(1)$ сопровождается изменениями

собственных значений, по модулю равными $\varepsilon^{1/n}$. Если ту же операцию проделать для матрицы (5.1), получим матрицу с характеристическим уравнением

$$(\mu-1)^n=\varepsilon^n,$$

откуда $\mu_i=1+\omega^i\varepsilon(i=1, \ldots, n)$, ω — первообразный корень из единицы. Таким образом, возмущения собственных значений по модулю равны возмущению элемента $(n, 1)$.

Процесс изменения чувствительности собственных значений по мере увеличения анормальности матрицы описывается теоремой, установленной в 1962 г. Энрици [120]. Прежде чем формулировать эту теорему, обсудим возможные количественные характеристики анормальности.

Если исходить из определения нормальной матрицы: $A^*A=AA^*$, то естественной мерой анормальности можно считать ту или иную норму *коммутатора* $C_A\equiv[A^*, A]\equiv A^*A-AA^*$. С другой стороны, можно использовать теорему Шура. Матрица A посредством унитарного подобия приводится к треугольному виду Δ:

$$P^*AP=\Delta. \tag{5.2}$$

Представим Δ в виде суммы $\Delta=\Lambda+T$, где Λ — диагональная матрица собственных значений $\lambda_1, \ldots, \lambda_n$, а T — строго верхняя треугольная матрица. Для нормальной матрицы A (и только в этом случае) $T=0$. Следовательно, мерой анормальности может служить норма матрицы T.

Здесь необходимо сделать следующее замечание. Как мы знаем, *форма Шура* Δ определена для данной матрицы неоднозначно. И даже если фиксирована конкретная матричная норма, значения $\|T\|$ для *разных* форм Шура одной и той же (анормальной) матрицы могут не совпадать. Исключением является выбор евклидовой нормы. В этом случае

$$\|T\|_E^2=\|A\|_E^2-\sum_{i=1}^{n}|\lambda_i|^2,$$

и оба слагаемых правой части не зависят от способа приведения к форме Шура.

В качестве меры анормальности в настоящей книге используется именно норма строго треугольной части T формы Шура; как правило, это будет евклидова либо спектральная норма. Разумеется, между обеими характеристиками анормальности имеются количественные соотношения. Например,

$$\frac{\|C_A\|_E^2}{6\|A\|_E^2}\leqslant\|T\|_E^2\leqslant\left(\frac{n^3-n}{12}\right)^{1/2}\|C_A\|_E.$$

Правая оценка найдена Энрици [120], левая — Эберляйн [96].

Теорема 5.1 (теорема Энрици). *Пусть $\Delta=\Lambda+T$ — форма Шура $n\times n$-матрицы A. Для каждого собственного значения μ_i возмущенной матрицы $B=A+F$ найдется собственное значение λ_{σ_i} матрицы A такое, что*

$$|\mu_i-\lambda_{\sigma_i}|\leqslant\frac{\|T\|_2}{\psi_n(\|T\|_2/\|F\|_2)}\leqslant\frac{\|T\|_E}{\psi_n(\|T\|_E/\|F\|_E)}. \tag{5.3}$$

Здесь $\psi_n(s)$ — обозначение (единственного) положительного корня уравнения $x^n+x^{n-1}+\ldots+x=s$ $(s>0)$.

Доказательство. Начинаем рассуждения по схеме, уже хорошо известной из теорем Бауэра—Файка и 4.9. Если μ_i есть в то же время собственное значение матрицы A, то (5.3) выполнено. В противном случае матрица $A-\mu_i I$ не вырождена, хотя вырождена матрица $B-\mu_i I=A+F-\mu_i I$. Совершая с обеими матрицами подобие $X\to P^*XP$, где P—матрица из соотношения (5.2), можем сказать, что матрица $\Lambda+T-\mu_i I$ не вырождена, а матрица $\Lambda+T+G-\mu_i I$, где $G=P^*FP$, вырождена. Поэтому вырождена и матрица

$$I+(\Lambda+T-\mu_i)^{-1}G.$$

В таком случае

$$1\leqslant\|(\Lambda+T-\mu_i I)^{-1}G\|_2\leqslant\|(\Lambda+T-\mu_i I)^{-1}\|_2\|G\|_2=$$
$$=\|(I+KT)^{-1}K\|_2\|G\|_2. \quad (5.4)$$

Здесь мы положили $K=(\Lambda-\mu_i I)^{-1}$; чтобы упростить дальнейшие записи, положим еще $\|T\|_2=b$, $\|K\|_2=1/\min\limits_j|\mu_i-\lambda_j|=a$. Заметим, что матрица KT строго верхняя треугольная, поэтому $(KT)^n=0$ и матрица $I-KT+(KT)^2-\ldots+(-1)^{n-1}(KT)^{n-1}$ будет обратной для матрицы $I+KT$ (что проверяется прямым перемножением). Продолжая цепочку соотношений (5.4), имеем

$$1\leqslant(1+ab+a^2b^2+\ldots+a^{n-1}b^{n-1})a\|G\|_2=$$
$$=(ab+a^2b^2+\ldots+a^nb^n)(\|F\|_2/b). \quad (5.5)$$

Равенство $\|F\|_2=\|G\|_2$ объясняется инвариантностью спектральной нормы относительно унитарных преобразований. Итак,

$$ab+a^2b^2+\ldots+a^nb^n\geqslant b/\|F\|_2. \quad (5.6)$$

Для положительных x функция $s=x^n+x^{n-1}+\ldots+x$ монотонно возрастает, поэтому обратная функция $x=\psi_n(s)$ также монотонно возрастающая. Из (5.6) следует, что

$$ab\geqslant\psi_n(\|T\|_2/\|F\|_2),$$

т. е.

$$\min_\sigma|\mu_i-\lambda_\sigma|\leqslant\frac{\|T\|_2}{\psi_n(\|T\|_2/\|F\|_2)}.$$

Положим $\|T\|_E=c$. Поскольку $\|T\|_2\leqslant\|T\|_E$, то

$$x=\psi_n(b/\|F\|_2)\leqslant y=\psi_n(c/\|F\|_2).$$

Из равенств $x+x^2+\ldots+x^n=b/\|F\|_2$, $y+y^2+\ldots+y^n=c/\|F\|_2$ вытекает неравенство

$$b/x=\|F\|_2(1+x+\ldots+x^{n-1})\leqslant\|F\|_2(1+y+\ldots+y^{n-1})=c/y.$$

Так как $y=\psi_n(c/\|F\|_2)\geqslant\psi_n(c/\|F\|_E)$, мы приходим к правому неравенству в (5.3).

Замечание 5.2. Если $T^p=0$ при $p<n$ (так заведомо будет, например, если $t_{i,i+1}=0$ $(i=1, \ldots, n-1)$), то в оценках (5.3) индекс n можно заменить на p.

Замечание 5.3. Вид неравенств (5.3) можно упростить ценой некоторого загрубления [116, с. 201]. Если в (5.5) $a\leqslant 1$, то

$$\min_{\sigma}|\mu_i-\lambda_\sigma|=1/a\leqslant(1+b+b^2+\ldots+b^{n-1})\,\|F\|_2\equiv\theta.$$

Если же $a>1$, то

$$1\leqslant(1+b+b^2+\ldots+b^{n-1})\,a^n\,\|F\|_2,$$

откуда

$$\min_{\sigma}|\mu_i-\lambda_\sigma|\leqslant\theta^{1/n}.$$

Итак, во всех случаях

$$\min_{\sigma}|\mu_j-\lambda_\sigma|\leqslant\max\{\theta,\ \theta^{1/n}\}.$$

Замечание 5.4. Теорема Энрици не требует, чтобы возмущение F было малым, хотя практический интерес представляет только в такой ситуации.

Замечание 5.5. При малых s справедливо приближенное равенство $\psi_n(s)\approx s$. Поэтому при малой анормальности матрицы A (точнее, если мало отношение меры анормальности $\|T\|$ к норме возмущения $\|F\|$) правая часть в (5.3) приближенно равна $\|F\|$, и теорема Энрици, по существу, дает тот же результат, что и теорема Бауэра — Файка.

Пример 5.6 (см. [42, с. 162]). Для матрицы (5.1), которую можно считать собственной формой Шура, $\|T\|_2=\varepsilon$ (мы предполагаем, что $\varepsilon>0$). Если положить $\|F\|_2=\delta$, то величина $\xi=|\mu_i-1|$ не превосходит значения, определяемого равенством (см. (5.5))

$$\xi=(1+\varepsilon/\xi+\varepsilon^2/\xi^2+\ldots+\varepsilon^{n-1}/\xi^{n-1})\,\delta.$$

Ясно, что $\xi>\delta$. С другой стороны,

$$\xi<\delta\left(1+\frac{\varepsilon}{\xi}+\frac{\varepsilon^2}{\xi^2}+\ldots+\frac{\varepsilon^{n-1}}{\xi^{n-1}}+\ldots\right)=\frac{\delta}{1-\varepsilon/\xi},$$

т. е. $\xi<\delta+\varepsilon$. Итак, для любого i справедливо неравенство $|\mu_i-1|<\delta+\varepsilon$.

Этот же вывод можно получить из теоремы Бауэра — Файка, если рассматривать матрицу $J_{n,\varepsilon}(1)+F$ как результат возмущения единичной матрицы I_n. Действительно, единичная матрица нормальна, а норма суммарного возмущения, состоящего из F и наддиагональной части матрицы $J_{n,\varepsilon}$, не превосходит суммы норм, т. е. $\delta+\varepsilon$.

Замечание 5.7. Для существенно анормальных матриц теорема Энрици может давать значительно лучшие оценки, чем теорема Бауэра — Файка. Пусть, например (см. [116, с. 201]),

$$A=\begin{bmatrix}1 & 2 & 3\\ 0 & 4 & 5\\ 0 & 0 & 4.001\end{bmatrix},\quad F=\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ 0.001 & 0 & 0\end{bmatrix}.$$

Матрица P, состоящая из собственных векторов матрицы A, имеет вид

$$P=\begin{bmatrix}1 & 0.666... & -0.666...\\ 0 & 1 & 1\\ 0 & 0 & -0.0002...\end{bmatrix}.$$

Поэтому (см. § 2) $\|P\|_2\geqslant(p_{22}^2+p_{23}^2)^{1/2}=\sqrt{2}$, $\|P^{-1}\|_2>1/\min\limits_i|p_{ii}|=$ $=5000$, $\operatorname{cond}_2 P>5000\sqrt{2}$, и радиусы кругов Бауэра — Файка не меньше чем $5\sqrt{2}\approx 7.07$. В то же время $\|T\|_2<6$, $\psi_3(1000\|T\|_2)>17$, и правая часть первой оценки (5.3) меньше чем $6/17=0.35...$ В действительности матрица $B=A+F$ имеет собственные значения 1.0001..., 4.0582..., 3.9427...

Замечание 5.8. Возможно и обратное: для сильно анормальной матрицы оценка Энрици может быть много грубее оценки Бауэра — Файка. Приведем соответствующий пример из [42, с. 162]. В случае матрицы

$$A=\begin{bmatrix}1 & 2\\ 0 & 3\end{bmatrix} \tag{5.7}$$

неравенства (5.3) означают: для каждого собственного значения μ возмущенной матрицы $B=A+F(\|F\|_2=\delta)$ справедливо неравенство либо

$$|\mu-1|\leqslant 2/\psi_2(2/\delta),$$

либо

$$|\mu-3|\leqslant 2/\psi_2(2/\delta).$$

Но $\psi_2(2/\delta)=4/[\delta(1+\sqrt{1+8/\delta})]$ и $2/\psi_2(2/\delta)=(\delta+\sqrt{\delta^2+8\delta})/2$. При малых δ радиус кругов Энрици приближенно равен $(2\delta)^{1/2}$. С другой стороны, спектральное число обусловленности матрицы

$$P=\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}$$

есть $(3+\sqrt{5})/2$, и, согласно теореме Бауэра — Файка, выполняется одно из неравенств

$$|\mu-1|\leqslant(3+\sqrt{5})\delta/2,\quad |\mu-3|\leqslant(3+\sqrt{5})\delta/2. \tag{5.8}$$

Для малых δ оценки, получаемые из теоремы Энрици, очень слабы по сравнению с (5.8).

Замечание 5.9. При больших s корень $\psi_n(s)$ ведет себя приблизительно как $s^{1/n}$. Поэтому для больших отношений $\|T\|/\|F\|$ правая часть оценки Энрици приближенно равна $\|T\|^{1-1/n}\|F\|^{1/n}$. Результат, найденный выше для матрицы (5.7), есть частный случай этого соотношения.

Любопытное обобщение теоремы Энрици установлено в [128]. Пусть Q — невырожденная матрица, трансформирующая матрицу A к блочнодиагональному виду

$$Q^{-1}AQ = K = D_1 \oplus D_2 \oplus \ldots \oplus D_s. \tag{5.9}$$

Каждый диагональный блок $D_i (1 \leqslant i \leqslant s)$ имеет только одно собственное значение λ_i кратности m_i $(m_1 + m_2 + \ldots + m_s = n)$ и разным блокам отвечают разные λ. Пусть l — индекс матрицы A. Меру анормальности будем, как и выше, обозначать через $c = \| T \|_E$.

Теорема 5.10 (E. Jiang). *Для каждого собственного значения μ_i возмущенной матрицы $B = A + F$ найдется собственное значение λ_{σ_i} матрицы A такое, что*

$$|\mu_i - \lambda_{\sigma_i}| \leqslant \frac{c}{\psi_l\left(\dfrac{c}{\| Q^{-1}FQ \|_2}\right)} \leqslant \frac{c}{\psi_l\left(\dfrac{c}{\mathrm{cond}_2 Q \cdot \| F \|_2}\right)}. \tag{5.10}$$

Доказательство. Не ограничивая общности, можно считать каждый блок D_i верхней треугольной матрицей. Действительно, в противном случае мы могли бы произвести дополнительное унитарное преобразование, приводящее все D_i к форме Шура. Спектральное число обусловленности трансформирующей матрицы Q при этом не изменится.

Рассуждая, как во всех теоремах подобного типа, приходим к заключению: если μ_i не является собственным значением матрицы A, то должны выполняться неравенства

$$1 \leqslant \| (K - \mu_i I)^{-1} Q^{-1} F Q \|_2 \leqslant \| (K - \mu_i I)^{-1} \|_2 \| Q^{-1} F Q \|_2. \tag{5.11}$$

Пусть σ — номер диагонального блока, на котором достигается $\max\limits_{1 \leqslant j \leqslant s} \| (D_j - \mu_i I)^{-1} \|_2$. Тогда $\| (K - \mu_i I)^{-1} \|_2 = \| (D_\sigma - \mu_i I)^{-1} \|_2$.

Матрицу D_σ можно представить в виде суммы $\lambda_\sigma I_{m_\sigma} + H_\sigma$, где H_σ — строго верхняя треугольная матрица. Тогда

$$(D_\sigma - \mu_i I)^{-1} = (\lambda_\sigma - \mu_i)^{-1} \left(I + \frac{1}{\lambda_\sigma - \mu_i} H_\sigma\right)^{-1} = \frac{1}{\lambda_\sigma - \mu_i} \sum_{k=0}^{l-1} \left(\frac{-H_\sigma}{\lambda_\sigma - \mu_i}\right)^k.$$

Так как $\| H_\sigma \|_2 \leqslant \| H_\sigma \|_E \leqslant c$, то

$$\| (D_\sigma - \mu_i I)^{-1} \|_2 \leqslant \frac{1}{c} \sum_{k=1}^{l} \left(\frac{c}{|\lambda_\sigma - \mu_i|}\right)^k.$$

Подставляя эту оценку в (5.11), получаем

$$\sum_{k=1}^{l} \left(\frac{c}{|\lambda_\sigma - \mu_i|}\right)^k \geqslant \frac{c}{\| Q^{-1}FQ \|_2}.$$

Отсюда, как и в теореме Энрици, выводим, что

$$\frac{c}{|\lambda_\sigma - \mu_i|} \geqslant \psi_l\left(\frac{c}{\| Q^{-1}FQ \|_2}\right).$$

Это и есть нужное неравенство (5.10).

Замечание 5.11. Если матрицу A можно привести к блочно диагональному виду (5.9) посредством унитарной матрицы Q (т. е. если корневые подпространства матрицы A попарно ортогональны), то оценка (5.10) становится особенно похожей внешне на оценку Энрици

$$|\mu_i-\lambda_{\sigma_i}|\leqslant\frac{c}{\psi_l(c/\|F\|_2)}.$$

Здесь наглядней всего смысл достигнутого уточнения: при малых значениях $\varepsilon=\|F\|_2$ (и $c\neq 0$) собственные значения приобретают возмущения порядка $O(\varepsilon^{1/l})$, а не $O(\varepsilon^{1/n})$. Напомним, что индекс l есть наибольший из порядков жордановых клеток в жордановой форме матрицы A. Поэтому полученный качественный результат в действительности уже нам известен: другими средствами он был установлен в теореме 4.9.

Для нормальных матриц имеется еще один замечательный локализационный результат — теорема Виландта — Хофмана [123].

Теорема 5.12 (теорема Виландта — Хофмана). *Пусть A и B — нормальные $n\times n$-матрицы с собственными значениями соответственно $\lambda_1, \ldots, \lambda_n$ и $\mu_1, \ldots, \mu_n$. Между собственными значениями обеих матриц можно установить взаимно однозначное соответствие $\mu_i\leftrightarrow\lambda_{\tau_i}$ $(i=1, \ldots, n)$, при котором*

$$\left[\sum_{i=1}^{n}|\mu_i-\lambda_{\tau_i}|^2\right]^{1/2}\leqslant\|B-A\|_E.$$

Доказательство теоремы 5.12 можно найти помимо [123] в книгах [49, с. 368—369; 42, с. 105—108]. Во второй из этих книг рассмотрен только случай эрмитовых матриц A, B.

Теорема Виландта — Хофмана широко используется в качестве теоремы о возмущениях собственных значений при решении спектральных задач для эрмитовых и вещественных симметричных матриц. Однако если A — нормальная, но неэрмитова матрица, то условие, чтобы возмущенная матрица $B=A+F$ также была нормальна, редко удается удовлетворить. В этом случае может быть полезен следующий результат.

Теорема 5.13 (см. [192]). *Пусть A — нормальная матрица, а B — диагонализуемая $n\times n$-матрица с собственными значениями соответственно $\lambda_1, \ldots, \lambda_n$ и $\mu_1, \ldots, \mu_n$. Пусть P трансформирует B к диагональному виду: $P^{-1}BP=\mathrm{M}=\mathrm{diag}(\mu_1, \ldots, \mu_n)$. Тогда можно установить взаимно однозначное соответствие между собственными значениями обеих матриц: $\mu_i\leftrightarrow\lambda_{\tau_i}$ $(i=1, \ldots, n)$, при котором*

$$\left[\sum_{i=1}^{n}|\mu_i-\lambda_{\tau_i}|^2\right]^{1/2}\leqslant\mathrm{cond}_2 P\cdot\|B-A\|_E. \tag{5.12}$$

Пусть теперь обе матрицы A, B анормальные (и даже, возможно, недиагонализуемые); обозначим меру анормальности (относительно евклидовой нормы) каждой из них соответственно через c_A, c_B.

Теорема 5.14 (см. [182]). *Между собственными значениями* $\lambda_1, \ldots, \lambda_n$ *и* $\mu_1, \ldots, \mu_n$ *матриц* A *и* B *можно установить взаимно однозначное соответствие*: $\mu_i \leftrightarrow \lambda_{\tau_i}$ $(i=1, \ldots, n)$, *при котором*

$$\left[\sum_{i=1}^{n} |\mu_i - \lambda_{\tau_i}|^2\right]^{1/2} \leqslant c_A + c_B + \| B - A \|_E.$$

Доказательство. Пусть P_A и P_B — унитарные матрицы, трансформирующие A и B к их формам Шура:

$$P_A^* A P_A = \Delta_A = \Lambda_A + T_A,$$
$$P_B^* B P_B = \Delta_B = \Lambda_B + T_B,$$
$$\| T_A \|_E = c_A, \quad \| T_B \|_E = c_B.$$

Тогда

$$P_A \Lambda_A P_A^* - P_B \Lambda_B P_B^* = P_B T_B P_B^* + A - B - P_A T_A P_A^*.$$

В левой части последнего равенства стоит разность нормальных матриц с теми же спектрами, что и у матриц A, B. По теореме Виландта — Хофмана для некоторого упорядочения собственных значений

$$\left[\sum_{i=1}^{n} |\mu_i - \lambda_{\tau_i}|^2\right]^{1/2} \leqslant \| P_A \Lambda_A P_A^* - P_B \Lambda_B P_B^* \|_E \leqslant$$
$$\leqslant \| P_B T_B P_B^* \|_E + \| P_A T_A P_A^* \|_E + \| B - A \|_E =$$
$$= \| T_A \|_E + \| T_B \|_E + \| B - A \|_E = c_A + c_B + \| B - A \|_E.$$

ДОПОЛНЕНИЯ К § 5

1. В связи с темой настоящего параграфа естественно поставить вопрос: какова нормальная матрица, ближайшая (относительно евклидовой нормы) к данной анормальной матрице A? Можно думать, что такой матрицей будет $P\Lambda P^*$, где Λ — диагональная часть формы Шура Δ (см. (5.2)), а P — унитарная трансформирующая матрица. Другими словами, для треугольной матрицы Δ ближайшей нормальной матрицей является ее диагональ.

Оказывается [167], что это не так. Более того, если в Δ хотя бы один элемент t_{ij} $(j>i)$ отличен от нуля, то расстояние от Δ до множества $\mathcal{N}$ нормальных матриц всегда строго меньше евклидовой меры анормальности $\| T \|_E$. Матрицы (нетреугольные), диагональ которых есть ближайшая нормальная матрица, имеют весьма специальный вид; именно каждая из $n(n-1)/2$ главных подматриц такой матрицы посредством диагонального сдвига и умножения на число $e^{i\varphi}$ может быть преобразована в матрицу типа

$$\begin{bmatrix} d & s \\ -\bar{s} & -d \end{bmatrix}, \qquad d \geqslant 0.$$

Значения сдвигов и показатели φ разные для разных подматриц.

2. Наряду с мерами анормальности, введенными в основном тексте параграфа, существует множество других. Так, в [104] устанавливаются двусторонние соотношения между двумя десятками характеристик анормальности, построенных с использованием спектральной и евклидовой норм.

3. Еще более общий — по сравнению с теоремой 5.13 — случай рассмотрен в [216]: обе матрицы A, B могут быть анормальными, но приводятся

к диагональному виду посредством соответственно матриц Q и P. Доказано, что между собственными значениями обеих матриц можно установить взаимно однозначное соответствие $\mu_i \leftrightarrow \lambda_{\tau_i}$ $(i=1, ..., n)$, при котором

$$\left[\sum_{i=1}^{n} |\mu_i - \lambda_{\tau_i}|^2\right]^{1/2} \leqslant \operatorname{cond}_2 P \operatorname{cond}_2 Q \| B-A \|_E.$$

§ 6. Число обусловленности собственного значения

Напомним (см. § 2), что под обусловленностью вычислительной задачи понимают чувствительность ее решения к малому изменению входных данных. Если зависимость между решением и входной информацией описывается гладкими функциями, то в качестве меры обусловленности можно принять норму матрицы Якоби, составленной из частных производных компонент решения по отдельным элементам входа. Так, из теории алгебраических функций следует, что простой корень многочлена

$$p(a_1, ..., a_n;\ x)=x^n+a_1 x^{n-1}+ ... +a_{n-1}x+a_n$$

является аналитической функцией его коэффициентов $a_1, ..., a_n$ (покуда остается простым). При этом производная $\partial x/\partial a_i$ может быть вычислена по формуле

$$\frac{\partial x}{\partial a_i}=-\frac{\partial p}{\partial a_i}\Big/\frac{\partial p}{\partial x}=-\frac{x^{n-i}}{p'_x}.$$

Норма градиента $\left(\frac{\partial x}{\partial a_1}, \frac{\partial x}{\partial a_2}, ..., \frac{\partial x}{\partial a_n}\right)^T$ есть естественная мера обусловленности задачи вычисления данного простого корня многочлена $p(x)$. Если же вычисляются все корни и все они простые, то обусловленность такой задачи можно характеризовать нормой $n\times n$-матрицы Якоби $J=\left[\frac{\partial x_i}{\partial a_j}\right]$.

В связи с тем что задача вычисления корней многочлена математически эквивалентна задаче вычисления собственных значений матрицы (собственные значения матрицы суть корни ее характеристического многочлена; с другой стороны, по любому многочлену $p(x)$ легко составляется матрица (см. § 2), для которой $p(x)$ (или его кратное) является характеристическим многочленом), остановимся на ней несколько подробней.

Вычислительная практика показала, что корни многих многочленов очень чувствительны к малым изменениям коэффициентов. Это относится и к многочленам с вещественными и хорошо разделенными корнями — ситуация, обычно считающаяся очень благоприятной. Приведем хорошо известный пример, принадлежащий Уилкинсону (см. [42, с. 371; 48, с. 31—33]). Многочлен

$$p(x)=(x-1)(x-2)...(x-19)(x-20)=x^{20}-210x^{19}+ ... \qquad (6.1)$$

имеет простые корни 1, 2, ..., 19, 20. Значения производных $\partial x/\partial a_1$ для всех корней приведены в табл. 1. Некоторые из них по

абсолютной величине превышают 10^9. Если к значению -210 коэффициента a_1 добавить 2^{-23}, что соответствует ошибке представления числа с плавающей точкой (такой же величины, как число 210) в двоичной вычислительной машине с 30-разрядным полем для мантиссы, то корни изменятся; новые их значения, округленные до пяти десятичных разрядов после запятой (точки), приведены в табл. 2. Мы видим, что некоторые корни, первоначально вещественные и отстоявшие друг от друга не менее чем на единицу, превратились в пять комплексных сопряженных пар; при этом мнимые части их отнюдь не малы. К таким изменениям приводит ошибка $2^{-23} \approx 10^{-7}$ всего лишь в одном коэффициенте многочлена (6.1). Однако, как ни велики эти изменения, они вполне укладываются в пределы, допускаемые значениями производных в 10, ..., 19-й строках табл. 1.

Таблица 1

x	$\partial x/\partial a_1$	x	$\partial x/\partial a_1$
1	-8.2×10^{-18}	11	-4.6×10^{7}
2	8.2×10^{-11}	12	2.0×10^{8}
3	-1.6×10^{-6}	13	-6.1×10^{8}
4	2.2×10^{-3}	14	1.3×10^{9}
5	-6.1×10^{-1}	15	-2.1×10^{9}
6	5.8×10^{1}	16	2.4×10^{9}
7	-2.5×10^{3}	17	-1.9×10^{9}
8	6.0×10^{4}	18	1.0×10^{9}
9	-8.3×10^{5}	19	-3.1×10^{8}
10	7.6×10^{6}	20	4.3×10^{7}

Таблица 2

Вычисленные корни		
1.00000	6.00001	$10.09527 \pm 0.64350\,i$
2.00000	6.99970	$11.79363 \pm 1.65233\,i$
3.00000	8.00727	$13.99236 \pm 2.51883\,i$
4.00000	8.91725	$16.73074 \pm 2.81262\,i$
5.00000	20.84691	$19.50244 \pm 1.94033\,i$

Многочлен (6.1) вполне мог бы быть характеристическим многочленом некоторой вещественной симметричной матрицы A. Его корни, будучи аналитическими функциями коэффициентов, которые в свою очередь являются многочленами от a_{ij}, гладко зависят от элементов матрицы. При этом чувствительность корней к изменениям элементов матрицы совсем иная, чем к изменениям коэффициентов многочлена: согласно теореме Бауэра—Файка, изменения собственных значений симметричной матрицы по модулю не превосходят нормы матрицы-возмущения.

Наличие и распространенность подобных примеров вызвали на рубеже 1940—1950-х годов переоценку применявшихся тогда методов

численного решения спектральных задач. Эти методы были по преимуществу методами более или менее эффективного — с точки зрения числа арифметических операций — построения характеристического многочлена. Не говоря уже о том, что это построение основывалось на численно неустойчивых алгоритмах, оно заменяло спектральную задачу плохо обусловленной во многих случаях (даже при вполне приемлемом уровне обусловленности исходной задачи) задачей вычисления корней многочлена.

Тот факт, что порочность такого подхода была осознана именно в эти годы, объясняется тем, что появление электронных вычислительных машин сделало возможным решение спектральных задач значительно более высокого, чем прежде, порядка. Развиваемые с тех пор методы предусматривают вычисление собственных значений, минуя построение характеристического многочлена.

Вернемся к способам оценки обусловленности вычислительной задачи. Во многих случаях, несмотря на гладкую зависимость решения от входных данных, вычисление нужных производных затруднено; в других — соответствие между входом и решением задачи вообще недифференцируемо. Пусть, однако, удается показать, что для всех достаточно малых возмущений входной информации соответствующие возмущения решения подчиняются закону

$$\| \Delta x \| \leqslant K \| \Delta A \|. \tag{6.2}$$

Здесь через A обозначены скаляр, вектор или матрица, составляющие входные данные задачи, через x — ее решение; число K не зависит от возмущения ΔA для всех ΔA, ограниченных условием $\| \Delta A \| \leqslant \varepsilon_0$, где ε_0 — некоторое положительное число. Такое число K и можно принять за меру обусловленности задачи. Другой вариант этого подхода получается при замене абсолютных возмущений в (6.2) относительными:

$$\frac{\| \Delta x \|}{\| x \|} \leqslant C \frac{\| \Delta A \|}{\| A \|}. \tag{6.3}$$

Число обусловленности матрицы, которым мы оперировали в двух предыдущих параграфах, вводится примерно таким образом при анализе задачи решения линейной системы или задачи обращения матрицы.

Теорема Бауэра — Файка в форме (4.4) имеет как раз вид неравенства (6.2). Поэтому число $\operatorname{cond} P$ — число обусловленности матрицы P по отношению к задаче обращения — есть в то же время мера обусловленности задачи вычисления собственных значений диагонализуемой матрицы A. Однако у этой меры есть ряд недостатков. Во-первых, она не определена для недиагонализуемых матриц; между тем у таких матриц могут быть простые или, более общо, полупростые собственные значения, на поведение которых собственные значения с бо́льшими индексами не влияют. Во-вторых, в случае диагонализуемой матрицы число $\operatorname{cond} P$ оценивает суммарную обусловленность задачи вычисления *всех* собственных значений, не

улавливая различий в чувствительности разных корней. Именно по этой причине для задачи вычисления конкретного собственного значения число $\operatorname{cond} P$ не является наилучшей возможной оценкой.

Ниже мы введем более точную характеристику — так называемое *число обусловленности*,— пригодную для простого собственного значения λ произвольной матрицы A независимо от того, диагонализуема она или нет. Представим возмущенную матрицу B в виде $B(\varepsilon)=A+\varepsilon F$, где матрица F фиксирована, а число ε меняется в окрестности нуля. Тогда существует аналитическая функция $\lambda(\varepsilon)$, которая для всякого ε из указанной окрестности является собственным значением матрицы $B(\varepsilon)$, а при $\varepsilon=0$ совпадает с интересующим нас числом λ. Пока $\lambda(\varepsilon)$ остается простым собственным значением, соответствующий собственный вектор $x(\varepsilon)$ определен с точностью до скалярного множителя. Выбрав гладкое условие нормировки, мы можем обеспечить, чтобы и семейство $x(\varepsilon)$ аналитически зависело от ε. Дифференцируя равенство

$$B(\varepsilon)x(\varepsilon)=\lambda(\varepsilon)x(\varepsilon),$$

получаем

$$Fx(\varepsilon)+B(\varepsilon)x'(\varepsilon)=\lambda'(\varepsilon)x(\varepsilon)+\lambda(\varepsilon)x'(\varepsilon).$$

При $\varepsilon=0$ это дает

$$Fx+Ax'(0)=\lambda'(0)x+\lambda x'(0) \tag{6.4}$$

($x=x(0)$ — собственный вектор матрицы A для собственного значения λ). Умножая скалярно обе части равенства на левый собственный вектор y для того же собственного значения λ, имеем

$$(Fx, y)+\lambda(x'(0), y)=\lambda'(0)(x, y)+\lambda(x'(0), y),$$

откуда

$$\lambda'(0)=(Fx, y)/(x, y).$$

Мы вывели формулу для производной собственного значения λ в направлении матрицы-возмущения F. Преобразуем ее к виду

$$\lambda'(0)=\frac{\|x\|_2\|y\|_2}{(x, y)}\,\frac{(Fx, y)}{\|x\|_2\|y\|_2}, \tag{6.5}$$

преимущество которого двоякое: во-первых, разделены сомножители, зависящий и не зависящий от F; во-вторых, теперь нет нужды считать x (как и y) нормированным вектором.

Число

$$k(\lambda)=\frac{\|x\|_2\|y\|_2}{|(x, y)|} \tag{6.6}$$

называется *числом обусловленности* (простого) собственного значения λ. В советской алгебраической литературе используется также термин *коэффициент перекоса, отвечающий числу* λ. Основанием для

такого названия послужило то, что геометрически $k(\lambda)$ есть величина, обратная косинусу угла между левым и правым собственными векторами для λ.

С помощью числа обусловленности можно оценить дифференциал функции $\lambda(\varepsilon)$ в нуле:

$$|d\lambda(0)| = |\lambda'(0)|\,|\varepsilon| \leqslant k(\lambda)\,\|\varepsilon F\|_2.$$

Это неравенство, по типу аналогичное (6.2), еще раз показывает, что $k(\lambda)$ — естественная мера обусловленности задачи вычисления простого собственного значения λ.

Пусть A — матрица с простым спектром. Интересно сравнить для этого случая две возможные меры обусловленности: $\operatorname{cond} P$, где P — матрица из собственных векторов для A, и индивидуальные числа обусловленности $k(\lambda_1)$, ..., $k(\lambda_n)$. Поскольку числа $k(\lambda_i)$ определяются посредством евклидовой длины, разумно для P пользоваться спектральным числом обусловленности.

Теорема 6.1. *Справедливы неравенства*

$$\max_i k(\lambda_i) \leqslant \min_P \operatorname{cond}_2 P \leqslant \sum_{i=1}^{n} k(\lambda_i). \tag{6.7}$$

Доказательство. Для вычисления $k(\lambda_i)$ можно взять векторы p_i и q_i — столбцы с номером i соответственно в P и $Q = P^{-*}$. Тогда $(p_i, q_i) = 1$ и

$$k(\lambda_i) = \|p_i\|_2\,\|q_i\|_2 \leqslant \|P\|_2\,\|P^{-1}\|_2 = \operatorname{cond}_2 P.$$

Пусть, напротив, для каждого i выбраны нормированные правый и левый собственные векторы x_i и y_i; тогда $|(x_i, y_i)| = [k(\lambda_i)]^{-1}$. Положим

$$p_i = [k(\lambda_i)]^{1/2} x_i, \quad i = 1, \ldots, n; \qquad P = [p_1 | p_2 | \ldots | p_n].$$

Столбцы q_1, ..., q_n матрицы $Q = P^{-*}$ составляют базис, двойственный к базису p_1, ..., p_n, так что, в частности, $(p_i, q_i) = 1$. С другой стороны, каждый вектор q_i лишь скалярным множителем α_i отличается от соответствующего вектора y_i. Из равенства $|(x_i, y_i)| = [k(\lambda_i)]^{-1}$ вытекает, что $|\alpha_i| = [k(\lambda_i)]^{1/2}$. Но

$$\|P\|_2^2 \leqslant \|P\|_E^2 = \sum_{i=1}^{n} \|p_i\|_2^2 = \sum_{i=1}^{n} k(\lambda_i),$$

$$\|P^{-1}\|_2^2 \leqslant \|P^{-1}\|_E^2 = \sum_{i=1}^{n} \|q_i\|_2 = \sum_{i=1}^{n} k(\lambda_i).$$

Перемножая эти неравенства, получим правое соотношение (6.7).

Замечание 6.2. Обе меры обусловленности задачи вычисления спектра диагонализуемой матрицы — и $\operatorname{cond}_2 P$, и числа $k(\lambda_i)$ — унитарно инвариантны, т. е. сохраняют свои значения при переходе от A к унитарно подобной матрице $B = Q^* A Q$.

Пример 6.3. Числа обусловленности собственных значений одной и той же матрицы могут очень сильно различаться по величине.

Вот знаменитый пример, построенный в 1958 г. Фрэнком [110]. Все собственные значения хессенберговой матрицы

$$F_n = \begin{bmatrix} n & n-1 & n-2 & \dots & 3 & 2 & 1 \\ n-1 & n-1 & n-2 & \dots & 3 & 2 & 1 \\ & n-2 & n-2 & \dots & 3 & 2 & 1 \\ & & \dots & \dots & \dots & \dots & \dots \\ & & & & 2 & 2 & 1 \\ & & & & & 1 & 1 \end{bmatrix}$$

вещественны и положительны [119, 200]. При нечетном n имеется простое собственное значение, равное 1. Все остальные собственные значения (а в случае четного n — все собственные значения) расположены парами вида $(\lambda, 1/\lambda)$. Собственные значения, бóльшие 1, хорошо обусловлены; младшие же собственные числа уже при умеренных значениях n обусловлены очень плохо. Так, для $n=12$ числа обусловленности трех наименьших собственных значений 0.08122..., 0.04950..., 0.03102... равны соответственно $2.6... \times 10^9$, $3.8... \times 10^9$, $1.8... \times 10^9$. С ростом n обусловленность младших собственных чисел еще ухудшается. Постепенно ухудшается и обусловленность собственного значения 1. В результате при $n=27$ приближение к $\lambda=1$, вычисленное на машине CRAY-1 по программе RG пакета EISPACK, уже почти не имеет верных знаков: $\tilde{\lambda}=1.1209...$ (см. [119]). Дело здесь не в программе, известной своим высоким качеством, а в чувствительности собственного значения к малым возмущениям элементов матрицы.

Пусть x и y — нормированные правый и левый собственные векторы, относящиеся к одному и тому же собственному значению λ. Положим

$$s(\lambda) \equiv |(x, y)| = [k(\lambda)]^{-1}. \tag{6.7a}$$

В выкладках, приведших к формуле (6.6), мы неявно опирались на то, что $s(\lambda) \neq 0$. Это действительно так, если λ простое. В самом деле, вектор x, будучи ортогонален левым корневым подпространствам для всех прочих собственных значений, не может быть ортогонален еще и вектору y: в этом случае он был бы ортогонален всему пространству, что возможно лишь при $x=0$.

Иная ситуация с кратными собственными значениями.

Теорема 6.4. *Если λ — кратное собственное значение матрицы A, то для него найдутся ортогональные правый и левый собственные векторы.*

Доказательство. Обозначим через m алгебраическую кратность собственного значения λ, через l его индекс. Пусть вначале $l=1$. Фиксируем произвольный левый собственный вектор y для λ. Если $x_1, \dots, x_m$ — линейно независимые правые собственные векторы, относящиеся к λ, то искомой парой будут векторы (x, y), где $x=\alpha_1 x_1 + \dots + \alpha_m x_m$ и

$$\alpha_1 (x_1, y) + \dots + \alpha_m (x_m, y) = 0.$$

При $m \geqslant 2$ из этого условия всегда можно определить нетривиальный набор $\alpha_1, \dots, \alpha_m$.

Пусть теперь $l>1$, и пусть в жордановой форме матрицы A

$$P^{-1} A P = J$$

первой стоит жорданова клетка порядка l с числом λ на диагонали. Тогда 1-й столбец матрицы P и l-й столбец матрицы $Q=P^{-*}$ суть правый и левый собственные векторы для λ. Из соотношений двойственности следует, что они ортогональны.

Итак, если $s(\lambda)=0$, то соответствующее собственное значение λ необходимо кратное. Любопытно, что справедлив следующий асимптотический вариант этого утверждения.

Теорема 6.5 (см. [209]). *Пусть x и y — нормированные правый и левый собственные векторы матрицы A, относящиеся к простому собственному значению λ, и пусть $s(\lambda)=\varepsilon<1$. Тогда найдется матрица B, для которой λ — кратное собственное значение и при этом*

$$\|B-A\|_2 \leqslant \frac{\varepsilon}{\sqrt{1-\varepsilon^2}}\|A\|_2. \tag{6.8}$$

Доказательство. Пусть P — унитарная матрица, первым столбцом которой является x. Тогда после преобразования $C=P^*AP$ получим матрицу вида

$$C=\begin{bmatrix}\lambda & c^*\\ 0 & C_{n-1}\end{bmatrix}, \tag{6.9}$$

где $c\in\mathbf{C}^{n-1}$, а C_{n-1} — квадратная матрица порядка $n-1$. Нормированным правым собственным вектором матрицы (6.9) для собственного значения λ можно считать координатный вектор e_1; нормированный левый собственный вектор w можно представить как

$$w=\begin{pmatrix}\varepsilon\\ z\end{pmatrix}, \qquad z\in\mathbf{C}^{n-1}, \qquad \|z\|_2=\sqrt{1-\varepsilon^2}.$$

Тот факт, что первая компонента равна ε, объясняется инвариантностью скалярного произведения при унитарных подобиях и возможностью изменить его аргумент за счет умножения векторов на числа с модулем 1. Так как левый собственный вектор w есть не что иное, как обычный собственный вектор (для собственного значения $\bar{\lambda}$) матрицы C^*, должно выполняться равенство

$$c\varepsilon+C_{n-1}^*z=\bar{\lambda}z,$$

или

$$\left(C_{n-1}^*+\frac{\varepsilon}{1-\varepsilon^2}cz^*\right)z=\bar{\lambda}z. \tag{6.10}$$

Если положить $\hat{C}_{n-1}=C_{n-1}+\dfrac{\varepsilon}{1-\varepsilon^2}zc^*$, то, согласно (6.10), λ является собственным значением для $\hat{C}_{n-1}$ и по меньшей мере двукратным собственным значением для матрицы

$$\hat{C}=\begin{bmatrix}\lambda & c^*\\ 0 & \hat{C}_{n-1}\end{bmatrix}.$$

При этом

$$\| \hat{C} - C \|_2 = \| \hat{C}_{n-1} - C_{n-1} \|_2 = \left\| \frac{\varepsilon}{1-\varepsilon^2} z c^* \right\|_2 = \frac{\varepsilon}{1-\varepsilon^2} \| z \|_2 \| c \|_2 =$$

$$= \frac{\varepsilon}{\sqrt{1-\varepsilon^2}} \| c \|_2 \leqslant \frac{\varepsilon}{\sqrt{1-\varepsilon^2}} \| C \|_2 = \frac{\varepsilon}{\sqrt{1-\varepsilon^2}} \| A \|_2.$$

Совершая обратное преобразование $A = PCP^*$ и учитывая унитарную инвариантность спектральной нормы, видим, что в качестве B можно взять матрицу $P\hat{C}P^*$.

Итак, если некоторое собственное значение плохо обусловлено, то это всегда связано с наличием близкой матрицы, имеющей кратный корень.

Пример 6.6 (см. [42, с. 93]). Для матрицы

$$A = \begin{bmatrix} 20 & 20 & & & & & \\ & 19 & 20 & & & & \\ & & 18 & 20 & & & \\ & & & \ddots & \ddots & & \\ & & & & \ddots & \ddots & \\ & & & & & 2 & 20 \\ & & & & & & 1 \end{bmatrix}$$

правый и левый собственные векторы, отвечающие собственному значению $\lambda_r = r$, имеют компоненты соответственно

$$x_r = \left(1, \frac{20-r}{-20}, \frac{(20-r)(19-r)}{20^2}, \ \ldots, \ \frac{(20-r)!}{(-20)^{20-r}}, \ 0, \ \ldots, \ 0\right)^{\tau},$$

$$y_r = \left(0, \ \ldots, \ 0, \ \frac{(r-1)!}{20^{r-1}}, \ \ldots, \ \frac{(r-1)(r-2)}{20^2}, \ \frac{r-1}{20}, \ 1\right)^{\tau}.$$

Эти векторы нормированы в норме $\| \cdot \|_\infty$, но, учитывая быстрое убывание компонент, мы можем при оценке порядка числа обусловленности игнорировать различие между нормами $\| \cdot \|_\infty$ и $\| \cdot \|_2$. Тогда

$$k(r) \approx 20^{19} / [(20-r)!\,(r-1)!].$$

Наибольшее значение, равное приблизительно 4×10^{12}, эта величина принимает при $r = 10$ и $r = 11$.

Внесем в позицию (20, 1) матрицы A возмущение ε с таким расчетом, чтобы возмущенная матрица B имела двойное собственное значение $\mu = 10.5$ — результат слияния двух ближайших корней $\lambda_{10} = 10$ и $\lambda_{11} = 11$. Характеристическим уравнением матрицы B будет

$$(20-\lambda)(19-\lambda)\ldots(2-\lambda)(1-\lambda) = 20^{19}\varepsilon.$$

Поэтому нужное ε определяется равенством

$$(0.5 \cdot 1.5 \cdot \ldots \cdot 9.5)^2 = 20^{19}\varepsilon,$$

откуда $\varepsilon \approx 8 \times 10^{-14}$. На таком расстоянии от A находится матрица B с кратным собственным значением. Это расстояние значительно

меньше значения, указываемого оценкой (6.8). Дело не только в приблизительности оценки, но и в том, что мы приняли в качестве кратного корня матрицы B число, промежуточное между 10 и 11. В теореме 6.5 кратным становилось собственное значение самой матрицы A (для нашего примера 10 или 11), что требует, вообще говоря, большего возмущения.

Треугольный определитель порядка $n-1$, стоящий в правом верхнем углу матрицы $B-\mu I$, не равен нулю; следовательно, $\operatorname{def}(B-\mu I)=1$, и геометрическая кратность собственного значения μ меньше алгебраической. В жордановой форме матрицы B имеется клетка порядка 2 с числом μ на диагонали.

Мы подробно разобрали вопрос об обусловленности простого собственного значения. А что можно сказать о кратных?

Вспомним (см. § 4), что на возмущения матричных элементов, имеющие порядок ε, собственное значение индекса m реагирует возмущениями порядка $\varepsilon^{1/m}$. Если исходить из определения обусловленности, связанного с (6.2), то число обусловленности собственного значения с индексом $m>1$ следует считать бесконечным.

Пусть теперь λ — полупростое собственное значение алгебраической кратности $l>1$.

Теорема 6.7. *В некоторой окрестности O нуля существуют l функций $\lambda_1(\varepsilon)$, ..., $\lambda^l(\varepsilon)$ таких, что: а) для любого $\varepsilon\in O$ $\lambda_i(\varepsilon)$ $(i=1, \ldots, l)$ — собственное значение матрицы $B(\varepsilon)=A+\varepsilon F$; б) при $\varepsilon\to 0$ справедливы разложения*

$$\lambda_i(\varepsilon)=\lambda_i+a_i\varepsilon+O\left(|\varepsilon|^{1+1/l}\right), \tag{6.11}$$

причем a_1, ..., a_l суть собственные значения матрицы $P_\lambda F P_\lambda$, где P_λ — спектральный проектор матрицы A, отвечающий λ.

Доказательство теоремы 6.7 дано в [29, с. 231].

Таким образом, кратное собственное значение λ кратности $l>1$ порождает l непрерывных ветвей, гладко входящих при $\varepsilon=0$ в λ. При этом значением производной в нуле является соответствующее собственное число матрицы $P_\lambda F P_\lambda$.

Замечание 6.8. Для простого собственного значения λ с правым и левым собственными векторами x, y, нормированными условием $(x, y)=1$, матрица $P_\lambda F P_\lambda$ имеет вид

$$xy^*Fxy^*=(y^*Fx)xy^*.$$

Единственное ненулевое собственное число этой матрицы равно $a=(y^*Fx)(y^*x)=(Fx, y)$. Подставляя значение a в (6.11), снова приходим к формуле (6.5) для $\lambda'(0)$.

ДОПОЛНЕНИЯ К § 6

1. Формулы для чисел обусловленности различных алгебраических задач (в том числе и задачи вычисления простого собственного значения) даны в [113]. Как и в (6.3), берутся относительные возмущения. Среди прочих рассматривается выбор специальной нормы, учитывающей относительные возмущения отдельных компонент входа.

2. Если λ — простое собственное значение, x и y — соответствующие правый и левый собственные векторы, причем $(x, y)=1$, то для спектральиого проектора $P_\lambda = xy^*$ имеем

$$P_\lambda P_\lambda^* = xy^* yx^* = \| y \|_2^2 xx^*,$$

$$\| P_\lambda \|_2 = \left[\lambda_{\max}(P_\lambda P_\lambda^*)\right]^{1/2} = \| y \|_2 \left[\lambda_{\max}(xx^*)\right]^{1/2} = \| y \|_2 \| x \|_2 = \frac{\| y \|_2 \| x \|_2}{(x, y)} = k(\lambda).$$

3. Пусть A — матрица с простым спектром $\lambda_1, \ldots, \lambda_n$. Поскольку (см. § 1)

$$P_{\lambda_1} + \ldots + P_{\lambda_n} = I,$$

то

$$k(\lambda_i) = \| P_{\lambda_i} \|_2 = \| I - \sum_{j \neq i} P_{\lambda_j} \|_2 \leqslant 1 + \sum_{j \neq i} \| P_{\lambda_j} \|_2 = 1 + \sum_{j \neq i} k(\lambda_j).$$

Отсюда вытекает следующий важный вывод [210]: если какое-то собственное зиачение λ_i очень плохо обусловлено, т. е. число $k(\lambda_i)$ очень велико, то среди прочих $k(\lambda_j)$ найдется хотя бы еще одно большое число. Иными словами, не может быть так, что у матрицы плохо обусловлено *только одно* собственное значение.

4. Оценка (6.8) для (спектрального) расстояния от матрицы A с простым спектром до множества матриц, имеющих кратные собственные значения, не учитывает возможную близость собственных чисел самой матрицы A и плохую обусловлениость по крайней мере еще одного (кроме λ) собственного значения $\hat{\lambda}$. Кроме того, для матрицы B, входящей в эту оценку, кратным является собственное значение λ. Между тем ближе к A может оказаться матрица с кратным корнем μ, находящимся, например, в точке $\mu = (\lambda + \hat{\lambda})/2$. Для того чтобы попасть в μ, числу $\hat{\lambda}$ нужно проделать меньший путь, чем при необходимости идти до λ. Все эти соображения обсуждаются в [210].

5. В [200] показано, что собственные значения λ матрицы Фрэнка F_n связаны с собственными значениями μ симметричной трехдиагональной матрицы

$$S = \begin{bmatrix} 0 & \sqrt{n-1} & & & & & & \\ \sqrt{n-1} & 0 & \sqrt{n-2} & & & & & \\ & \sqrt{n-2} & 0 & \ddots & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & \ddots & 0 & \sqrt{2} & & \\ & & & & \sqrt{2} & 0 & 1 & \\ & & & & & 1 & 0 & \end{bmatrix} \tag{6.12}$$

соотношением

$$\mu = (\lambda - 1)/\sqrt{\lambda}.$$

Спектр матрицы S симметричен относительно иуля. Каждой паре чисел μ_i, $-\mu_i$ отвечают числа

$$\lambda_i^+ = \left(\frac{\mu_i}{2} + \sqrt{\left(\frac{\mu_i}{2}\right)^2 + 1}\right)^2, \qquad \lambda_i^- = \left(\frac{\mu_i}{2} - \sqrt{\left(\frac{\mu_i}{2}\right)^2 + 1}\right)^2,$$

для которых $\lambda_i^+\lambda_i^- = 1$. Там же, в [200], даны нижние оценки чисел обусловленности, выраженные в терминах собственных значений λ_i, μ_i, крайних компонент собственных векторов матрицы S и некоторых функций от n.

В действительности в [200] рассмотрен более общий класс хессенберговых матриц, включающий в себя F_n и порождаемый по определенным правилам симметричными трехдиагональными матрицами.

Любопытно отметить, что с точностью до множителя $\sqrt{2}$ матрица (6.12) составлена из коэффициентов трехчленных формул, связывающих многочлены Эрмита $H_n(x)$ разных порядков.

§ 7. Какую информацию дает невязка?

Предположим, что число μ и нормированный в евклидовой метрике вектор p рассматриваются как приближенная собственная пара матрицы A. Качество этих приближений можно оценивать посредством *невязки*

$$r \equiv r(\mu, p) = Ap - \mu p. \tag{7.1}$$

Действительно, точной собственной паре отвечает нулевая невязка. Естественно думать поэтому, что малая длина вектора r свидетельствует о малой погрешности в μ и p или хотя бы только в μ.

Для эрмитовых матриц это действительно так.

Теорема 7.1. *Если $A = A^*$, то найдется собственное значение λ матрицы A такое, что*

$$|\lambda - \mu| \leqslant \|r\|_2. \tag{7.2}$$

Доказательство теоремы можно найти, например, в [35, § 4.5]. Но можно рассматривать ее и как частный случай доказываемой ниже теоремы 7.2.

Теорема 7.2. *Если A — диагонализуемая матрица, причем $P^{-1}AP = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, то найдется собственное значение λ_i такое, что*

$$|\lambda_i - \mu| \leqslant \operatorname{cond}_2 P \, \|r\|_2. \tag{7.3}$$

Доказательство. Если μ — собственное значение матрицы A, то доказывать нечего. В противном случае, исходя из равенств

$$p = (A - \mu I)^{-1}(A - \mu I)p, \qquad (A - \mu I)^{-1} = P^{-1}(\Lambda - \mu I)^{-1}P,$$

имеем

$$1 = \|p\|_2 \leqslant \|(A - \mu I)^{-1}\|_2 \, \|(A - \mu I)p\|_2 \leqslant$$

$$\leqslant \|P^{-1}\|_2 \, \|(\Lambda - \mu I)^{-1}\|_2 \, \|P\|_2 \, \|r\|_2 = \frac{1}{\min\limits_i |\mu - \lambda_i|} \operatorname{cond}_2 P \, \|r\|_2;$$

отсюда и следует (7.3).

Хотя для эрмитовой и, более обшо, для нормальной матрицы A неравенство (7.3) переходит в (7.2), в других ситуациях пользоваться им затруднительно. При решении частичной проблемы собственных значений мы, как правило, не имеем не только приближения к матрице P, но и сколько-нибудь обоснованной оценки для числа $\operatorname{cond}_2 P$.

Для плохо обусловленной матрицы P правая часть оценки (7.3) не будет мала даже при малой невязке r. Понятно, что еще хуже обстоит дело в случае недиагонализуемой матрицы A. Приведем в связи с этим поучительный пример, заимствованный из [132].

Пример 7.3. Элементы $n\times n$-матрицы A определяются формулами

$$a_{ij}=\begin{cases}0, & i\geqslant j,\\ 1, & i<j.\end{cases}$$

Жорданова форма этой матрицы есть жорданова клетка $J_n(0)$.

Примем в качестве μ число 1, в качестве приближенного собственного вектора — вектор $\hat{p}=(2^{n-2},\ 2^{n-3},\ ...,\ 2,\ 1,\ 1)^{\mathrm{T}}$. Тогда

$$A\hat{p}-1\cdot\hat{p}=(0,\ ...,\ 0,\ -1)^{\mathrm{T}}=\hat{r}. \tag{7.4}$$

Однако о величине невязки осмысленно можно судить, только относя ее длину к длине порождающего вектора $\hat{p}$. Иными словами, обе части в (7.4) нужно разделить на $\|\hat{p}\|_2=[(4^{n-1}+2)/3]^{1/2}$. Итак,

$$\|r\|_2=\|\hat{r}\|_2/\|\hat{p}\|_2=\left[(4^{n-1}+2)/3\right]^{-1/2},$$

и уже при умеренных n это — очень малое число, несмотря на то что μ отстоит на 1 от единственного собственного значения 0 матрицы A.

Заметим, что для этого же «приближения» $\mu=1$ будет мала и *левая невязка* s:

$$q^*A-1\cdot q^*=s^*,$$

отвечающая приближенному левому собственному вектору $q=\hat{q}/\|\hat{q}\|_2$, где

$$\hat{q}=(1,\ 1,\ 2,\ ...,\ 2^{n-3},\ 2^{n-2})^{\mathrm{T}},\quad \|\hat{q}\|_2=\|\hat{p}\|_2.$$

Действительно, $s=(-1,\ 0,\ ...,\ 0)^{\mathrm{T}}\cdot\|\hat{q}\|_2^{-1}$ и $\|s\|_2=\|r\|_2$.

И все же, несмотря на примеры, подобные рассмотренному, для многих анормальных матриц невязка дает полезную информацию о погрешности приближенного собственного значения. В пояснение этого тезиса начнем со следующего утверждения.

Теорема 7.4. *Пусть* μ, p *— приближенная собственная пара для матрицы* A, *причем вектор* p *нормированный и* r *— соответствующая невязка. Найдется матрица* F *такая, что*: *а)* $\|F\|_2=\|r\|_2$; *б)* μ *и* p *— точные собственное значение и собственный вектор для матрицы* $B=A+F$.

Доказательство. Из равенства $Ap-\mu p=r$ следует $\mu p=Ap-r=$ $=Ap-r(p^*p)=Ap-(rp^*)p=(A-rp^*)p$. Полагая $F=-rp^*$, имеем

$$\|F\|_2=\|r\|_2\,\|p\|_2=\|r\|_2.$$

Если для числа μ известен не только приближенный правый собственный вектор p, но и приближенный левый собственный вектор q, то теорему 7.4 можно усилить.

Теорема 7.5 (см. [132]). *Пусть* μ, p, q *— приближенная собственная тройка для матрицы* A, *причем векторы* p *и* q *нормированы,*

r и s — соответствующие им невязки и $(p, q) \neq 0$. Найдется матрица F такая, что: а) $\|F\|_2 = \max\{\|r\|_2, \|s\|_2\}$; б) μ, p и q — точные собственное значение и правый и левый собственные векторы для матрицы $B = A + F$.

Доказательство. Искомая матрица F должна одновременно удовлетворять условиям

$$(A+F)p = \mu p, \qquad q^*(A+F) = \mu q^*. \tag{7.5}$$

Построим какие-либо унитарные матрицы P и Q, первыми столбцами которых являются соответственно p и q: $P = [p|\check{P}]$, $Q = [q|\check{Q}]$. Матрицу $Z = -Q^*FP$ представим в соответствующей блочной форме:

$$Z = \begin{bmatrix} z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}. \tag{7.6}$$

Тогда, согласно формулам (7.5),

$$\begin{aligned} z_{11} &= q^*(-Fp) = q^*(Ap - \mu p) = q^*r = (-q^*F)p = s^*p, \\ Z_{12} &= (-q^*F)\check{P} = s^*\check{P}, \\ Z_{21} &= \check{Q}^*(-Fp) = \check{Q}^*r. \end{aligned} \tag{7.7}$$

Если теперь мы, наоборот, определим z_{11}, Z_{12}, Z_{21} правыми частями соотношений (7.7) — при произвольном выборе $n \times (n-1)$-матриц $\check{P}$ и $\check{Q}$ с ортонормированными столбцами, ортогональными к p и q соответственно, и при *произвольном* выборе клетки Z_{22}, то, положив

$$F = -QZP^*, \tag{7.8}$$

обеспечим выполнение условий (7.5). Действительно, в силу (7.7) первый столбец z_1 матрицы Z равен Q^*r. Переписывая (7.8) в виде $FP = -QZ$ и приравнивая первые столбцы, получаем $Fp = -QQ^*r = \mu p - Ap$. Аналогично проверяется второе условие (7.5).

Пока что показана возможность выбора F в соответствии с утверждением б). Неоднозначность выбора используем для минимизации нормы этой матрицы. Поскольку спектральная норма унитарно инвариантна, достаточно минимизировать норму матрицы Z. Евклидова длина первого столбца $z_1 = Q^*r$ равна $\|r\|_2$. Точно так же евклидова длина первой строки, т. е. произведения s^*P, равна $\|s\|_2$. Поэтому $\|Z\|_2 \geqslant \max\{\|r\|_2, \|s\|_2\}$, как бы мы ни выбирали блок Z_{22}.

Для завершения доказательства нам достаточно сослаться на следующий результат, установленный в [131]:

$$\min_{Z_{22}} \left\| \begin{bmatrix} z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \right\|_2 = \max\left\{ \left\| \begin{matrix} z_{11} \\ Z_{21} \end{matrix} \right\|_2, \ \| z_{11} \quad Z_{12} \|_2 \right\} = \max\{\|r\|_2, \ \|s\|_2\}. \tag{7.9}$$

Для симметричной матрицы Z доказательство формулы (7.9) можно найти в [35, § 11.11].

Замечание 7.6. В действительности экстремальное соотношение (7.9) доказано в [131] для случая, когда z_{11} — клетка произвольного

порядка k $(1 \leqslant k < n)$. Впрочем, и сама теорема 7.5 может быть перенесена на ситуацию, когда заданы группа приближенных собственных значений и соответствующие приближенные правые и левые собственные векторы. Об этом мы поговорим в п. 4 дополнений к § 8.

Обсудим теперь практическое значение доказанной теоремы. Предположим, что у нас есть основания считать μ простым собственным значением матрицы B. Если обе невязки r и s малы, то мала и матрица F. Будем считать исходную матрицу A возмущением матрицы B. Тогда при переходе от B к A собственное значение μ претерпевает возмущение, линейный член которого ограничен по модулю величиной $k(\mu)\|F\|_2$, где, согласно (6.6),

$$k(\mu) = 1/|(p, q)|. \tag{7.10}$$

Итак, хотя мы и не можем дать строгую оценку погрешности приближенного собственного значения μ, примерную ее величину определить можно. Вот почему при решении спектральных задач желательно вычисление не только правых, но и левых собственных векторов, даже если сама по себе постановка задачи этого не требует.

Замечание 7.7. Число $k(\mu)$ из формулы (7.10) есть в то же время приближенное значение числа обусловленности $k(\lambda)$ собственного значения λ матрицы A, соответствующего приближению μ.

§ 8*. Об обусловленности собственных векторов и инвариантных подпространств

Изучение обусловленности собственных векторов начнем со случая, когда матрица A имеет простой спектр $\lambda_1, \ldots, \lambda_n$. Пусть $x_1, \ldots, x_n$ — соответствующие собственные векторы единичной евклидовой длины. Как и в § 6, можно считать, что в некоторой окрестности нуля существуют скалярные функции $\lambda_1(\varepsilon), \ldots, \lambda_n(\varepsilon)$ и векторные функции $x_1(\varepsilon), \ldots, x_n(\varepsilon)$, аналитические в этой окрестности, представляющие собой собственные значения и собственные векторы матрицы $B(\varepsilon) = A + \varepsilon F$ и совпадающие при $\varepsilon = 0$ с $\lambda_1, \ldots, \lambda_n$ и $x_1, \ldots, x_n$. При $\varepsilon \neq 0$, однако, мы не будем требовать нормированности векторов $x_i(\varepsilon)$.

Дифференцируя равенство $B(\varepsilon)x_m(\varepsilon) = \lambda_m(\varepsilon)x_m(\varepsilon)$ и полагая затем $\varepsilon = 0$, получаем соотношение, аналогичное (6.4):

$$Ax'_m(0) + Fx_m = \lambda'_m(0)x_m + \lambda_m x'_m(0).$$

Подставляя сюда разложение вектора $x'_m(0)$:

$$x'_m(0) = \sum_{i=1}^{n} \alpha_i x_i, \tag{8.1}$$

находим

$$\sum_{i=1}^{n} \alpha_i(\lambda_i - \lambda_m)x_i + Fx_m = \lambda'_m(0)x_m.$$

Умножая скалярно обе части на нормированный левый собственный вектор y_i $(i \neq m)$, подобранный так, чтобы скалярное произведение

(x_i, y_i) было положительным, и учитывая соотношения биортогональности, приходим к формуле для коэффициентов α_i:

$$\alpha_i = (Fx_m, y_i)/[(\lambda_m - \lambda_i)(x_i, y_i)], \qquad i \neq m.$$

Поскольку собственные векторы определены с точностью до скалярного множителя, равенство (8.1) не позволяет вычислить α_m. Можно положить $\alpha_m = 0$. Тогда производная вектора $x_m(\varepsilon)$ при $\varepsilon = 0$ имеет вид

$$x'_m(0) = \sum_{\substack{i=1 \\ i \neq m}}^{n} \frac{(Fx_m, y_i)}{\lambda_m - \lambda_i} k(\lambda_i) x_i. \tag{8.2}$$

Эта формула показывает, что вектор x_m может быть очень чувствительным к малым возмущениям элементов матрицы при наличии плохо обусловленных собственных значений (множители $k(\lambda_i)$); кроме того, на чувствительность влияет наличие собственных значений, близких к λ_m (знаменатели $\lambda_m - \lambda_i$).

Оба указанных фактора связаны. В силу теоремы 6.5 матрица A, имеющая плохо обусловленное собственное значение, близка к матрице B, в спектре которой есть кратные точки. В как угодно малой окрестности последней можно найти матрицу с простыми собственными значениями, среди которых есть очень близкие. Таким образом, можно сказать, что плохая обусловленность собственного вектора матрицы A всегда связана либо с наличием у A близких собственных значений, либо с близостью A к матрице с этим свойством.

Заметим, что, хотя число обусловленности $k(\lambda_m)$ не входит в формулу (8.2), плохая обусловленность λ_m сказалась бы в присутствии еще по крайней мере одного плохо обусловленного собственного значения (см. п. 3 дополнений к § 6). В этом случае сумма в правой части (8.2) содержит хотя бы одно большое слагаемое (разве что для данной матрицы-возмущения F скалярное произведение в числителе окажется малым).

Пример 8.1 (см. [116]). Матрица

$$A = \begin{bmatrix} 1.01 & 0.01 \\ 0 & 0.99 \end{bmatrix}$$

имеет правые собственные векторы $x_1 = (1, 0)^T$, $x_2 = (-0.5, 1)^T$ и левые собственные векторы $y_1 = (1, 0.5)^T$, $y_2 = (0, 1)^T$. Векторы y_1 и x_2 выбраны из условий $(x_1, y_1) = 1$, $(x_2, y_2) = 1$. Оба числа обусловленности $k(1.01)$, $k(0.99)$ равны $\sqrt{5}/2 \approx 1.118$, т. е. собственные значения обусловлены хорошо. Это и естественно, поскольку матрица A получена малым возмущением диагональной матрицы. В то же время для близкой матрицы

$$A + F = \begin{bmatrix} 1.01 & 0.01 \\ 0 & 1 \end{bmatrix}$$

вектор $\tilde{x}_2$ имеет вид $(-1, 1)^T$. И после нормировки векторы x_2, $\tilde{x}_2$ сильно различаются, что объясняется близостью собственных значений.

При выводе формулы (8.2) предполагалась простота спектра матрицы A. Однако если λ_m — простое собственное значение, то, как отмечалось в § 6, в малой окрестности нуля $\lambda_m(\varepsilon)$ и $x_m(\varepsilon)$ являются аналитическими функциями от ε независимо от того, просты или нет прочие собственные значения матрицы A. Поэтому можно получить формулу для производной $x'_m(0)$, справедливую во всех случаях (см., например, [42, гл. 2, § 24; 113, теорема 3.1]). Правда, эта формула не будет иметь столь прозрачного истолкования, как (8.2).

То, что собственные векторы, отвечающие близким собственным значениям, *должны* быть очень чувствительны к малым возмущениям матрицы, становится вполне понятным в случае, если рассмотреть предельную ситуацию: матрица имеет кратное собственное значение λ, которому соответствует собственное подпространство $\mathscr{L}_\lambda$ размерности $l \geqslant 2$. Пусть $x_1, \ldots, x_l$ — базис этого подпространства. Если возмущение εF не меняет ни λ, ни собственное подпространство $\mathscr{L}_\lambda$, то ниоткуда не следует, что для матрицы $B = A + \varepsilon F$ выбранный по произволу базис $\tilde{x}_1, \ldots, \tilde{x}_l$ подпространства $\mathscr{L}_\lambda$ будет близок к первоначальному базису $x_1, \ldots, x_l$.

Объединим в одну группу близкие собственные значения матрицы A так, чтобы эта группа, называемая *кластером*, была достаточно удалена от прочих собственных значений. Если следить за поведением собственных векторов, ассоциированных с числами из кластера, то можно заметить следующую закономерность: хотя отдельные собственные векторы из рассматриваемой совокупности сильно меняются при малых возмущениях матрицы, натянутое на них подпространство мало чувствительно к этим возмущениям. Так будет и для нормальных, и для анормальных матриц.

Пример 8.2 (см. [185]). Для матрицы $A = \operatorname{diag}(1, 1, 2)$ рассмотрим две возмущенные матрицы:

$$B_1 = \frac{1}{1+2\varepsilon^2}\begin{bmatrix} 1+3\varepsilon^2+\varepsilon^3 & \varepsilon-\varepsilon^2+\varepsilon^3 & -\varepsilon-\varepsilon^2 \\ \varepsilon-\varepsilon^2+\varepsilon^3 & 1+3\varepsilon^2+\varepsilon^3 & \varepsilon+\varepsilon^2 \\ -\varepsilon-\varepsilon^2 & \varepsilon+\varepsilon^2 & 2+2\varepsilon^2-2\varepsilon^3 \end{bmatrix},$$

$$B_2 = \frac{1}{5(1+5\varepsilon^2)}\begin{bmatrix} 5-3\varepsilon+30\varepsilon^2-20\varepsilon^3 & 4\varepsilon+10\varepsilon^2+10\varepsilon^3 & 5\varepsilon(1-\varepsilon) \\ 4\varepsilon+10\varepsilon^2+10\varepsilon^3 & 5+3\varepsilon+45\varepsilon^2-5\varepsilon^3 & 10\varepsilon(1-\varepsilon) \\ 5\varepsilon(1-\varepsilon) & 10\varepsilon(1-\varepsilon) & 10+25\varepsilon^2+25\varepsilon^3 \end{bmatrix}.$$

Обе матрицы имеют одни и те же собственные значения $1+\varepsilon$, $1-\varepsilon$, 2. Собственные векторы обеих матриц, нормированные так, чтобы наибольшие по модулю компоненты были равны 1, образуют (по столбцам) матрицы

$$P_1 = \begin{bmatrix} 1 & 1 & -\varepsilon \\ 1 & -1 & \varepsilon \\ 0 & 2\varepsilon & 1 \end{bmatrix}, \qquad P_2 = \begin{bmatrix} \frac{1}{2} & 1 & \varepsilon \\ 1 & -\frac{1}{2} & 2\varepsilon \\ -\frac{5}{2}\varepsilon & 0 & 1 \end{bmatrix}.$$

В то время как собственные векторы для $\lambda=2$ близки друг к другу (и к координатному вектору e_3), собственные векторы матриц B_1 и B_2, отвечающие $\lambda=1+\varepsilon$ или $\lambda=1-\varepsilon$, различаются между собой величинами порядка 1. Однако для обеих матриц подпространство, натяиутое на два первых собственных вектора, близко к координатному подпространству векторов e_1 и e_2.

Таким образом, если у матрицы есть кластер, то больший смысл имеет вычисление не отдельных собственных векторов для входящих в кластер собственных значений, а всего инвариантного подпространства $\mathscr{L}$, порождаемого ими. При таком подходе мы можем вместо плохо обусловленного базиса подпространства $\mathscr{L}$, образуемого собственными векторами, искать хорошо обусловленный, например ортонормированный, базис.

Остальная часть параграфа посвящена вопросу об изменении инвариантных подпространств при малых возмущениях матрицы. Ее можно рассматривать как изложение результатов очень важной работы [185], а также работ [89, 183].

Мы начнем с обсуждения способов описания близости подпространств. Классическим понятием, введенным в [16], является *раствор подпространств*. Пусть $\mathscr{L}$, $\mathscr{M}$ — заданные подпространства с единичными сферами соответственно $S_{\mathscr{L}}$ и $S_{\mathscr{M}}$. Раствором подпространств $\mathscr{L}$ и $\mathscr{M}$ называется число

$$\delta(\mathscr{L}, \mathscr{M})=\max\{\sup_{x\in S_{\mathscr{L}}}\inf_{y\in\mathscr{M}}\|x-y\|,\quad \sup_{y\in S_{\mathscr{M}}}\inf_{x\in\mathscr{L}}\|y-x\|\}. \tag{8.3}$$

Если $\mathscr{L}=\{0\}$, считаем, что $S_{\mathscr{L}}=\{0\}$. Внутренние инфимумы в формуле (8.3) суть расстояние $d(x, \mathscr{M})$ от вектора $x\in S_{\mathscr{L}}$ до подпространства $\mathscr{M}$ и расстояние $d(y, \mathscr{L})$ от вектора $y\in S_{\mathscr{M}}$ до подпространства $\mathscr{L}$. В рассматриваемом конечномерном случае символы inf и sup можно заменить соответственно на min и max.

Хотя определение раствора подпространств имеет смысл, как бы ни была определена норма в C^n, мы ограничимся случаем евклидовой метрики.

Теорема 8.3. *Если $P_{\mathscr{L}}$ и $P_{\mathscr{M}}$ — ортопроекторы на подпространства $\mathscr{L}$ и $\mathscr{M}$, то*

$$\delta(\mathscr{L}, \mathscr{M})=\|P_{\mathscr{L}}-P_{\mathscr{M}}\|_2. \tag{8.4}$$

Доказательство. Если x — произвольный вектор из $\mathscr{L}$, то

$$\|(P_{\mathscr{L}}-P_{\mathscr{M}})x\|_2=\|P_{\mathscr{L}}x-P_{\mathscr{M}}x\|_2=\|x-P_{\mathscr{M}}x\|_2=d(x, \mathscr{M}),$$

поэтому

$$\|P_{\mathscr{L}}-P_{\mathscr{M}}\|_2=\max_{x\in\mathbf{C}^n,\ \|x\|_2=1}\|(P_{\mathscr{L}}-P_{\mathscr{M}})x\|_2\geqslant$$

$$\geqslant\max_{x\in S_{\mathscr{L}}}\|(P_{\mathscr{L}}-P_{\mathscr{M}})x\|_2=\max_{x\in S_{\mathscr{L}}}d(x, \mathscr{M}).$$

Точно так же

$$\|P_{\mathscr{L}}-P_{\mathscr{M}}\|_2\geqslant\max_{y\in S_{\mathscr{M}}}d(y, \mathscr{L}).$$

Следовательно,

$$\|P_{\mathscr{L}}-P_{\mathscr{M}}\|_2\geqslant\delta(\mathscr{L}, \mathscr{M}). \tag{8.5}$$

Докажем теперь обратное неравенство. Для произвольного $x \in \mathbf{C}^n$ имеем

$$(P_{\mathscr{L}} - P_{\mathscr{M}})x = P_{\mathscr{L}}x - P_{\mathscr{M}}x = (P_{\mathscr{L}}x - P_{\mathscr{L}}P_{\mathscr{M}}x) - (P_{\mathscr{M}}x - P_{\mathscr{L}}P_{\mathscr{M}}x) = \\ = P_{\mathscr{L}}(I - P_{\mathscr{M}})x - (I - P_{\mathscr{L}})P_{\mathscr{M}}x.$$

Полагая $u = P_{\mathscr{L}}(I - P_{\mathscr{M}})x$, $v = -(I - P_{\mathscr{L}})P_{\mathscr{M}}x$, заключаем, что $u \in \mathscr{L}$, $v \in \mathscr{L}^{\perp}$, а потому

$$\|(P_{\mathscr{L}} - P_{\mathscr{M}})x\|_2^2 = \|u\|_2^2 + \|v\|_2^2. \tag{8.6}$$

Оценим сверху каждое из слагаемых правой части. Если $x \notin \mathscr{M}^{\perp}$ и $t = P_{\mathscr{M}}x / \|P_{\mathscr{M}}x\|_2$, то $t \in \mathscr{M}$ и

$$\|v\|_2 = \|(I - P_{\mathscr{L}})P_{\mathscr{M}}x\|_2 = \|(I - P_{\mathscr{L}})t\|_2 \|P_{\mathscr{M}}x\|_2 = \\ = d(t, \mathscr{L}) \|P_{\mathscr{M}}x\|_2 \leqslant \max_{y \in S_{\mathscr{M}}} d(y, \mathscr{L}) \|P_{\mathscr{M}}x\|_2 \leqslant \delta(\mathscr{L}, \mathscr{M}) \|P_{\mathscr{M}}x\|_2. \tag{8.7}$$

Если $x \perp \mathscr{M}$, то $P_{\mathscr{M}}x = 0 = v$, и неравенство $\|v\|_2 \leqslant \delta(\mathscr{L}, \mathscr{M}) \|P_{\mathscr{M}}x\|_2$ по-прежнему выполнено. Далее,

$$\|u\|_2 = \|P_{\mathscr{L}}(I - P_{\mathscr{M}})x\|_2 = \|P_{\mathscr{L}}(I - P_{\mathscr{M}})(I - P_{\mathscr{M}})x\|_2 \leqslant \\ \leqslant \|P_{\mathscr{L}}(I - P_{\mathscr{M}})\|_2 \|(I - P_{\mathscr{M}})x\|_2,$$

$$\|P_{\mathscr{L}}(I - P_{\mathscr{M}})\|_2 = \|(I - P_{\mathscr{M}})^* P_{\mathscr{L}}^*\|_2 = \|(I - P_{\mathscr{M}})P_{\mathscr{L}}\|_2 = \\ = \max_{\|z\|_2 = 1} \|(I - P_{\mathscr{M}})P_{\mathscr{L}}z\|_2 = \max_{\substack{w \in \mathscr{L} \\ \|w\|_2 \leqslant 1}} \|(I - P_{\mathscr{M}})w\|_2 = \\ = \max_{w \in S_{\mathscr{L}}} \|w - P_{\mathscr{M}}w\|_2 = \max_{w \in S_{\mathscr{L}}} d(w, \mathscr{M}) \leqslant \delta(\mathscr{L}, \mathscr{M}),$$

откуда

$$\|u\|_2 \leqslant \delta(\mathscr{L}, \mathscr{M}) \|(I - P_{\mathscr{M}})x\|_2. \tag{8.8}$$

Подставляя в (8.6) оценки (8.7) и (8.8), получаем

$$\|(P_{\mathscr{L}} - P_{\mathscr{M}})x\|_2^2 \leqslant \delta^2(\mathscr{L}, \mathscr{M}) [\|P_{\mathscr{M}}x\|_2^2 + \|(I - P_{\mathscr{M}})x\|_2^2] = \\ = \delta^2(\mathscr{L}, \mathscr{M}) \|x\|_2^2.$$

Следовательно,

$$\|P_{\mathscr{L}} - P_{\mathscr{M}}\|_2 = \max_{\|x\|_2 = 1} \|(P_{\mathscr{L}} - P_{\mathscr{M}})x\|_2 \leqslant \delta(\mathscr{L}, \mathscr{M}).$$

Вместе с неравенством (8.5) это доказывает теорему.

Следствие 8.4. *Раствор ортогональных дополнений $\mathscr{L}^{\perp}$ и $\mathscr{M}^{\perp}$ равен раствору самих подпространств $\mathscr{L}$ и $\mathscr{M}$:*

$$\delta(\mathscr{L}^{\perp}, \mathscr{M}^{\perp}) = \|P_{\mathscr{L}^{\perp}} - P_{\mathscr{M}^{\perp}}\|_2 = \|(I - P_{\mathscr{L}}) - (I - P_{\mathscr{M}})\|_2 = \\ = \|P_{\mathscr{M}} - P_{\mathscr{L}}\|_2 = \delta(\mathscr{L}, \mathscr{M}).$$

Следствие 8.5. *Раствор есть метрика на множестве подпространств из $\mathbf{C}^n$.*

Действительно, положительность и симметрия раствора очевидны из (8.3), а неравенство треугольника легко следует из (8.4).

Замечание 8.6. В качестве метрики раствор полезен только при сопоставлении подпространств одинаковой размерности. Пусть, например, $\dim \mathscr{L} > \dim \mathscr{M}$. Тогда в $\mathscr{L}$ найдется нормированный вектор x_0, ортогональный к $\mathscr{M}$. Поэтому $\delta(\mathscr{L}, \mathscr{M}) = 1$. Это соотношение справедливо для любых двух подпространств разной размерности.

Теорема 8.7 (см. [89]). *Пусть заданы подпространства $\mathscr{L}$ и $\mathscr{M}$, причем $\dim \mathscr{L} = \dim \mathscr{M} = l \leqslant n/2$. Найдутся унитарные матрицы $X = [X_1 | X_2]$ и $Y = [Y_1 | Y_2]$ такие, что $n \times l$-матрицы X_1, Y_1 базисные в $\mathscr{L}$ и $\mathscr{M}$ и*

$$Y^* X = \begin{bmatrix} \Gamma & \Sigma & 0 \\ -\Sigma & \Gamma & 0 \\ 0 & 0 & I_{n-2l} \end{bmatrix}. \tag{8.9}$$

При этом $l \times l$-матрицы Γ, Σ диагональные и неотрицательные:

$$\Gamma = \operatorname{diag}(\gamma_1, ..., \gamma_l), \quad \Sigma = \operatorname{diag}(\sigma_1, ..., \sigma_l), \quad \gamma_i \geqslant 0, \quad \sigma_j \geqslant 0 \quad \forall i, j.$$

Доказательство. Выберем в $\mathscr{L}$ произвольный ортонормированный базис $\hat{x}_1, ..., \hat{x}_l$, положим $\hat{X}_1 = [\hat{x}_1 | ... | \hat{x}_l]$ и достроим (опять по произволу) $\hat{X}_1$ до унитарной матрицы $\hat{X} = [\hat{X}_1 | \hat{X}_2]$. Аналогичным образом, исходя из ортонормированного базиса $\hat{y}_1, ..., \hat{y}_l$ подпространства $\mathscr{M}$, построим унитарную матрицу $\hat{Y} = [\hat{Y}_1 | \hat{Y}_2]$. Матрицу $\hat{Y}^* \hat{X}$ представим в блочном виде:

$$\hat{Y}^* \hat{X} = \begin{bmatrix} \hat{Y}_1^* \hat{X}_1 & \hat{Y}_1^* \hat{X}_2 \\ \hat{Y}_2^* \hat{X}_1 & \hat{Y}_2^* \hat{X}_2 \end{bmatrix} \equiv \begin{bmatrix} \hat{Z}_{11} & \hat{Z}_{12} \\ \hat{Z}_{21} & \hat{Z}_{22} \end{bmatrix}.$$

Заметим, что если в каждом из подпространств $\mathscr{L}$, $\mathscr{L}^\perp$, $\mathscr{M}$, $\mathscr{M}^\perp$ перейти к новому ортонормированному базису, то, повторяя наши построения, мы приходим к унитарным матрицам $\tilde{X} = [\tilde{X}_1 | \tilde{X}_2] = [\hat{X}_1 R_1 | \hat{X}_2 R_2]$, $\tilde{Y} = [\tilde{Y}_1 | \tilde{Y}_2] = [\hat{Y}_1 S_1 | \hat{Y}_2 S_2]$ и

$$\tilde{Y}^* \tilde{X} = \begin{bmatrix} \tilde{Y}_1^* \tilde{X}_1 & \tilde{Y}_1^* \tilde{X}_2 \\ \tilde{Y}_2^* \tilde{X}_1 & \tilde{Y}_2^* \tilde{X}_2 \end{bmatrix} = \begin{bmatrix} S_1^* \hat{Z}_{11} R_1 & S_1^* \hat{Z}_{12} R_2 \\ S_2^* \hat{Z}_{21} R_1 & S_2^* \hat{Z}_{22} R_2 \end{bmatrix}. \tag{8.10}$$

Здесь R_1, R_2, S_1, S_2 — матрицы перехода от старых базисов подпространств к их новым базисам.

Пусть $\hat{Z}_{11} = U \Gamma V^*$ — сингулярное разложение матрицы $\hat{Z}_{11}$. Для определенности условимся, что диагональные элементы матрицы Γ упорядочены по возрастанию. Положим $S_1 = U$, $R_1 = V$; тогда $\tilde{Z}_{11} = S_1^* \hat{Z}_{11} R_1 = \Gamma$. Из унитарности матрицы $\tilde{Y}^* \tilde{X}$ вытекают соотношения

$$\Gamma^* \Gamma + \tilde{Z}_{21}^* \tilde{Z}_{21} = I_l, \qquad \Gamma \Gamma^* + \tilde{Z}_{12} \tilde{Z}_{12}^* = I_l. \tag{8.11}$$

Здесь $\tilde{Z}_{21} = S_2^* \hat{Z}_{21} V$, $\tilde{Z}_{12} = U^* \hat{Z}_{12} R_2$. Первое из этих равенств означает, что столбцы матрицы $\tilde{Z}_{21}$ попарно ортогональны, второе говорит о том, что попарно ортогональны строки матрицы $\tilde{Z}_{12}$. Эта ортогональность имеет место независимо от выбора унитарных матриц S_2 (в первом случае) и R_2 (во втором).

Рассмотрим вначале более простой случай, когда среди $\gamma_1, ..., \gamma_l$ нет чисел, равных 1. Согласно (8.11), это предположение гарантирует, что матрица $\tilde{Z}_{21}$ имеет полный столбцовый ранг, а матрица $\tilde{Z}_{12}$ — полный строчный. Возьмем теперь в качестве S_2^* произведение левосторонних отражений или вращений, которые приводят матрицу $\hat{Z}_{21}V$ к «треугольному» виду. Поскольку попарная ортогональность столбцов в ходе приведения сохраняется, «треугольный» вид будет в действительности «диагональным». При этом всегда можно обеспечить отрицательность диагональных элементов. В результате матрица $\tilde{Z}_{21}$ приобретет вид

$$\tilde{Z}_{21}=\begin{bmatrix}\Phi\\0\end{bmatrix}\begin{matrix}\}l\\\}n-2l\end{matrix}, \tag{8.12}$$

где $\Phi=\operatorname{diag}(\varphi_1, ..., \varphi_l)$, $\varphi_i<0\ \forall i$. Матрицу R_2 подбираем аналогичным образом, с тем чтобы обеспечить представление

$$\tilde{Z}_{12}=[\underbrace{\Sigma}_{l}\quad\underbrace{0}_{n-2l}], \tag{8.13}$$

$\Sigma=\operatorname{diag}(\sigma_1, ..., \sigma_l)$, $\sigma_i>0\ \forall i$. Подставляя матрицы (8.12) и (8.13) в соотношения (8.11), находим

$$\Phi^2=I_l-\Gamma^2=\Sigma^2, \tag{8.14}$$

откуда $\Phi=-\Sigma$. Теперь можно представить матрицу (8.10) в более дробной форме:

$$\tilde{Y}^*\tilde{X}=\begin{bmatrix}\Gamma & \Sigma & 0\\ -\Sigma & W & F\\ 0 & G & H\end{bmatrix}\begin{matrix}\}l\\\}l\\\}n-2l\end{matrix}$$
$$\underbrace{}_{l}\ \underbrace{}_{l}\ \underbrace{}_{n-2l}$$

Из ортогональности первых двух блочных строк следует, что $W=\Gamma$, а тогда из условий

$$\Sigma^2+WW^*+FF^*=I_l,\qquad \Sigma^2+W^*W+G^*G=I_l$$

и равенств (8.14) выводим следующее: $F=0$, $G=0$. Поэтому H должна быть унитарной матрицей порядка $n-2l$. Умножая последние $n-2l$ столбцов матрицы $\tilde{Y}$ на матрицу H (или последние $n-2l$ столбцов матрицы $\tilde{X}$ на матрицу H^*), мы и получим нужные матрицы X, Y и представление (8.9).

Пусть теперь последние $l-m$ чисел γ_i равны единице. Тогда матрицу (8.10) можно представить таким образом:

$$\tilde{Y}^*\tilde{X}=\begin{bmatrix}\Gamma_m & 0 & \tilde{Z}_{12}^{(1)} & \tilde{Z}_{12}^{(2)}\\ 0 & I_{l-m} & 0 & 0\\ \tilde{Z}_{21}^{(1)} & 0 & \tilde{Z}_{22}^{(1)} & \tilde{Z}_{22}^{(2)}\\ \tilde{Z}_{21}^{(3)} & 0 & \tilde{Z}_{22}^{(3)} & \tilde{Z}_{22}^{(4)}\end{bmatrix}. \tag{8.15}$$

В этом представлении учтено, что строки и столбцы унитарной матрицы $\tilde{Y}^*\tilde{X}$ нормированы. Поскольку $\gamma_1\leqslant\gamma_2\leqslant...\leqslant\gamma_m<1$, то матрицы

$[\tilde{Z}_{12}^{(1)}|\tilde{Z}_{12}^{(2)}]$ и $[(\tilde{Z}_{21}^{(1)})^{\tau}|(\tilde{Z}_{21}^{(3)})^{\tau}]$ имеют полный строчный ранг, что позволяет выбрать R_2 и S_2 по аналогии с предыдущим случаем. Матрица (8.15) превращается в

$$\begin{bmatrix} \Gamma_m & 0 & \Sigma_m & 0 & 0 \\ 0 & I_{l-m} & 0 & 0 & 0 \\ -\Sigma_m & 0 & W_{11} & W_{12} & F_1 \\ 0 & 0 & W_{21} & W_{22} & F_2 \\ 0 & 0 & G_1 & G_2 & H \end{bmatrix}, \quad \Sigma_m = \operatorname{diag}(\sigma_1, \ldots, \sigma_m), \ \sigma_i > 0, \ i = 1, \ldots, m.$$

Как и выше, показываем, что $W_{11}=\Gamma_m$, $W_{12}=0$, $F_1=0$, $W_{21}=0$, $G_1=0$. Матрица порядка $n_1=n-l-m$

$$\begin{bmatrix} W_{22} & F_2 \\ G_2 & H \end{bmatrix}$$

унитарная, и остается повторить заключительную часть прежнего рассуждения, заменяя в ней число $n-2l$ на n_1.

Каноническими углами между подпространствами $\mathscr{L}$ и $\mathscr{M}$ называются углы θ_i ($i=1, \ldots, l$), лежащие в пределах от 0 до $\pi/2$ и определяемые соотношениями

$$\cos\theta_i = \gamma_i, \tag{8.16}$$

где числа $\gamma_1, \ldots, \gamma_l$ образуют матрицу Γ в (8.9). Если положить $\Theta=\operatorname{diag}(\theta_1, \ldots, \theta_l)$, то, согласно (8.16), $\Gamma=\cos\Theta$, $\Sigma=\sin\Theta$.

Замечание 8.8. Из доказательства теоремы 8.7 видно, что γ_i ($i=1, \ldots, l$) суть сингулярные числа матрицы $\hat{Y}_1^* \hat{X}_1$, соответствующей выбору в $\mathscr{L}$ и $\mathscr{M}$ ортонормированных базисов $\hat{x}_1, \ldots, \hat{x}_l$ и $\hat{y}_1, \ldots, \hat{y}_l$. В действительности те же числа получатся, если взять *любые* ортонормированные базисы этих подпространств. В самом деле, новому выбору отвечают матрицы $\check{X}_1=\hat{X}_1 R$ и $\check{Y}_1=\hat{Y}_1 S$, где R, S—унитарные $l\times l$-матрицы. Но сингулярные числа матриц $\hat{Y}_1^* \hat{X}_1$ и $\check{Y}_1^* \check{X}_1 = S^* \hat{Y}_1^* \hat{X}_1 R$ одинаковы.

Замечание 8.9. Условие теоремы 8.7 $l \leqslant n/2$ не является ограничением, так как при $l>n/2$ мы можем вместо $\mathscr{L}$ и $\mathscr{M}$ рассматривать их ортогональные дополнения и искать канонические углы между $\mathscr{L}^{\perp}$ и $\mathscr{M}^{\perp}$.

Пример 8.10. Пусть x и y—нормированные векторы. Применяя к натянутым на них одномерным подпространствам $\mathscr{L}$ и $\mathscr{M}$ теорему 8.7 при $l=1$, получаем

$$Y^*X = \begin{bmatrix} \gamma & \sigma & \\ -\sigma & \gamma & 0 \\ 0 & & I_{n-2} \end{bmatrix}.$$

Здесь $\gamma=|y^*x|=|(x, y)|$, и канонический угол θ из формулы (8.16) совпадает с обычным углом φ между векторами x и y, если этот последний угол острый; в противном случае углы θ и φ дополняют друг друга до π.

Установим теперь связь между двумя введенными мерами близости подпространств.

Теорема 8.11 (см. [185]). *Пусть в условиях теоремы* 8.7 $P_{\mathscr{L}}$ *и* $P_{\mathscr{M}}$ *— ортопроекторы на подпространства* $\mathscr{L}$ *и* $\mathscr{M}$. *В таком случае спектр оператора* $P_{\mathscr{L}}-P_{\mathscr{M}}$ *составляют числа* $\pm\sin\theta_i$ $(i=1, \ldots, l)$, *дополненные* $n-2l$ *нулями.*

Доказательство. Выбирая базисные матрицы подпространств, построенные в теореме 8.7, мы можем записать ортопроекторы в виде $P_{\mathscr{L}}=X_1X_1^*$, $P_{\mathscr{M}}=Y_1Y_1^*$. Тогда

$$P_{\mathscr{L}}-P_{\mathscr{M}}=X_1X_1^*-Y_1Y_1^*=X(X^*X_1X_1^*X-X^*Y_1Y_1^*X)X^*=$$

$$=X\left(\begin{bmatrix} I_l \\ 0 \\ 0 \end{bmatrix}\begin{bmatrix} I_l & 0 & 0 \end{bmatrix}-\begin{bmatrix} \Gamma \\ \Sigma \\ 0 \end{bmatrix}\begin{bmatrix} \Gamma & \Sigma & 0 \end{bmatrix}\right)X^*=$$

$$=X\begin{bmatrix} I_l-\Gamma^2 & -\Gamma\Sigma & 0 \\ -\Sigma\Gamma & -\Sigma^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}X^*.$$

Поскольку матрицы Γ и Σ диагональные, то в спектре матрицы $P_{\mathscr{L}}-P_{\mathscr{M}}$ содержатся спектры матриц

$$\Psi_i=\begin{bmatrix} 1-\gamma_i^2 & -\gamma_i\sigma_i \\ -\gamma_i\sigma_i & -\sigma_i^2 \end{bmatrix}, \qquad i=1, \ldots, l.$$

Так как $1-\gamma_i^2=\sigma_i^2$, то собственными значениями матрицы Ψ_i будут числа σ_i и $-\sigma_i$, т. е. $\sin\theta_i$ и $-\sin\theta_i$.

Следствие 8.12. *Если канонические углы между подпространствами* $\mathscr{L}$ *и* $\mathscr{M}$ *упорядочены по убыванию*: $\theta_1\geqslant\theta_2\geqslant\ldots\geqslant\theta_l$, *то*

$$\begin{aligned}&\delta(\mathscr{L}, \mathscr{M})=\|P_{\mathscr{L}}-P_{\mathscr{M}}\|_2=\sin\theta_1,\\ &\|P_{\mathscr{L}}-P_{\mathscr{M}}\|_E=(2\sin^2\theta_1+\ldots+2\sin^2\theta_l)^{1/2}.\end{aligned} \tag{8.17}$$

Действительно, сингулярные числа эрмитовой матрицы $P_{\mathscr{L}}-P_{\mathscr{M}}$ совпадают с модулями ее собственных значений.

Наконец, рассмотрим еще один способ характеризовать близость подпространств. Пусть $X=[X_1|X_2]$ — унитарная матрица, причем X_1 — базисная матрица подпространства $\mathscr{L}$ размерности l. Базисную матрицу Y_1 подпространства $\mathscr{M}$ будем искать в виде

$$Y_1=(X_1+X_2P)(I_l+P^*P)^{-1/2}, \tag{8.18}$$

где P — некоторая $(n-l)\times l$-матрица. Ее норма и будет характеристикой близости $\mathscr{L}$ и $\mathscr{M}$. При $P=0$ матрицы X_1 и Y_1 совпадают.

И эта новая характеристика оказывается связанной с каноническими углами между подпространствами $\mathscr{L}$, $\mathscr{M}$. Канонические углы определяются сингулярными числами γ_i произведения $Y_1^*X_1$ (см. замечание 8.8), т. е. в нашем случае сингулярными числами матрицы $(I_l+P^*P)^{-1/2}$. Если $\tau_1, \ldots, \tau_l$ — сингулярные числа матрицы P, то

$$\cos\theta_i=\gamma_i=(1+\tau_i^2)^{-1/2}, \qquad i=1, \ldots, l,$$

т. е.

$$\tau_i=\operatorname{tg}\theta_i, \qquad i=1, \ldots, l.$$

В результате приходим к неравенству

$$\delta(\mathscr{L}, \mathscr{M}) = \sin\theta_1 \leqslant \operatorname{tg}\theta_1 = \|P\|_2.$$

Перейдем теперь к вопросу о том, как оценить качество $n \times l$-матрицы X_1 — базисной матрицы приближенного инвариантного подпространства матрицы A. Вспомним, что качество приближенного собственного вектора x часто характеризуют посредством нормы соответствующей невязки

$$r(\mu, x) = Ax - \mu x,$$

причем оптимальным выбором числа μ — при фиксированном векторе x — является здесь отношение Рэлея

$$\rho(x) = (Ax, x)/(x, x).$$

Для вектора x единичной евклидовой длины отношение Рэлея перестает быть отношением: $\rho(x) = (Ax, x)$. Подобно этому, вводя понятие *матрицы-невязки*

$$R(M, X_1) \equiv AX_1 - X_1 M, \tag{8.19}$$

где M — произвольная $l \times l$-матрица, можно оценивать качество матрицы X_1 посредством спектральной или евклидовой нормы $\|R(M, X_1)\|$ и стремиться выбрать M наилучшим образом с точки зрения нормы невязки.

Теорема 8.13 (см. [131]). *Пусть X_1 есть $n \times l$-матрица с ортонормированными столбцами. Минимум нормы (спектральной или евклидовой) матрицы-невязки* (8.19) *достигается тогда, когда M есть матричное отношение Рэлея для X_1, т. е. когда*

$$M = X_1^* A X_1. \tag{8.20}$$

Доказательство. Пусть $X = [X_1 | X_2]$ — унитарная $n \times n$-матрица. Тогда

$$\|R(M, X_1)\| = \|X^* R(M, X_1)\| = \left\| \begin{bmatrix} X_1^* \\ X_2^* \end{bmatrix} AX_1 - \begin{bmatrix} X_1^* \\ X_2^* \end{bmatrix} X_1 M \right\| = \begin{bmatrix} X_1^* AX_1 - M \\ X_2^* AX_1 \end{bmatrix}.$$

Для обеих норм — спектральной и евклидовой — норма невязки не может быть меньше, чем $\|X_2^* AX_1\|$, и в обоих случаях это значение достигается, если $M = X_1^* AX_1$. Для евклидовой нормы точка минимума единственна.

Наша ближайшая цель заключается в следующем. Имея матрицу X_1 — базисную матрицу *приближенного* инвариантного подпространства матрицы A, мы хотели бы оценить величину матрицы P в соотношении (8.18), где Y_1 определяет *точное* инвариантное подпространство. Оценка должна выражаться через норму невязки, порождаемой матрицей X_1, и некоторые другие величины, зависящие от A и X_1. Какие именно величины, будет видно из приводимого ниже эскизного описания метода, с помощью которого мы вслед за [185] найдем нужную оценку.

Достроим заданную матрицу X_1 (считая ее столбцы ортонормированными) до унитарной матрицы $X=[X_1|X_2]$. Представим матрицу X^*AX в блочном виде:

$$B=X^*AX=\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \tag{8.21}$$

где

$$B_{ij}=X_i^*AX_j, \qquad i,\ j=1,\ 2. \tag{8.22}$$

Таким образом, B_{11} есть матричное отношение Рэлея для X_1, а величина $\|B_{21}\|$, согласно теореме 8.13, дает норму соответствующей невязки. Будем искать унитарную матрицу

$$Y=XU=X\begin{bmatrix} I_l & -P^* \\ P & I_{n-l} \end{bmatrix}\begin{bmatrix} (I_l+P^*P)^{-1/2} & 0 \\ 0 & (I_{n-l}+PP^*)^{-1/2} \end{bmatrix} \tag{8.23}$$

с таким расчетом, чтобы матрица $C=Y^*AY$ имела *точную* блочно треугольную форму

$$C=Y^*AY=\begin{bmatrix} C_{11} & C_{12} \\ 0 & C_{22} \end{bmatrix}. \tag{8.24}$$

Первые l столбцов матрицы Y как раз и дадут Y_1 из формулы (8.18).

Если в условие $C_{21}=Y_2^*AY_1=0$ подставить выражение (8.23) и использовать (8.21), то придем к уравнению для P:

$$PB_{11}-B_{22}P=B_{21}-PB_{12}P. \tag{8.25}$$

В связи с этим уравнением определим на пространстве матриц с размерами $(n-l)\times l$ линейный матричный *оператор Сильвестра*

$$TP=PB_{11}-B_{22}P \tag{8.26}$$

и квадратичный оператор

$$\varphi P=PB_{12}P. \tag{8.27}$$

Предполагая «невязку» B_{21} малой, мы интересуемся «малым» решением уравнения (8.25). Для его разыскания естественно применить метод последовательных приближений

$$TP_{i+1}=B_{21}-\varphi P_i, \qquad i=0,\ 1,\ 2,\ \ldots \tag{8.28}$$

В качестве начального приближения можно взять значение P_0, определяемое уравнением

$$TP_0=B_{21}. \tag{8.29}$$

Для успешности такого подхода нужно, чтобы, во-первых, оператор T был не вырожден и, во-вторых, значение нормы обратного оператора T^{-1} позволяло применить к процессу (8.28) принцип сжимающих отображений.

Хорошо известно (см., например, [12, гл. VIII, § 1]), что невырожденность оператора T равносильна тому, что матрицы B_{11} и B_{22} не имеют общих собственных значений. Что касается $\|T^{-1}\|$, то величина этой нормы зависит от разделенности спектров и степени

анормальности матриц B_{11}, B_{22}. Введем в связи с этим такое определение. *Разделенностью матриц* B_{11} и B_{22} назовем число

$$\operatorname{sep}(B_{11}, B_{22}) = \min_{X \neq 0} \frac{\|XB_{11} - B_{22}X\|}{\|X\|}. \tag{8.30}$$

Норму в этой формуле можно брать евклидову или спектральную. Понятно, что в случае невырожденного оператора T справедливо равенство

$$\operatorname{sep}(B_{11}, B_{22}) = \|T^{-1}\|^{-1}, \tag{8.31}$$

для вырожденного оператора $\operatorname{sep}(B_{11}, B_{22}) = 0$.

Пример 8.14 (см. [116]). Пусть

$$B_{11} = \begin{bmatrix} 3 & 10 \\ 0 & 1 \end{bmatrix}, \quad B_{22} = \begin{bmatrix} 0 & -20 \\ 0 & 3.01 \end{bmatrix}.$$

Оценим разделенность этих матриц.

Известно [19, § 11], что в естественном базисе пространства 2×2-матриц

$$E_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad E_{12} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad E_{21} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad E_{22} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

оператору Сильвестра соответствует матрица $T_{(E)} = B_{11} \otimes I_2 - I_2 \otimes B_{22}^{\mathrm{T}}$, т. е. для нашего примера — матрица

$$T_{(E)} = \begin{bmatrix} 3 & 0 & 10 & 0 \\ 20 & -0.01 & 0 & 10 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 20 & -2.01 \end{bmatrix}.$$

Обратный оператор T^{-1} имеет матрицу (значения элементов приводятся с четырьмя десятичными знаками)

$$T_{(E)}^{-1} = \begin{bmatrix} 0.3333 & 0 & -3.333 & 0 \\ 666.6 & -100 & -5671 & -497.5 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0.9950 & -0.4975 \end{bmatrix}.$$

Если для вычисления разделенности использовать евклидову норму, то $\|T^{-1}\|_E < 5800$ и (см. (8.31))

$$\operatorname{sep}_E(B_{11}, B_{22}) > 0.00017. \tag{8.32}$$

Переходя к оператору $\varphi X = XB_{12}X$, отметим такие его свойства:

а) $\|\varphi X\| \leqslant \|B_{12}\| \|X\|^2$;

б) $\|\varphi X - \varphi Y\| \leqslant 2\|B_{12}\| \max\{\|X\|, \|Y\|\} \|X - Y\|$.

В этих неравенствах можно брать как спектральную, так и евклидову норму.

Свойство б) проверяется посредством следующей выкладки:

$$\|\varphi X - \varphi Y\| = \|XB_{12}X - YB_{12}Y\| \leqslant \|XB_{12}X - XB_{12}Y\| + \\ + \|XB_{12}Y - YB_{12}Y\| \leqslant \|X\| \|B_{12}\| \|X - Y\| + \|X - Y\| \|B_{12}\| \|Y\|.$$

Вернемся к матрице (8.21).

Теорема 8.15 (см. [185]). *Пусть оператор T из формулы* (8.26) *не вырожден. Положим* $\gamma=\|B_{21}\|$, $\eta=\|B_{12}\|$, $\delta=\|T^{-1}\|^{-1}$, $\varkappa=\gamma\eta/\delta^2$, *и пусть* $\varkappa<1/4$. *Определим* τ $(0<\tau<1)$ *как меньший из двух корней квадратного уравнения.*

$$z=\varkappa(1+z)^2. \tag{8.33}$$

Тогда найдется $(n-l)\times l$*-матрица* P *такая, что*

$$\|P\|\leqslant\frac{\gamma}{\delta}(1+\tau)<2\frac{\gamma}{\delta} \tag{8.34}$$

и матрица Y_1 *в формуле* (8.18) *базисная для инвариантного подпространства матрицы* A. *Норму во всех соотношениях нужно брать спектральную либо евклидову.*

Доказательство. Пусть последовательность матриц $\{P_i\}$ построена в соответствии с предписанием (8.28), (8.29). Покажем, что все P_i принадлежат шару

$$\mathscr{B}=\{X \mid \|X\|\leqslant\frac{\gamma}{\delta}(1+\tau)\}.$$

Так как $\|P_0\|=\|T^{-1}B_{21}\|\leqslant\gamma/\delta$, то при $i=0$ утверждение выполнено. Для прочих i имеем

$$\|P_i\|=\|T^{-1}B_{21}-T^{-1}\varphi P_{i-1}\|\leqslant\|T^{-1}\|(\|B_{21}\|+\|\varphi P_{i-1}\|)\leqslant$$

$$\leqslant\frac{1}{\delta}(\gamma+\eta\|P_{i-1}\|^2).$$

Если $P_{i-1}\in\mathscr{B}$, то $\|P_{i-1}\|\leqslant\gamma(1+\tau)/\delta$ и

$$\|P_i\|\leqslant\frac{1}{\delta}\left[\gamma+\eta\frac{\gamma^2}{\delta^2}(1+\tau)^2\right]=\frac{\gamma}{\delta}[1+\varkappa(1+\tau)^2]=\frac{\gamma}{\delta}(1+\tau),$$

т. е. и $P_i\in\mathscr{B}$. Итак, справедливы индуктивный переход и вместе с ним доказываемое утверждение.

Теперь проверим, что отображение θ

$$\theta X=T^{-1}(B_{21}-\varphi X),$$

связанное с процессом (8.28), является сжимающим на шаре $\mathscr{B}$. В самом деле,

$$\|\theta X-\theta Y\|=\|T^{-1}(\varphi Y-\varphi X)\|\leqslant\frac{1}{\delta}\|\varphi Y-\varphi X\|\leqslant$$

$$\leqslant\frac{2\eta}{\delta}\max\{\|X\|,\|Y\|\}\|X-Y\|\leqslant\frac{2\gamma\eta}{\delta^2}(1+\tau)\|X-Y\|<4\varkappa\|X-Y\|, \text{ если } X\neq Y.$$

Итак, применим принцип сжимающих отображений, откуда и следует теорема 8.15.

Теорема 8.15 позволит нам получить оценку возмущения инвариантного подпространства под действием малого возмущения матрицы. Пусть X_1 — *точная* (а не приближенная, как до сих пор) базисная

матрица инвариантного подпространства матрицы A. Если $X=[X_1|X_2]$, как и прежде, унитарная матрица, то в формуле (8.21) $B_{21}=0$. Наряду с A рассмотрим возмущенную матрицу $\tilde{A}=A+F$, тогда

$$\tilde{B}=X^*\tilde{A}X=B+\begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}, \quad G_{ij}=X_i^*FX_j. \tag{8.35}$$

По отношению к $\tilde{A}$ матрица X_1 будет задавать приближенное инвариантное подпространство. В соответствии с теоремой 8.15 можно оценить норму матрицы P, преобразующей X_1 по формуле (8.18) в базисную матрицу Y_1 *точного* инвариантного подпространства матрицы $\tilde{A}$. Нужно только величины, связанные с матрицей B, заменить одноименными величинами для $\tilde{B}$. Таким образом, теперь $\gamma=\|\tilde{B}_{21}\|=\|G_{21}\|$, $\eta=\|\tilde{B}_{12}\|\leqslant\|B_{12}\|+\|G_{12}\|$, $\delta=\|\tilde{T}^{-1}\|^{-1}=\operatorname{sep}(\tilde{B}_{11}, \tilde{B}_{22})=$ $=\operatorname{sep}(B_{11}+G_{11}, B_{22}+G_{22})$. Заметим, что из определения разделенности (8.30) тривиально вытекает

Теорема 8.16 (см. [185]). *Справедливы неравенства*

$$\operatorname{sep}(B_{11}, B_{22})+\|G_{11}\|+\|G_{22}\|\geqslant\operatorname{sep}(\tilde{B}_{11}, \tilde{B}_{22})\geqslant$$

$$\geqslant\operatorname{sep}(B_{11}, B_{22})-\|G_{11}\|-\|G_{22}\|.$$

Объединяя все сказанное, получаем главный результат этого параграфа.

Теорема 8.17 (см. [185]). *Пусть X_1 — базисная матрица инвариантного подпространства матрицы A, так что для унитарной матрицы $X=[X_1|X_2]$*

$$B=X^*AX=\begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}. \tag{8.36}$$

Пусть $\tilde{A}=A+F$ — возмущенная матрица и (8.35) — блочное представление матрицы $\tilde{B}=X^*\tilde{A}X$, аналогичное представлению (8.36). Положим

$$\delta=\operatorname{sep}(B_{11}, B_{12})-\|G_{11}\|-\|G_{22}\|,$$

и пусть $\delta>0$. Если

$$\frac{\|G_{21}\|(\|B_{12}\|+\|G_{12}\|)}{\delta^2}<\frac{1}{4}, \tag{8.37}$$

то существует матрица P с нормой

$$\|P\|\leqslant 2\,\frac{\|G_{21}\|}{\delta}, \tag{8.38}$$

такая, что столбцы матрицы $Y_1=(X_1+X_2P)(I+P^*P)^{-1/2}$ составляют ортонормированный базис инвариантного подпространства матрицы $A+F$.

Замечание 8.18. Наиболее просто условия и утверждение теоремы 8.17 выглядят для эрмитовых матриц A и $\tilde{A}$. В этом случае B_{11}, B_{22} также эрмитовы и

$$\operatorname{sep}_2(B_{11}, B_{22})=\min_{i,j}|\lambda_i(B_{11})-\lambda_j(B_{22})|. \tag{8.39}$$

Кроме того, $B_{12}=0$ и $\|G_{12}\|=\|G_{21}\|$. Если $\|F\|=\varepsilon$, то

$$\delta \geqslant \min_{i,j} |\lambda_i(B_{11})-\lambda_j(B_{22})|-2\varepsilon.$$

Условие (8.37) принимает вид $\|G_{21}\|<\delta/2$.

Вспомним, что $\|G_{21}\|$ интерпретируется как норма невязки, отвечающей приближенной базисной матрице X_1, а $\|P\|_2$ — как тангенс наибольшего канонического угла θ_1 между инвариантными подпространствами матриц A и $A+F$. Таким образом, в эрмитовом случае неравенство (8.38) связывает $\operatorname{tg}\theta_1$ с величиной невязки и разделенностью спектров подматриц $\tilde{B}_{11}$ и $\tilde{B}_{22}$. Оно является прямым обобщением хорошо известных оценок для $l=1$ (см., например, [35, § 11.7]).

Пример 8.19 (см. [116]). Пусть $A=B$ и B_{11}, B_{22} те же, что и в примере 8.14. Положим

$$B_{12}=\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \qquad B_{21}=0.$$

Матрицу возмущения $F=G$ выберем так, чтобы $G_{11}=G_{12}=G_{22}=0$,

$$G_{21}=10^{-6}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Роль X_1 играет 4×2-матрица, образованная первыми двумя координатными векторами e_1, e_2; матрица Y_1 имеет вид (при округлении до четырех десятичных разрядов)

$$Y_1=\begin{bmatrix} -0.9999 & -0.0003 \\ 0.0003 & -0.9999 \\ -0.0005 & -0.0026 \\ 0.0000 & 0.0003 \end{bmatrix}.$$

Тангенсы канонических углов между X_1 и Y_1 укладываются в оценку (8.38):

$$\|P\|_E \leqslant 2\,\frac{\|G_{21}\|_E}{\operatorname{sep}_E(B_{11},\,B_{22})} < \frac{4\times 10^{-6}}{17\times 10^{-5}} \approx 0.024.$$

ДОПОЛНЕНИЯ К § 8.

1. Поскольку в (8.21) норма $\|B_{21}\|$ предполагается малой, собственные значения диагональных блоков B_{11} и B_{22} являются *приближениями* к собственным значениям матрицы A. Если же вычислена матрица P, о которой говорится в теореме 8.15, то можно найти *точные* собственные числа. Действительно, спектр матрицы A есть объединение спектров матриц C_{11} и C_{22} в (8.24), при этом

$$C_{11}=(I_l+P^*P)^{1/2}(B_{11}+B_{12}P)(I_l+P^*P)^{-1/2},$$

$$C_{22}=(I_{n-l}+PP^*)^{-1/2}(B_{22}-PB_{12})(I_{n-l}+PP^*)^{1/2}.$$

Матрицы C_{11} и C_{22} подобны соответственно матрицам $B_{11}+B_{12}P$ и $B_{22}-PB_{12}$. Поэтому, если нужны только собственные значения, можно ограничиться построением двух последних матриц.

Точно так же из теоремы 8.17 вытекает, что спектр матрицы $A+F$ есть объединение спектров матриц $B_{11}+G_{11}+(B_{12}+G_{12})P$ и $B_{22}+G_{22}-P(B_{12}+G_{12})$.

Если матрица P не известна, оценки ее нормы, устанавливаемые в теоремах 8.15 и 8.17, позволяют оцеиить точность приближений к собственным числам, получаемым из блоков B_{11}, B_{22}. Например, собственные значения матрицы B_{11} можно рассматривать как возмущения собственных значений матрицы $B_{11}+B_{12}P$ в теореме 8.15 или матрицы $B_{11}+G_{11}+(B_{12}+G_{12})P$ в теореме 8.17.

Недостатком этих оценок является то, что граница для матрицы P зависит от разделенности δ; зависимость от δ будет перенесена и на оценки возмущений в собствеиных числах. Между тем плохо отделенные собственные значения могут быть мало чувствительны к возмущениям матрицы; так заведомо будет для нормальных матриц и матриц, близких к нормальным.

Зависимость от δ неизбежна, если для оценивания погрешностей в собственных значениях привлекаются оценки для инвариантных подпространств, где присутствие разделенности в самом деле обязательно. Однако, что касается собственных чисел, эту зависимость можно ослабить, перенеся ее в члены второго порядка.

Пусть в условиях теоремы 8.17 спектр блока B_{11} отделен от спектра B_{22}. Наряду с инвариантным подпространством $\mathscr{L}$ матрицы A, натянутым на столбцы матрицы X_1,—*правым* инвариантным подпространством—рассмотрим *левое* инвариантное подпространство $\mathscr{M}$, отвечающее $\sigma(B_{11})$. Другими словами, $\mathscr{M}$ есть инвариантное подпространство матрицы A^* для собственных значений из $\sigma(B_{11}^*)$. Базисную матрицу левого инвариантного подпространства составим из векторов, двойственных к столбцам X_1; обозначим эту матрицу через Z_1. Тогда, как легко видеть,

$$B_{11}=Z_1^* AX_1. \tag{8.40}$$

Напоминаем, что $Z_1^* X_1=I_l$. Матричное произведение (8.40) называют наряду с (8.20) матричным отношением Рэлея.

В [185] доказана

Теорема. *Пусть* θ_1 *— наибольший канонический угол между подпространствами* $\mathscr{L}$ *и* $\mathscr{M}$. *Положим* $\delta=\operatorname{sep}(B_{11}, B_{22})-2\sec\theta_1\|F\|_2$. *Пусть*

$$2\sec\theta_1\|F\|_2<\delta.$$

Тогда для матрицы $\tilde{A}=A+F$ *найдутся правое и левое инвариантное подпространства* $\tilde{\mathscr{L}}$, $\tilde{\mathscr{M}}$ *и соответствующее матричное отношение Рэлея* $\tilde{B}_{11}$ *такие, что*

$$\|B_{11}-\tilde{B}_{11}\|\leqslant\sec\theta_1\|F\|_2\left[1+\frac{2\sec\theta_1\|F\|_2}{\delta}\right]<2\sec\theta_1\|F\|_2.$$

Членом первого порядка в этой оценке является произведение $\sec\theta_1\cdot\|F\|_2$. Это точный эквивалент оценки первого порядка $k(\lambda)\|F\|_2$ для погрешности в простом собственном значении λ. Коэффициент перекоса $k(\lambda)$, напомним, тоже есть секанс угла между правым и левым собственными векторами.

2. Наиболее простое выражение разделенность имеет для эрмитовых матриц B_{11}, B_{22} (см. (8.39)). Для матриц, не являющихся нормальными, явную формулу разделенности в терминах только собственных чисел получить нельзя.

Некоторые соотношения, связанные с разделенностью, указаны в [185]. Именно

$$\operatorname{sep}(XBX^{-1}, YCY^{-1})\geqslant\frac{\operatorname{sep}(B, C)}{\operatorname{cond}X\operatorname{cond}Y},$$

$$\operatorname{sep}(XBX^*, YCY^*)=\operatorname{sep}(B, C),$$

если матрицы X и Y унитарные. Норма может быть спектральной или евклидовой.

3. Величина σ в теореме Пауэрса (см. п. 14 дополнений к § 3) с точностью до обращения есть наихудшая из разделенностей диагональных блоков, но в определении разделенности берутся нормы $\|\cdot\|_1$ либо $\|\cdot\|_\infty$, а не спектральная или евклидова норма.

4. Пусть для матрицы A найдены *приближенные* правое инвариантное подпространство $\mathscr{L}$ с базисной матрицей P и левое инвариантное подпространство $\mathscr{M}$ с базисной матрицей Q; оба подпространства имеют одинаковую размерность l; столбцы базисных матриц считаем ортонормированными. При каких условиях существует матрица F такая, что $\mathscr{L}$ и $\mathscr{M}$ — *точные* инвариантные подпространства (правое и левое) для возмущенной матрицы $B=A+F$? Этот вопрос исследован в [132], а ответ для случая $l=1$ изложен в § 7.

Постановка вопроса предполагает, что оба подпространства соответствуют одной и той же группе собственных значений. Поэтому прежде всего необходимо, чтобы матрица Q^*P была невырожденной. Далее, качество $\mathscr{L}$ и $\mathscr{M}$ как приближенных инвариантных подпространств будем характеризовать посредством норм матриц-невязок

$$R\equiv AP-PC, \qquad S^*=Q^*A-DQ^*,$$

где C и D — некоторые $l\times l$ — матрицы, которые можно выбирать произвольно. Согласно теореме 8.13, наиболее выгоден выбор матричных отношений Рэлея $C=P^*AP$, $D=Q^*AQ$. Однако чтобы удовлетворить желаемые равенства

$$(A+F)P=PC, \qquad Q^*(A+F)=DQ^*, \tag{8.41}$$

необходимо согласование в выборе C и D. Именно разрешимость относительно F уравнений (8.41) равносильна условию

$$C=(Q^*P)^{-1}D(Q^*P). \tag{8.42}$$

Если (8.42) выполнено, то для решения F с минимальной нормой справедливо равенство

$$\|F\|_2=\max\{\|R\|_2, \|S\|_2\}.$$

Заметим, что в одномерном случае уравнение (8.42) удовлетворяется автоматически. При $l>1$ матрицы C и D обязаны быть подобными.

5. Теоремы 7.1 и 7.2 могут быть многими разными способами обобщены на ситуацию, когда вместо одного приближенного собственного вектора p имеется приближенное инвариантное подпространство, определяемое столбцами $n\times l$-матрицы P. Приведем формулировки двух таких обобщений.

Теорема (см. [35, § 11.10]). *Пусть A — эрмитова $n\times n$-матрица, T — произвольная эрмитова матрица порядка l, P — $n\times l$-матрица ранга l с нормированными (в норме $\|\cdot\|_2$) столбцами;*

$$R(T, P)=AP-PT. \tag{8.43}$$

Если $\theta_1, \ldots, \theta_l$ — собственные значения матрицы T, то найдутся l собственных значений $\lambda_{\sigma_1}, \ldots, \lambda_{\sigma_l}$ матрицы A такие, что будут выполняться неравенства

$$|\theta_i-\lambda_{\sigma_i}|\leqslant\sqrt{2}\|P^{-1}\|_2\|R(T, P)\|_2, \qquad i=1, \ldots, l.$$

Теорема (см. [141]). *Пусть A — $n\times n$-матрица, в общем случае недиагонализуемая, и H трансформирует A к жордановой форме; пусть T — произвольная матрица порядка l, а P есть $n\times l$-матрица ранга l. Если θ, s — собственная пара матрицы T, то найдется собственное значение λ матрицы A такое, что*

$$|\lambda-\theta|^r/(1+|\lambda-\theta|)^{r-1}\leqslant\operatorname{cond}_2 H\|R(T, P)s\|_2/\|Ps\|_2.$$

Здесь $r=\operatorname{ind}\lambda$.

ГЛАВА 3. СТЕПЕННОЙ МЕТОД И ОБРАТНЫЕ ИТЕРАЦИИ

Два метода, описываемых в данной главе, принадлежат к числу простейших в вычислительной линейной алгебре. Несмотря на это, метод обратных итераций имеет большое практическое значение. Кроме того, очень важна идея степенной итерации, лежащая в основе обоих методов. В большей или меньшей степени она используется более сложными алгоритмами из двух следующих глав.

§ 9. Степенной метод

Степенной метод (или прямые итерации, или алгоритм Стодолы, или итерации Мизеса — все это названия одного и того же метода, многократно переоткрывавшегося в различных приложениях) предназначен для вычисления старшего собственного значения матрицы и соответствующего собственного вектора. Очень простой по своей идее и предъявляющий минимальные требования к объему памяти, он тем не менее нечасто используется в практических вычислениях. Причиной является его небыстрая, а нередко и весьма медленная сходимость. И все же включение этого метода в учебники вычислительной линейной алгебры вполне оправданно: ведь на той же, в сущности, идее основаны многие, гораздо более эффективные спектральные алгоритмы. Степенной метод позволяет описать эту идею в наиболее чистом виде.

Предположим, что матрица A диагонализуема, и пусть $\lambda_1, \ldots, \lambda_t$ — все *различные* ее собственные значения. Будем считать их пронумерованными в порядке убывания модулей, при этом пусть модуль числа λ_1 строго больше остальных:

$$|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_t|. \tag{9.1}$$

Всякий вектор $x^0 \in \mathbf{C}^n$ можно представить в виде

$$x^0 = u_1 + u_2 + \ldots + u_t, \tag{9.2}$$

где u_i — собственный вектор, относящийся к λ_i, либо $u_i = 0$. Предположим, что выбранный для степенного метода *начальный вектор* x^0 *имеет ненулевую компоненту* u_1.

Рассмотрим последовательность векторов $\{A^k x^0\}$. Применяя A^k к обеим частям равенства (8.2), получаем

$$A^k x^0 = \lambda_1^k u_1 + \lambda_2^k u_2 + \ldots + \lambda_t^k u_t = \\ = \lambda_1^k \left[u_1 + (\lambda_2/\lambda_1)^k u_2 + \ldots + (\lambda_t/\lambda_1)^k u_t\right]. \tag{9.3}$$

Для наименьшего угла φ между векторами $A^k x^0$ и u_1 имеем

$$\cos\varphi=\frac{|(A^k x^0,\ u_1)|}{\|u_1\|_2\,\|A^k x^0\|_2}=\frac{\left|\ \|u_1\|_2^2+(\lambda_2/\lambda_1)^k(u_2,\ u_1)+\ \ldots\ +(\lambda_t/\lambda_1)^k(u_t,\ u_1)\ \right|}{\|u_1\|_2\,\|u_1+(\lambda_2/\lambda_1)^k u_2+\ \ldots\ +(\lambda_t/\lambda_1)^k u_t\|_2}.$$

В силу неравенств (9.1) $\cos\varphi\to 1$ при $k\to\infty$, т. е. векторы $A^k x^0$ *сходятся по направлению* к собственному вектору u_1, отвечающему доминирующему собственному значению λ_1. Сходимость имеет скорость геометрической прогрессии со знаменателем $|\lambda_2/\lambda_1|$. При больших k выполняется приближенное равенство $A^{k+1}x^0\approx\lambda_1 A^k x^0$; поэтому в качестве приближения к λ_1 можно взять, например, отношение одноименных компонент соседних векторов последовательности $\{A^k x^0\}$ или же число

$$\frac{(A^{k+1}x^0,\ A^k x^0)}{(A^k x^0,\ A^k x^0)}.$$

Множитель λ_1^k в правой части (9.3) показывает, что при $|\lambda_1|\neq 1$ длины векторов $A^k x^0$ либо бесконечно возрастают, либо сходятся к нулю. Поэтому при практическом проведении степенного метода эти векторы время от времени или даже на каждой итерации нормируют. В последнем случае метод имеет такую вычислительную схему:

Степенной метод (9.4)

1. Выбрать нормированный вектор x^0.
2. Для $k=1, 2, \ldots$

вычислить вектор $y^k=Ax^{k-1}$;
найти норму $\|y^k\|$;
положить $x^k=y^k/\|y^k\|$.

Ясно, что нормировка не изменяет направления и свойства сходимости векторов. Выбор нормы произволен. Для выхода из итерационной процедуры можно использовать стабилизацию разрядов в компонентах последовательных векторов x^k: когда нужная точность достигнута, итерации прекращают.

Предположение о диагонализуемости матрицы A не является обязательным для сходимости степенного метода. Подчас метод можно использовать и тогда, когда имеется несколько старших собственных чисел. Различные варианты протекания метода подробно проанализированы в книгах [45, § 53; 42, гл. IX]. Мы не станем повторять этот анализ, а ограничимся указанием одного из возможных подходов к нему.

Пусть $J=J_1\oplus J_2\oplus\ldots\oplus J_s$ — жорданова форма матрицы A, так что $A=PJP^{-1}$ для некоторой невырожденной матрицы P. Тогда

$$A^k x^0=(PJP^{-1})^k x^0=PJ^k P^{-1}x^0.$$

Полагая $P^{-1}x^0=y^0$, видим, что вместо последовательности $\{A^k x^0\}$ можно изучать поведение последовательности $J^k y^0$.

Удобно считать, что еще до применения процесса (9.4) *сама матрица* A нормирована делением на старшее собственное значение λ_1 (если старших собственных значений несколько, то на одно из них). Разумеется, это предположение нужно только для теоретического

исследования — ведь λ_1 есть именно то, что должен вычислить степенной метод!

Итак, старшим собственным значением (нормированных) матриц A и J является 1, хотя могут присутствовать и другие собственные числа с модулем 1. Пусть старшим λ отвечают жордановы клетки $J_1, \ldots, J_\omega$ с наименьшими номерами. Если разбить вектор y^0 на подвекторы $y_1^0, \ldots, y_s^0$ в соответствии с порядками жордановых клеток, то аналогичное разбиение вектора $y^k = J^k y^0$ имеет вид

$$y^k = \begin{bmatrix} y_1^k \\ \ldots \\ y_\omega^k \\ y_{\omega+1}^k \\ \ldots \\ y_s^k \end{bmatrix} = \begin{bmatrix} J_1^k y_1^0 \\ \ldots \\ J_\omega^k y_\omega^0 \\ J_{\omega+1}^k y_{\omega+1}^0 \\ \ldots \\ J_s^k y_s^0 \end{bmatrix}. \tag{9.5}$$

Рассмотрим поведение степеней жордановой клетки $J_m(\lambda)$:

$$J_m(\lambda) = \begin{bmatrix} \lambda & 1 & & & & \\ & \lambda & 1 & & & \\ & & \ddots & & & \\ & & & \lambda & 1 & \\ & & & & \lambda & \end{bmatrix}.$$

Можно проверить по индукции или с применением формулы бинома Ньютона к представлению $J_m(\lambda) = \lambda I_m + E$, где

$$E = \begin{bmatrix} 0 & 1 & & & & \\ & 0 & 1 & & & \\ & & \ddots & & & \\ & & & 0 & 1 & \\ & & & & 0 & \end{bmatrix},$$

что справедливо следующее правило для возведения жордановой клетки в натуральную степень k $(k \geqslant m)$:

$$J_m^k(\lambda) = \begin{bmatrix} \lambda^k & k\lambda^{k-1} & \frac{k(k-1)}{2}\lambda^{k-2} & \ldots & C_k^{m-1}\lambda^{k-m+1} \\ & \lambda^k & k\lambda^{k-1} & \ldots & C_k^{m-2}\lambda^{k-m+2} \\ & & \ldots\ldots\ldots & \ldots & \ldots\ldots\ldots \\ & & & \lambda^k & k\lambda^{k-1} \\ & & & & \lambda^k \end{bmatrix} \tag{9.6}$$

Из (9.6) следует, в частности, что $J_m^k(\lambda) \to 0$, если $|\lambda| < 1$.

Возвращаясь к (9.5), можно сделать такие выводы:

1. Подвекторы $y_{\omega+1}^k, \ldots, y_s^k$ с ростом k сходятся к нулевым векторам соответствующей размерности.

2. Если старшее собственное значение только одно и ему отвечают только клетки порядка 1, то подвекторы $y_1^k, \ldots, y_\omega^k$ не зависят от k, а сами векторы y^k сходятся при $k \to \infty$ к вектору

$$\begin{bmatrix} y_1^0 \\ \ldots \\ y_\omega^0 \\ 0 \\ \ldots \\ 0 \end{bmatrix}.$$

Стало быть, последовательность $\{A^k x^0\}$ сходится (если A не нормировалась, то сходится по направлению) к собственному вектору для числа 1 (для λ_1). Предположение о диагонализуемости матрицы A оказалось ненужным: важно лишь, что доминирующее собственное значение единственно и его индекс равен 1. Однако сходимость к пределу может оказаться медленнее, чем в диагонализуемом случае, из-за биномиального коэффициента C_k^{m-1} в клетке $J_m(\lambda_2)$.

3. Пусть снова имеется только одно доминирующее собственное число, но на этот раз его индекс m больше 1. Если $z^0 = (\zeta_1^0, \ldots, \zeta_m^0)^T$, то

$$z^k = J_m^k(1)\, z^0 = \begin{bmatrix} \zeta_1^0 + k\zeta_2^0 + C_k^2 \zeta_3^0 + \ldots + C_k^{m-1} \zeta_m^0 \\ \zeta_2^0 + k\zeta_3^0 + \ldots + C_k^{m-2} \zeta_m^0 \\ \ldots \\ \zeta_m^0 \end{bmatrix}.$$

При больших k максимальный модуль имеет первая компонента. Если $\zeta_m^0 \neq 0$, она растет со скоростью примерно $C_k^{m-1}\zeta_m^0$, и отношение ее модуля к модулю следующей по величине, а именно второй, компоненты ведет себя как $O\left(\frac{1}{k}\right)$. С этой очень малой скоростью — скоростью гармонической последовательности — векторы z^k сходятся к координатному вектору e_1, т. е. к собственному вектору клетки J_m. То же самое будет происходить, если $\zeta_{l+1}^0 = \ldots = \zeta_m^0 = 0$, но $\zeta_l^0 \neq 0$ и $l > 1$.

Итак, и в этом случае итерации степенного метода сходятся (по направлению) к собственному вектору для доминирующего собственного числа; однако сходимость будет очень медленной. При этом компоненты векторов $\{A^k x^0\}$ по корневым подпространствам для $\lambda \neq \lambda_1$ сходятся к нулю (имеется в виду — после нормировки векторов) со скоростью, регулируемой отношениями $|\lambda_i/\lambda_1|$ и порядками жордановых клеток для λ_i. Наиболее медленной оказывается сходимость компоненты по доминирующему корневому подпространству, которая «разворачивается» в сторону собственного вектора со скоростью всего лишь гармонической последовательности.

4. Если собственных значений со старшим модулем несколько, то в J присутствуют клетки для чисел $\lambda = e^{i\varphi}(\varphi \neq 0)$. Как следует из (9.6), одновременная сходимость по направлению последовательностей $z^k = J_m^k(1) z^0$ и $w^k = J_r^k(\lambda) w^0$ невозможна (при z^0, $w^0 \neq 0$). Таким образом, последовательность $\{A^k x^0\}$ не сходится по направлению ни к какому вектору. Тем не менее компоненты векторов $A^k x^0$ по корневым подпространствам для λ_i ($|\lambda_i| < 1$) в относительном смысле убывают со скоростью геометрических прогрессий. С ростом k векторы $A^k x^0$, не придерживаясь определенного направления, приближаются тем не менее к инвариантному подпространству матрицы A, складывающемуся из корневых подпространств для старших собственных чисел. Если размерность d этого подпространства невысока, то при большом k можно использовать последовательные итерации x^k, x^{k+1}, ..., x^{k+d} для приближенного вычисления старших собственных значений и отвечающих им собственных и корневых векторов. Однако намного эффективней в этих условиях работают методы одновременных итераций (см. § 14).

Обсудим теперь роль последнего требования в описании степенного метода. Если в разложении (9.2) компонента u_1 нулевая, но зато, например, $u_2 \neq 0$ и $|\lambda_2| > |\lambda_3|$, то, проводя прежние рассуждения, убеждаемся, что итерации степенного метода сходятся по направлению к собственному вектору u_2, ассоциированному со вторым по старшинству собственным значениям. Такое поведение может быть следствием неудачного выбора начального вектора x^0, а может быть и заранее запланированным. Действительно, предположим, что собственный вектор u_1 для доминирующего собственного значения λ_1 уже известен (например, был вычислен тем же степенным методом), и теперь нужен именно вектор u_2. Тогда как раз и необходим вектор x^0 с нулевой компонентой по доминирующему собственному подпространству. Если λ_1 — простое собственное значение и известен левый собственный вектор q_1 для него, то условие $(x^0, q_1) = 0$ обеспечивает (в силу соотношений двойственности) отсутствие компоненты u_1 в разложении (9.2).

Вернемся к ситуации неудачного выбора x^0, когда разыскивается собственный вектор для λ_1. Если даже в разложении (9.2) $u_1 = 0$, то вследствие ошибок округлений с ростом k в разложениях векторов x^k (см. (9.4)) появляется ненулевая компонента по доминирующему собственному подпространству. Вначале очень малая, она растет с большей скоростью, чем другие компоненты. Поэтому при достаточной протяженности процесса он сойдется в конечном счете к нужному собственному направлению. Однако такое поведение метода не является гарантированным. Может случиться, что нулевая компонента исходного разложения остается нулевой и в дальнейшем за счет специальных свойств матрицы A. Пример этого рода, когда A — центросимметричная матрица, приведен в [3, с. 398]. Даже если процесс развивается согласно данному выше описанию, он может прерваться выполнением условия выхода еще до того, как компонента в направлении u_1 стала сколько-нибудь ощутимой. Полученный вектор будет ошибочно воспринят пользователем как приближение к старшему собственному вектору, в то время как это — приближение к u_2.

Пример 9.1 (см. [212]). Для матрицы $A=PDP^{-1}$, где

$$D=\begin{bmatrix}1&0&0\\0&4&0\\0&0&6\end{bmatrix}, \quad P=\begin{bmatrix}1&0&1\\0&1&1\\1&1&1\end{bmatrix}$$

и, следовательно,

$$A=\begin{bmatrix}6&5&-5\\2&6&-2\\2&5&-1\end{bmatrix},$$

степенным методом вычисляется доминирующее собственное значение. Вычисления проводились в студенческом практикуме на микро-ЭВМ Apple посредством программного обеспечения, описанного в [213].

При задании начального вектора $x^0=(1, 2, 3)^T$ после 15 итераций были получены значения $\tilde{\lambda}=3.99926723$ и $x^{15}=(2.94530762\times 10^{-8}, 0.999999996, 0.999999996)^T$. Таким образом, вполне можно считать, что наибольшее собственное число матрицы A равно 4, а соответствующим собственным вектором является $(0, 1, 1)^T$. Объясняется этот результат тем, что вектор x^0 принадлежит инвариантному подпространству, натянутому на собственные векторы для $\lambda_2=4$ и $\lambda_3=1$:

$$x^0=2u_2+u_3, \quad u_2=(0, 1, 1)^T, \quad u_3=(1, 0, 1)^T.$$

Если продолжить процесс, то после 60 итераций получим $\tilde{\lambda}=5.54002294$, а после 100 итераций $\tilde{\lambda}=5.99999995$.

Приведем в заключение пример ситуации, где можно дать обоснованный рецепт выбора начального вектора. Пусть нужно найти доминирующее собственное значение λ_1 и соответствующий собственный вектор p_1 *положительной* матрицы A (т. е. матрицы, все элементы которой положительны). Хорошо известно (см., например, [12, гл. XIII, § 2]), что для положительной матрицы собственное значение с наибольшим модулем единственно. Оно простое и положительное, и все компоненты ассоциированного с ним собственного вектора имеют одинаковый знак, так что их можно считать положительными. Эти утверждения составляют содержание важной *теоремы Перрона*, вследствие чего λ_1 и p_1 называют соответственно *перроновым корнем* и *перроновым вектором* матрицы A.

Естественно, что теорема Перрона в равной мере применима к положительной матрице A^T. Поэтому перронову корню λ_1 отвечает наряду с правым собственным вектором p_1 положительный левый собственный вектор q_1. Для вектора x^0 с разложением (9.2) справедливо равенство $(x^0, q_1)=(u_1, q_1)$. Следовательно, в случае *неотрицательного* вектора x^0 скалярное произведение в левой части положительно и $u_1\neq 0$. Поэтому степенной метод, начатый с неотрицательного вектора x^0, обязательно сойдется к перронову вектору положительной матрицы.

ДОПОЛНЕНИЯ К § 9

1. Для задач высокого порядка ограничения, связанные с объемом памяти ЭВМ, иногда не оставляют альтернативы применению степенного метода, несмотря на его медленную сходимость (см., например, [144]). В этих случаях используют различные приемы ускорения [45, § 54; 42, гл. IX; 144].

2. В последние годы специалисты по дифференциальной геометрии обратили внимание на то, что многие методы вычисления собственных значений, включая степенной метод, могут быть истолкованы как итерационные процессы на однородных пространствах различных групп Ли. Установлены связи этих методов с векторными и матричными дифференциальными уравнениями Риккати, которые в минувшее десятилетие также вызвали интерес у геометров. В случае степенного метода эту связь можно описать следующим образом.

Каждому ненулевому вектору $x \in \mathbf{C}^n$ сопоставляем натянутое на этот вектор одномерное подпространство $\mathscr{X}$. Множество всех одномерных подпространств в $\mathbf{C}^n$ есть $(n-1)$-мерное комплексное аналитическое многообразие, называемое комплексным проективным пространством и обозначаемое через $\mathbf{CP}^{n-1}$. Всякий невырожденный линейный оператор - (матрица) $A: \mathbf{C}^n \to \mathbf{C}^n$ естественным образом интерпретируется как оператор на $\mathbf{CP}^{n-1}$: если $x \in \mathbf{C}^n$ и $\mathscr{X} = \operatorname{span}(x)$, то $\mathscr{A}\mathscr{X} = \operatorname{span}(Ax)$. Степенные итерации $\mathscr{X}^i = \mathscr{A}\mathscr{X}^{i-1} = \mathscr{A}^i\mathscr{X}^0$ суть последовательное действие $\mathscr{A}$ на подпространство $\mathscr{X}^0$.

Рассмотрим подмножество Γ в $\mathbf{CP}^{n-1}$, образованное подпространствами с базисными векторами вида $\begin{pmatrix} 1 \\ z \end{pmatrix}$, $z \in \mathbf{C}^{n-1}$. Если $\mathscr{X}^0 \in \Gamma$ и последовательное действие A на $\mathscr{X}^0$ не выводит из Γ, то подвекторы z_i соседних итераций связаны соотношением

$$z_{i+1} = \frac{A_{21} + A_{22} z_i}{a_{11} + A_{12} z_i}. \tag{9.7}$$

Здесь a_{11} — скаляр, а A_{12}, A_{21}, A_{22} — блоки в представлении

$$A = \begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

Формуле (9.7) можно придать вид

$$z_{i+1} - z_i = (A_{21} + A_{22} z_i - a_{11} z_i - z_i A_{12} z_i)/(a_{11} + A_{12} z_i),$$

что соответствует дискретизации векторного дифференциального уравнения Риккати

$$dz/dt = A_{21} + A_{22} z - a_{11} z - z A_{12} z. \tag{9.8}$$

При этом неподвижные точки процесса (9.7) суть положения равновесия уравнения (9.8), т. е. решения алгебраического уравнения Риккати

$$A_{21} + A_{22} z - a_{11} z - z A_{12} z = 0.$$

§ 10. Обратные итерации

Обратные итерации — это степенной метод, применяемый к обратной матрице A^{-1}. Так как собственные значения последней обратны собственным значениям матрицы A, то обратные итерации служат для вычисления младших по модулю собственных чисел. За исключением этого обстоятельства, в отношении сходимости обратных операций справедливо все сказанное в § 9 о степенном методе. Нужно заметить, правда, что *при одинаковой разделенности* старших и младших собственных значений сходимость обратных итераций может оказаться значительно более быстрой, чем в степенном методе. Предположим, например, что $|\lambda_1| = 1$, $|\lambda_2| = 0.9999$, $|\lambda_{n-1}| = 0.0002$, $|\lambda_n| = 0.0001$, где λ_{n-1} и λ_n — собственные значения с наименьшими

модулями. Оба процесса сходятся со скоростью геометрической прогрессии, но для степенного метода знаменатель прогрессии равен 0.9999, и для уменьшения ошибки в e раз потребуется почти 1000 итераций; в то же время в методе обратных итераций знаменатель прогрессии равен $|\lambda_n|/|\lambda_{n-1}|$, т. е. 0.5. С каждой новой итерацией мы получаем дополнительный верный двоичный разряд в приближенном собственном значении и компонентах приближенного собственного вектора.

Вычислительная схема метода обратных итераций может быть получена формальной заменой A на A^{-1} в схеме (9.4). Однако мы несколько видоизменим последнюю.

Метод обратных итераций: (10.1)

1. Выбрать нормированный вектор x^0.
2. Для $k=1, 2, \ldots$

вычислить нормированный вектор x^k из условия

$$Ax^k=\tau_k x^{k-1}, \tag{10.2}$$

где τ_k — скаляр.

Эта форма записи подчеркивает следующие важные положения. Во-первых, для проведения обратных итераций нет необходимости обращать матрицу A. Очередное приближение x^k определяется как решение линейной системы (10.2). На всех шагах метода системы имеют одну и ту же матрицу. Поэтому, вычислив до начала пропесса (10.1) треугольное разложение матрицы A (что требует значительно меньшей арифметической работы, чем ее обращение), мы найдем каждое x^k посредством сравнительно небольших вычислений. Во-вторых, скаляр τ_k есть, в сущности, нормирующий множитель для решения системы $Ay=x^{k-1}$, однако он внесен в саму эту систему. Так сделано потому, что в случае плохо обусловленной матрицы A вектор y может быть очень велик. Чтобы воспрепятствовать выходу компонент вычисляемого вектора за пределы разрядной сетки ЭВМ, и предусмотрена нормировка в ходе вычислений. Она основана на том обстоятельстве, что нам нужен не сам вектор y, а его направление.

Из сказанного вытекает, что в методе обратных итераций мы имеем двусмысленную ситуацию. Если младшее собственное значение матрицы A близко к нулю, то даже при не очень хорошей его отделенности от соседних собственных чисел знаменатель прогрессии $|\lambda_n|/|\lambda_{n-1}|$ будет достаточно мал, и итерации быстро сходятся. С другой стороны, матрица A в этом случае будет плохо обусловлена (относительно обращения), и решение y системы $Ay=x^{k-1}$ вычисляется с большой ошибкой. Не теряется ли вместе с этой ошибкой уточнение, которое должна давать очередная итерация?

Ответ вкратце таков (цитируем [45, с. 374]): «Направление, в котором решение y плохо определено, почти совпадает с направлением самого решения, и после нормировки неопределенность в решении исчезнет». Остановимся на этом несколько подробней.

Анализ погрешностей округлений в прямых методах решения линейных систем приводит к такому выводу: если система $Ay=x^{k-1}$ решается, например, методом Гаусса (с выбором главного элемента), то вместо y будет получен вектор $\tilde{y}$, являющийся *точным* решением *возмущенной системы*

$$(A+F)\tilde{y}=x^{k-1}, \tag{10.3}$$

причем для нормы *матрицы эквивалентного возмущения* F, как правило, выполняется оценка вида

$$\|F\| \leqslant f(n)\|A\|\beta^{-t}$$

с медленно растущей функцией $f(n)$. Иными словами (см. § 2), метод Гаусса является численно устойчивым методом решения линейных систем.

Полагая $x^k=\tilde{y}/\|\tilde{y}\|$, из (10.3) получаем

$$(A-\lambda_n I)x^k=-\lambda_n x^k-Fx^k+x^{k-1}/\|\tilde{y}\|,$$

откуда (в предположении, что нормы векторов и матриц согласованы) имеем неравенство

$$\|r(\lambda_n, x^k)\| \leqslant |\lambda_n|+\|F\|+1/\|\tilde{y}\|. \tag{10.4}$$

Следовательно, если λ_n настолько мало, что имеет тот же порядок величины, что и $\|F\|$, а $\tilde{y}$, наоборот, есть большой вектор, то x^k как приближенный собственный вектор для λ_n дает малую невязку. Это означает, как мы знаем из § 7, что λ_n и x^k составляют точную собственную пару для некоторой матрицы $\tilde{A}$, близкой к A. Близость матриц A и $\tilde{A}$ не всегда сопровождается близостью собственных векторов для общего собственного значения λ_n, но если задача вычисления собственных значений и векторов матрицы A хорошо обусловлена (не путать с плохой обусловленностью A в отношении обращения!), то такая близость есть. Значит, погрешности, содержащиеся в вычисленном векторе $\tilde{y}$, не изменяют существенно его направления, а потому вектор ошибки сам имеет примерно то же направление.

Плохая обусловленность (в смысле обращения) матрицы A чаще всего проявляется именно в том, что решения систем типа (10.3) очень велики. Получается, что это обстоятельство нам как раз на руку — ведь оно способствует уменьшению невязки (10.4)!

До сих пор речь шла только о вычислении старших (степенной метод) или младших (обратные итерации) собственных значений вместе с соответствующими собственными векторами. В 1944 г. Виландт [207] предложил использовать обратные итерации для уточнения приближения μ к *произвольному* собственному значению матрицы A. Если это приближение достаточно хорошее, то можно надеяться, что у *матрицы-сдвига* $A-\mu I$ есть собственное значение, по модулю значительно меньшее остальных. В таком случае обратные итерации с матрицей $A-\mu I$ быстро приведут к его вычислению, т. е. к поправке для μ. Заодно будет найден (или уточнен) приближенный собственный вектор.

Метод Виландта (независимо от него открытый многими вычислителями) оказался настолько успешным, что под названием «обратные итерации» сейчас обычно понимают именно его, а не тот частный случай ($\mu=0$), которым мы начали параграф. Иногда, правда, говорят об «обратных итерациях со сдвигом». Очень часто метод используют в ситуации, когда собственное значение уже вычислено с предельной возможной (при его обусловленности) точностью. Тем самым обратные итерации являются по преимуществу методом вычисления собственных векторов по ранее найденным приближенным собственным значениям.

Вычислительная схема метода Виландта получается из (10.1), если заменить A на $A-\mu I$.

Обратные итерации с постоянным сдвигом: (10.5)

1. Выбрать нормированный вектор x^0.
2. Для $k=1, 2, \ldots$

вычислить нормированный вектор x^k из условия

$$(A-\mu I)x^k=\tau_k x^{k-1}, \tag{10.6}$$

где τ_k — скаляр.

Таким образом, искать разложение теперь нужно для матрицы $A-\mu I$, а не A. Вычисление разложения часто облегчается тем, что еще до применения обратных итераций матрица преобразуется к одной из удобных специальных форм — трехдиагональной, хессенберговой и т. д. Впрочем, при решении систем (10.6), особенно для матриц очень высокого порядка, применяются и итерационные методы (см., например, [111]).

На метод Виландта переносится — с очевидными видоизменениями — анализ сходимости, проведенный выше для обратных итераций. Если λ_m — собственное значение матрицы A, ближайшее к μ, то вместо (10.4) получим оценку

$$\|r(\lambda_m, x^k)\| \leqslant |\lambda_m-\mu|+\|F\|+1/\|\tilde{y}\|.$$

Как и прежде, можем сделать отсюда вывод о том, что при хорошей обусловленности задачи вычисления собственных значений и собственных векторов матрицы A вектор x^k близок к собственному вектору для λ_m. Практика показывает, что для достижения приемлемой точности очень часто достаточно всего лишь одной итерации метода.

Если собственный вектор для λ_m обусловлен плохо, то требуется несколько итераций, и само поведение метода заметно усложняется. Как уже говорилось, обратные итерации часто применяют к вычислению собственного вектора для приближенного собственного значения, найденного с рабочей точностью. В этом случае μ в уравнении (10.6) можно приближенно отождествить с λ_m. Тогда вычисление длин невязок для последовательных приближений x^k не требует никакой дополнительной работы, ведь, согласно (10.6), $\|(A-\mu I)\times$ $\times x^k\|=\tau_k$. Казалось бы, естественно судить о характере сходимости метода по скорости убывания его невязок. Между тем замечено,

что для случаев плохой обусловленности типично такое поведение невязок: τ_1 очень мало, а последующие τ_k *гораздо больше*. Значит ли это, что метод перестает сходиться?

Анализу обратных итераций в плохо обусловленном случае посвящена важная работа [164]. Ниже мы изложим основные положения этого анализа.

Предположим, что вычисления в схеме (10.5) выполняются без ошибок округлений. Описанное выше поведение невязок не зависит от данного предположения.

Матрицу A будем считать диагонализуемой, а собственные значения — занумерованными так, что ближайшим к μ является число λ_1. Относительно плохой обусловленности предположим, что она проявляется в большом значении числа обусловленности $k(\lambda_1)$. Согласно (6.7), отсюда следует, что матрица P, составленная по столбцам из нормированных собственных векторов для A, будет плохо обусловленной. Поэтому некоторая подсистема ее столбцов, скажем столбцы $p_1, ..., p_r$, будет «почти» линейно зависимой. Последнее означает: найдутся числа $\alpha_1, ..., \alpha_r$ такие, что

$$\alpha_1 p_1 + ... + \alpha_r p_r = \varepsilon v_1, \qquad |\alpha_1|^2 + ... + |\alpha_r|^2 = 1, \tag{10.7}$$

где v_1 — нормированный вектор, а ε — малое положительное число. Включим v_1 в какой-нибудь ортонормированный базис $v_1, v_2, ..., v_n$ и разложим вектора v_i по базису из собственных векторов $p_1, p_2, ..., p_n$:

$$v_1 = \sum_{i=1}^{r} (\alpha_i/\varepsilon) p_i, \qquad v_j = \sum_{i=1}^{n} \gamma_{ji} p_i, \qquad j = 2, ..., n. \tag{10.8}$$

В то время как коэффициенты первого разложения в (10.8) велики (из-за ε в знаменателе), нет оснований считать большими коэффициенты γ_{ji} остальных разложений.

Пусть x^0 — начальный вектор обратных итераций. Выпишем и для него разложение по базису $p_1, ..., p_n$, но сделаем это в два этапа. Вначале разложим x^0 по ортонормированному базису $v_1, ..., v_n$:

$$x^0 = \sum_{j=1}^{n} \beta_j v_j. \tag{10.9}$$

Поскольку x^0 нормирован, скажем, в евклидовой норме, то

$$\sum_{j=1}^{n} |\beta_j|^2 = 1.$$

Вероятность того, что наугад взятый вектор x^0 имеет малую компоненту β_1, невелика. Поэтому будем считать, что $|\beta_1| \gg \varepsilon$. Подставляя в (10.9) равенства (10.8), получаем искомое разложение, которое запишем в виде

$$x^0 = \sum_{i=1}^{r} \left[\beta_1 \alpha_i/\varepsilon + \sum_{j=2}^{n} \beta_j \gamma_{ji} \right] p_i + \sum_{i=r+1}^{n} \left[\sum_{j=2}^{n} \beta_j \gamma_{ji} \right] p_i. \tag{10.10}$$

В полученном разложении коэффициенты при векторах $p_1, ..., p_r$ гораздо больше прочих коэффициентов. При этом их величина

определяется первыми слагаемыми $\beta_1\alpha_i/\varepsilon$ в скобках. Чтобы сумма (10.10) могла быть нормированным вектором, необходимо, чтобы при сложении происходило сильное взаимное уничтожение составляющих ее больших векторов. Это уничтожение объясняется именно тем, что коэффициенты разложения (10.10) при векторах $p_1, \ldots, p_r$ находятся в примерном отношении $\alpha_1 : \alpha_2 : \ldots : \alpha_r$ (см. (10.7)). Так будет для почти любого начального вектора x^0 — важно лишь, чтобы $|\beta_1| \gg \varepsilon$.

Теперь посмотрим, как ведут себя невязки метода обратных итераций. Чтобы упростить рассуждения, будем следить только за порядком величины невязок и не будем нормировать последовательные приближения метода. Тогда τ_k определяется отношением

$$\tau_k = \|(A-\mu I)^{-(k-1)}x^0\| / \|(A-\mu I)^{-k}x^0\|. \tag{10.11}$$

Умножая обе части равенства (10.10) на матрицу $(A-\mu I)^{-1}$, получаем

$$y^1 = (A-\mu I)^{-1}x^0 = \sum_{i=1}^{r} \frac{1}{\lambda_i-\mu}\left[\frac{\beta_1\alpha_i}{\varepsilon} + \sum_{j=2}^{n}\beta_j\gamma_{ji}\right]p_i + \\ + \sum_{i=r+1}^{n} \frac{1}{\lambda_i-\mu}\left[\sum_{j=2}^{n}\beta_j\gamma_{ji}\right]p_i. \tag{10.12}$$

Коэффициенты при $p_1, \ldots, p_r$ теперь равны приблизительно $\beta_1\alpha_i/[\varepsilon(\lambda_i-\mu)]$ $(i=1, \ldots, r)$ и не находятся в том специальном отношении $\alpha_1 : \alpha_2 : \ldots : \alpha_r$, которое гарантировало взаимное уничтожение. Поэтому весь вектор y^1 имеет тот же порядок величины, что и первое слагаемое в разложении (10.12), т. е. $O(\varepsilon^{-1}|\lambda_1-\mu|^{-1})$. Согласно (10.11),

$$\tau_1 = O(\varepsilon|\lambda_1-\mu|). \tag{10.13}$$

Для вектора y^2 разложение по базису $p_1, \ldots, p_n$ будет иметь вид

$$y^2 = (A-\mu I)^{-2}x^0 = \sum_{i=1}^{r} \frac{1}{(\lambda_i-\mu)^2}\left[\frac{\beta_1\alpha_i}{\varepsilon} + \sum_{j=2}^{n}\beta_j\gamma_{ji}\right]p_i + \\ + \sum_{i=r+1}^{n} \frac{1}{(\lambda_i-\mu)^2}\left[\sum_{j=2}^{n}\beta_j\gamma_{ji}\right]p_i.$$

Здесь снова не происходит взаимного уничтожения слагаемых, и вектор y^2 имеет порядок $O(\varepsilon^{-1}|\lambda_1-\mu|^{-2})$. Формула (10.11) дает

$$\tau_2 = O(|\lambda_1-\mu|).$$

Из-за отсутствия множителя ε длина второй невязки процесса *больше* длины первой (ср. с (10.13)). Аналогичное рассуждение показывает, что и для бóльших k длина невязки τ_k тоже будет порядка $O(|\lambda_1-\mu|)$. Правда, может случиться, что при некотором натуральном β выполняются равенства

$$(\lambda_i-\mu)^\beta = (\lambda_j-\mu)^\beta, \qquad i, j = 1, \ldots, r. \tag{10.14}$$

Тогда специальное соотношение коэффициентов при $p_1, \ldots, p_r$ восстановится, произойдет взаимное уничтожение и длина вектора y^β будет иметь порядок $O(|\lambda_1-\mu|^{-\beta})$. В то же время $\|y^{\beta-1}\| =$

$= O(\varepsilon^{-1}|\lambda_1-\mu|^{-\beta+1})$. Поэтому $\tau_\beta = O(|\lambda_1-\mu|/\varepsilon)$, т. е. невязка β-го приближения может стать большой. Правда, $\tau_{\beta+1}$ будет, как и τ_1, величиной порядка $O(\varepsilon|\lambda_1-\mu|)$, и далее цикл повторится. Этот вариант возможен только при наличии кластера собственных значений, равноудаленных от сдвига μ.

«Неправильное» поведение невязок не мешает последовательным итерациям метода сходиться к направлению собственного вектора для λ_1. Сходимость нарушается только в случае кластера (10.14).

Пример 10.1 (см. [164]). Матрица

$$A=\begin{bmatrix} 1 & 1 \\ 10^{-10} & 1 \end{bmatrix}$$

имеет собственные значения $\lambda_1 = 1-10^{-5}$, $\lambda_2 = 1+10^{-5}$ и собственные векторы $p_1=(1,\ -10^{-5})^T$, $p_2=(1,\ 10^{-5})^T$. Векторы p_1 и p_2 «почти» линейно зависимы:

$$-\frac{1}{\sqrt{2}}p_1+\frac{1}{\sqrt{2}}p_2=\sqrt{2}\times 10^{-5}\begin{pmatrix}0\\1\end{pmatrix}.$$

Следовательно, $\varepsilon=\sqrt{2}\times 10^{-5}$. Для вектора $x^0=(\cos\theta,\ \sin\theta)^T$ разложение по базису p_1, p_2 имеет вид

$$x^0=\frac{1}{2}(10^5\sin\theta+\cos\theta)p_2-\frac{1}{2}(10^5\sin\theta-\cos\theta)p_1.$$

Если $\sin\theta$ не очень мал, коэффициенты разложения будут большими.

В качестве сдвига возьмем $\mu=1-2\times 10^{-5}$. Может показаться, что такой сдвиг не слишком хорош, поскольку отличается от λ_1 уже в пятом десятичном разряде после запятой. Однако при работе в десятиразрядной десятичной арифметике на бóльшую точность вычисления λ_1 и нельзя рассчитывать, поскольку для матрицы собственных векторов $\operatorname{cond} P = O(10^5)$. Заметим, что μ является *точным* собственным значением матрицы

$$\begin{bmatrix} 1 & 1 \\ 4\times 10^{-10} & 1 \end{bmatrix},$$

получаемой из A возмущением порядка 10^{-10}.

Пусть $x^0=(0,\ 1)^T$, т. е. $\sin\theta=1$. Для трех первых обратных итераций в табл. 3 показаны ненормированные векторы $y^k=(A-\mu I)^{-1}x^{k-1}$, нормированные векторы x^k и отвечающие им невязки r_k. Поведение последних точно соответствует теоретическому анализу: первая невязка r_1 имеет порядок 10^{-10} (а вектор y^1 — норму порядка 10^{10}), длины следующих невязок — уже только порядка 10^{-5}, т. е. $\|r_k\|\approx\|r_1\|/\varepsilon$ $(k>1)$. Поскольку μ ближе к λ_1, векторы x^k сходятся, хотя и медленно, к собственному вектору p_1.

Этот же пример дает нам возможность проиллюстрировать особый случай (10.14) (см. табл. 4). Если взять значение $\mu=1$, равноудаленное от λ_1 и λ_2, то (10.14) выполняется при $\beta=2$. Снова в соответствии с теоретическим прогнозом имеет место цикл; при

этом длина невязки от значения 10^{-10} возрастает до 1, опять уменьшается до 10^{-10} и т. д. Векторы x^k поочередно совпадают с координатными векторами e_2 и e_1.

Таблица 3

x_0	y_1	x_1	r_1
0	$-\frac{1}{3}\times 10^{10}$	1	0
1	$\frac{2}{3}\times 10^{5}$	-2×10^{-5}	-3×10^{-10}
x_0	y_2	x_2	r_2
0	$\frac{4}{3}\times 10^{5}$	1	$\frac{3}{4}\times 10^{-5}$
1	$-\frac{5}{3}$	$-\frac{5}{4}\times 10^{-5}$	$-\frac{3}{2}\times 10^{-10}$
x_0	y_3	x_3	r_3
0	$\frac{13}{12}\times 10^{5}$	1	$\frac{12}{13}\times 10^{-5}$
1	$-\frac{7}{6}$	$-\frac{14}{13}\times 10^{-5}$	$-\frac{15}{13}\times 10^{-10}$

Обсудим теперь связь обратных итераций в вещественном случае с методом Ньютона для решения нелинейных систем [164]. Чтобы воспользоваться этим последним, придадим задаче вычисления собственного значения λ и нормированного собственного вектора x вид нелинейной системы

$$Ax-\lambda x=0, \qquad x^{\mathrm{T}}x=1. \tag{10.15}$$

Если в качестве начальных значений взять приближенную собственную пару μ, x^0, то для поправок $\Delta\mu$ и Δx^0 имеем систему

$$A(x^0+\Delta x^0)-(\mu+\Delta\mu)(x^0+\Delta x^0)=0,$$
$$(x^0+\Delta x^0)^{\mathrm{T}}(x^0+\Delta x^0)=1.$$

Поправки $d\mu$ и dx^0 по методу Ньютона получатся, если пренебречь в этих уравнениях величинами второго порядка:

$$(A-\mu I)(x^0+dx^0)=(d\mu)x^0, \tag{10.16}$$

$$(x^0)^{\mathrm{T}}dx^0=0. \tag{10.17}$$

Уравнение (10.17) показывает, что приращение вектора x^0 имеет ортогональное к x^0 направление. Что касается уравнения (10.16), то оно, в сущности, совпадает с уравнением (10.6) обратной итерации при $k=0$. Единственное отличие в том, что вектор $x^1=x^0+dx^0$ метода Ньютона не нормируется.

Таблица 4

x_0	y_1	x_1	r_1	y_2	x_2	r_2	y_3	x_3	r_3
0	10^{10}	1	0	0	0	1	10^{10}	1	0
1	0	0	10^{-10}	1	1	0	1	0	10^{-10}

Несмотря на фактическое совпадение основных уравнений, между обоими методами имеются и заметные расхождения. Разность x^1-x^0 в методе обратных итераций, вообще говоря, не ортогональна к x^0. Кроме того, метод Ньютона предполагает пересчет не только приближенного собственного вектора x^0, но и приближенного собственного значения μ. Что касается метода Виландта, то обратные итерации в нем идут с постоянным сдвигом μ.

Первое из отмеченных расхождений можно устранить изменением способа нормировки. Если x^0 — достаточно хорошее приближение к собственному вектору и ξ_m^0 — компонента вектора x^0, имеющая максимальный модуль, то и у последующих векторов x^k и у собственного вектора x m-я компонента будет если не максимальной, то одной из самых больших. Поэтому можно принять способ нормировки, состоящий в том, что во всех векторах x^k поддерживается значение 1 для m-й компоненты. Пусть для определенности $m=n$. Тогда в методе обратных итераций τ_k^{-1} превращается в значение n-й компоненты вектора y^k. Что касается метода Ньютона, то в формулировке (10.15) второе уравнение нужно заменить на $e_n^{\mathrm{T}}x=1$. Теперь, повторяя прежние рассуждения, снова получим (10.16), а вместо (10.17) — уравнение $e_n^{\mathrm{T}}dx^0=0$. Итак, приращение n-й компоненты в обоих методах равно нулю, и векторы x^1 совпадают.

Однако если значение μ не будет пересчитано, то последующие векторы x^k методов Виландта и Ньютона различны. Как известно, при некоторых условиях метод Ньютона асимптотически сходится квадратично. Достаточным условием является невырожденность якобиана системы на разыскиваемом решении. Посмотрим, что представляет собой якобиан нашей задачи и в каких случаях он не вырожден.

Поскольку $\xi_n=1$, то в системе только n неизвестных: μ и $n-1$ компонент вектора x. С помощью вектора-поправки

$$\delta=\begin{bmatrix} \{dx^{k-1}\}_1 \\ \cdots \\ \{dx^{k-1}\}_{n-1} \\ d\mu_{k-1} \end{bmatrix} \tag{10.18}$$

уравнение (10.16) (с заменой индекса 0 на $k-1$) можно записать в виде

$$\tilde{A}_{k-1}\delta=-(A-\mu_{k-1}I)x^{k-1}=-r_{k-1}. \tag{10.19}$$

Матрица $\tilde{A}_{k-1}$ получается из $A-\mu_{k-1}I$, если вместо последнего столбца поставить вектор $(-x^{k-1})$. Уравнение (10.19) означает, что якобианом задачи в точке (x^{k-1}, μ_{k-1}) является матрица $\tilde{A}_{k-1}$.

Покажем, что в случае простого собственного значения λ якобиан $\tilde{A}$ не может быть вырожден в точке (x, λ). Пусть, в самом деле,

$\tilde{A}y=0$ для некоторого ненулевого вектора y. В более подробной записи это означает, что

$$(A-\lambda I)\begin{pmatrix}\tilde{y}\\0\end{pmatrix}-\eta_n x=0, \tag{10.20}$$

где

$$y=\begin{pmatrix}\tilde{y}\\ \eta_n\end{pmatrix}\begin{matrix}\}n-1\\ \}1\end{matrix}.$$

Если $\eta_n=0$, то $z=(\tilde{y}^T|0)^T$ — собственный вектор матрицы A, относящийся к числу λ, причем не пропорциональный собственному вектору x, последняя компонента которого не равна нулю. Если же $\eta_n\neq 0$, то, умножая (10.20) на $A-\lambda I$, получим $(A-\lambda I)^2 z=0$, т. е. z — корневой вектор высоты 2 для λ. И в том, и в другом варианте λ должно быть не менее чем двукратным собственным значением матрицы A.

Итак, метод Ньютона для вычисления *простого* собственного значения λ и соответствующего собственного вектора x асимптотически сходится квадратично. Это было показано для конкретной нормировки $e_n^T x=1$, но верно и для других, например для евклидова условия $(x, x)=1$.

Учитывая связь методов Виландта и Ньютона, заключаем, что асимптотически квадратичную сходимость можно получить и для обратных итераций, если от шага к шагу изменять значение сдвига μ в соответствии с (10.18) и (10.19). В результате приходим к следующей вычислительной схеме.

Обратные итерации с переменным сдвигом: (10.21)

1. Выбрать нормированный вектор x^0 и начальное приближение μ_0 к собственному значению.
2. Для $k=1, 2, \ldots$

вычислить нормированный вектор x^k из условия

$$(A-\mu_{k-1}I)x^k=\tau_k x^{k-1}; \tag{10.22}$$

положить $\mu_k=\mu_{k-1}+\tau_k$.

Практическое значение процесса (10.21) не слишком велико. Повторим еще раз, что типичное использование обратных итераций — это вычисление собственного вектора по найденному с рабочей точностью собственному значению. В такой ситуации, во-первых, нет надобности поправлять сдвиги, во-вторых, обычно бывает достаточно одной-двух итераций с постоянным μ. Большее число итераций может потребоваться только для очень плохо обусловленного собственного значения. Но и здесь процесс с переменным сдвигом не обязательно экономичней: ведь каждая его итерация сопряжена с разложением — при применении к (10.22) прямого метода — новой матрицы $A-\mu_k I$. В итерациях с фиксированным сдвигом выполняется лишь одно разложение.

Однако в теоретическом плане обратные итерации с переменным сдвигом важны. Например, они помогают понять асимптотические свойства QR-алгоритма (см. § 12). Некоторые сведения о процессе (10.21) даны в пп. 1 и 2 дополнений к § 10.

ДОПОЛНЕНИЯ К § 10

1. Мы пришли к обратным итерациям с переменным сдвигом, рассматривая метод Ньютона для задачи (10.15). При этом асимптотически квадратичная сходимость была установлена только для случая, когда λ — простое собственное значение.

Связь обратных итераций с методом Ньютона осознана сравнительно недавно, тогда как обратные итерации — с фиксированным или переменными сдвигами — применяются более четырех десятилетий. На протяжении этого времени изучались различные способы выбора сдвигов, не обязательно совпадающие со сдвигом по Ньютону. Один из способов, когда в качестве μ_k берется отношение Рэлея текущей итерации x^k, подробно рассмотрен нниже в п. 2. Здесь же мы хотим обратить внимание читателя на новое обстоятельство, возникающее при пользовании обратными итерациями с переменным сдвигом, если собственное значение λ кратное и при этом его индекс $l>1$. Мы опираемся на результаты статьи [79].

Если $\mu_k \to \lambda$ по любому закону, то в случае простого собственного значения λ векторы x^k, определяемые уравнением (10.22), сходятся к соответствующему собственному вектору x. Не так, вообще говоря, будет для кратного λ с неединичным индексом. В этом можно убедиться на простейшем примере жордановой клетки

$$J=\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}.$$

Единственный — с точностью до пропорциональности — собственный вектор матрицы J — это координатный вектор e_1. Предположим, что в последовательности $\{\mu_k\}$ иет нулевых сдвигов. Тогда

$$x^k=\pi_k(J-\mu_{k-1}I)^{-1}(J-\mu_{k-2}I)^{-1}\ \dots\ (J-\mu_0)^{-1}x^0, \tag{10.23}$$

где π_k — нормирующий множитель. Можно записать (10.23), полагая $\zeta_i=-\mu_i^{-1}$ $(i=0, 1, 2, \dots)$, в виде

$$x^k=\nu_k(I+\zeta_{k-1}J)(I+\zeta_{k-2}J)\ \dots\ (I+\zeta_0 J)x^0=\nu_k(I+(\zeta_0+\zeta_1+\dots+\zeta_{k-1})J)x^0.$$

Здесь учтено, что $J^2=0$; ν_k — новый нормирующий множитель. Полагая $\sigma_k=\zeta_0+\zeta_1+\dots+\zeta_{k-1}$, получаем окончательно

$$x^k=\nu_k\begin{bmatrix}1 & \sigma_k\\ 0 & 1\end{bmatrix}x^0.$$

Если последовательность $\{\sigma_k\}$ имеет конечную предельную точку σ, то для векторов x^k существует вектор накопления, с точностью до нормирующего множителя равный

$$z=\begin{bmatrix}1 & \sigma\\ 0 & 1\end{bmatrix}x^0=\begin{bmatrix}1 & \sigma\\ 0 & 1\end{bmatrix}\begin{pmatrix}\xi_1^0\\ \xi_2^0\end{pmatrix}=\begin{pmatrix}\xi_1^0+\sigma\xi_2^0\\ \xi_2^0\end{pmatrix}.$$

Таким образом, при $\xi_2^0\neq 0$ вектор z не совпадает с собственным вектором e_1.

Приведем пример последовательности $\{\mu_k\}$, обеспечивающей конечную предельную точку для $\{\sigma_k\}$. Удобно записать члены последовательности

в терминах полярных представлений $\mu_j=\tau_j e^{i\varphi_j}$. Положим $\mu_0=\tau_0 e^{i\varphi_0}$, где $\tau_0>0$, $\cos\varphi_0\neq 0$. Далее для $j=2, 4, 6, \ldots$ полагаем $\tau_j=\tau_{j-2}/2$, а φ_j подбираем из условия $\mathrm{Re}\zeta_j=(\mathrm{Re}\zeta_{j-2})/2$ (или, иначе, из уравнения $4\cos\varphi_j=\cos\varphi_{j-2}$ с выбором решения в соответствующем квадранте). Наконец, $\mu_j=\bar{\mu}_{j-1}$ для $j=1, 3, 5, \ldots$ Построенная последовательность монотонно (по модулю) сходится к собственному значению нуль, в то же время

$$\sigma_{2m}=(\zeta_0+\zeta_1)+(\zeta_2+\zeta_3)+\ldots+(\zeta_{2m-2}+\zeta_{2m-1})=2(\mathrm{Re}\zeta_0+\mathrm{Re}\zeta_2+\ldots+\mathrm{Re}\zeta_{2m-2})=$$

$$=2\mathrm{Re}\zeta_0\left(1+\frac{1}{2}+\left(\frac{1}{2}\right)^2+\ldots+\left(\frac{1}{2}\right)^{m-1}\right),$$

т. е. $\{\sigma_{2m}\}$ имеет конечный предел $4\mathrm{Re}\zeta_0$.

Достаточное условие, предотвращающее подобного рода несходимость, сформулировано в той же работе [79]. Пусть сдвиги μ_j $(j=0, 1, 2, \ldots)$ сходятся к собственному значению λ индекса $l>1$. Положим $\mu_j-\lambda=\tau_j e^{i\varphi_j}$ $(\tau_j\geqslant 0)$. Если существуют целое число $k\geqslant 0$ и вещественное φ, такие, что

$$|\varphi-\varphi_j|\leqslant\frac{\pi}{4(l-1)}, \qquad j\geqslant k,$$

то (в точной арифметике) векторы x^j, генерируемые обратными итерациями со сдвигами μ_j, сходятся к собственному вектору.

Для вещественных λ и μ_j отсюда следует сходимость обратных итераций, если μ_j приближаются к λ с какой-нибудь определенной стороны. Как показывает практика, такое одностороннее приближение к λ обычно имеет место, когда в качестве сдвигов берутся отношения Рэлея.

2. Пусть x — приближенный собственный вектор матрицы A. Если нужно указать приближение к соответствующему собственному значению, то наилучшим — с точки зрения евклидовой длины вектора-невязки — будет выбор отношения Рэлея $\rho(x)=(Ax, x)/(x, x)$. Это мотивирует рассмотрение такого варианта схемы (10.21), в котором последний оператор заменяется на «положить $\mu_k=\rho(x^k)$». Для полученного метода будем пользоваться английской аббревиатурой RQI (Rayleigh Quotient Iteration — обратные итерации со сдвигами Рэлея).

Изучение метода RQI было начато в работах Темпля [134] и Крэндалла [85]. Однако первое исчерпывающее описание локального поведения RQI и некоторых его вариантов дано в цикле работ Островского [150]. В частности, Островский доказал, что для эрмитовой матрицы A сходимость RQI асимптотически кубическая. Это доказательство существенно использует стационарность отношения Рэлея на собственных векторах матрицы. В комплексной области отношение Рэлея $\rho(x)$ не будет аналитической функцией компонент ненулевого вектора x (из-за наличия операции сопряжения в определении скалярного произведения). Тем не менее для каждого ненулевого x и каждого направления v существует производная $\rho'(x, v)$ по направлению v:

$$\rho'(x, v)=\lim_{\substack{t\to 0\\ t\in R^+}}\{[\rho(x+tv)-\rho(x)]/t\}=v^*(A-\rho(x)I)x+x^*(A-\rho(x)I)v.$$

Стационарность означает, что $\rho'(x, v)=0$ для любого направления v. Она имеет место, если

$$(A-\rho(x)I)x=0, \qquad x^*(A-\rho(x)I)=0,$$

т. е. если x одновременно правый и левый собственный вектор для матрицы A. Совпадение правых и левых собственных векторов имеет место не только

для эрмитовых, но и для более общего класса нормальных матриц. Пользуясь этим, Парлетт [158] перенес на нормальные матрицы доказательство асимптотически кубической сходимости метода RQI.

Следующим шагом после работ Островского было описание глобального поведения RQI в эрмитовом случае, полученное в 1966 г. Каханом (см. [35, § 4.8, 4.9]). Оказалось, что нормы невязок в RQI монотонно убывают, и последовательность $\{\mu_k\}$ всегда сходится. Как правило, пары $\{\mu_k, x^k\}$ сходятся к точной собственной паре (λ, p), причем, как мы уже знаем, асимптотически скорость сходимости будет кубической. Однако имеется особый случай, когда числа μ_k сходятся лишь линейно к полусумме некоторых двух собственных значений — $\tilde{\lambda}$ и $\tilde{\tilde{\lambda}}$. Последовательность $\{x^k\}$ вообще не сходится, но подпоследовательности, образованные соответственно нечетными и четными членами, сходятся к биссекторам пары собственных векторов для $\tilde{\lambda}$ и $\tilde{\tilde{\lambda}}$. Малые возмущения векторов x^k выводят из этого особого случая и возвращают процесс к основному варианту с асимптотически кубической сходимостью.

И этот результат Кахана перенесен Парлеттом [158] на класс нормальных матриц. В приведенной выше формулировке нужно изменить только последнюю часть: в варианте с линейной сходимостью пределом последовательности $\{\mu_k\}$ является точка, равноудаленная от s $(s \geqslant 2)$ собственных значений матрицы A. Последовательность $\{x^k\}$ не сходится и в отличие от эрмитовой ситуации может не иметь предельного цикла. Особый режим сходимости снова неустойчив относительно возмущений векторов x^k.

В третьей из статей своего цикла Островский рассмотрел следующее обобщение RQI на анормальные матрицы.

Двусторонние итерации Островского: (10.24)

1. Выбрать нормированные векторы x^0, y^0 такие, что $(x^0, y^0) \neq 0$.
2. Для $k = 1, 2, \ldots$

вычислить обобщенное отношение Рэлея

$$\mu_{k-1} = \rho(x^{k-1}, y^{k-1}) = (Ax^{k-1}, y^{k-1})/(x^{k-1}, y^{k-1});$$

вычислить векторы x^k, y^k из условий

$$(A - \mu_{k-1} I) x^k = \tau_k x^{k-1}, \quad (A^* - \bar{\mu}_{k-1} I) y^k = \varkappa_k y^{k-1},$$

где τ_k, $\varkappa_k$ — нормирующие множители;
проверить, что $(x^k, y^k) \neq 0$; если это условие нарушено, процедура заканчивается безрезультатно.

Если $x^k \to p$, а $y^k \to q$, где p и q — правый и левый собственные векторы матрицы A, причем $(p, q) \neq 0$, то μ_k сходится к соответствующему собственному числу λ и скорость сходимости асимптотически кубическая. Однако глобальной сходимости процесс Островского не имеет, что видно хотя бы из возможности досрочного аварийного выхода.

Парлетт [158] рассматривает другое анормальное обобщение метода RQI.

Попеременные RQI *итерации:* (10.25)

1. Выбрать нормированный вектор x^0.
2. Для $k = 1, 3, 5, \ldots$

вычислить отношение Рэлея
$\mu_{k-1} = \rho(x^{k-1}) = (Ax^{k-1}, x^{k-1})$;

вычислить вектор x^k из условия

$$(A - \mu_{k-1} I) x^k = \tau_k x^{k-1};$$

вычислить отношение Рэлея

$\mu_k = \rho(x^k)$;

вычислить вектор x^{k+1} из условия

$(A^* - \bar{\mu}_k I) x^{k+1} = \varkappa_{k+1} x^k$.

Как и выше, τ_k и $\varkappa_{k+1}$ — нормирующие множители.

Для процесса (10.25) в [158] доказаны монотонное убывание длин невязок (отдельно по подпоследовательностям четных и нечетных членов) и глобальная сходимость $\mu_k \to \rho$, $x^{2k-1} \to p$, $x^{2k} \to q$, где p и q — либо правый и левый собственные векторы матрицы A для собственного значения ρ, либо правый и левый сингулярные векторы матрицы $A - \rho I$, отвечающие одному и тому же сингулярному числу. Однако асимптотическая сходимость метода для анормальных матриц очень медленная — в лучшем случае линейная.

3. В [164] дано еще одно объяснение тому, что для плохо обусловленного собственного значения λ первая итерация метода Виландта может дать лучший — в смысле длины невязки — результат, чем следующие за ней. Пусть $v_1, \ldots, v_n$ — правые, а $u_1, \ldots, u_n$ — левые сингулярные векторы матрицы $A - \mu I$; нумерация тех и других отвечает упорядочеиию сиигулярных чисел по убыванию: $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_n$. Матрица $A - \mu I$ есть малое возмущение вырожденной матрицы $A - \lambda I$; разыскиваемый собственный вектор p отвечает в последней матрице числу нуль; соответствующий левый собственный вектор обозначим через q; и p, и q считаем нормированными. По предположению число $|(p, q)|$ мало. Поскольку σ_n, v_n, u_n суть возмущения соответственно для 0, p и q, то мало и число $|(u_n, v_n)|$.

Если качество приближенного собственного вектора оценивать по величине невязки, которую он дает относительно приближенного собственного значения μ, то наилучшим следует признать вектор v_n, поскольку

$$\min_{\|z\|_2 = 1} \|(A - \mu I) z\|_2 = \sigma_n = \|(A - \mu I) v_n\|_2.$$

Соответственно этому наилучшим начальным приближением для метода Вилаидта нужно считать u_n, потому что

$(A - \mu I) v_n = \sigma_n u_n$.

Предположим, что, найдя в первой итерации вектор v_n, мы тем не менее продолжаем процесс (10.5). Чтобы оценить результат второй итерации, разложим v_n по левому сингулярному базису:

$$v_n = \alpha_1 u_1 + \ldots + \alpha_n u_n, \qquad |\alpha_1|^2 + \ldots + |\alpha_n|^2 = 1.$$

В этом разложении $\alpha_n = (v_n, u_n)$ есть малый коэффициент. Поэтому в векторе

$$(A - \mu I)^{-1} v_n = \alpha_1 (A - \mu I)^{-1} u_1 + \ldots + \alpha_n (A - \mu I)^{-1} u_n = \frac{\alpha_1}{\sigma_1} v_1 + \ldots + \frac{\alpha_n}{\sigma_n} v_n,$$

с точностью до нормирующего множителя, совпадающем с вектором x^2 метода Виландта, компонента по «наилучшему» направлению v_n может не быть доминирующей, или, во всяком случае, ее доминирование ослаблено вследствие малости α_n.

Для наглядиости мы рассмотрели случай $x^0 = u_n$. Однако сходные явления имеют место для почти любого начального вектора; достаточно, чтобы компонента вектора x^0 в направлении u_n не была слишком мала.

4. Необходимо упомянуть о вкладе, внесенном в осмысление обратных итераций работами Вараха [197—199].

5. В работе [78] рассматривается обобщение метода Ньютона на случай вычисления базиса инвариантного подпространства, соответствующего кластеру плохо обусловленных собственных значений.

ГЛАВА 4. QR-АЛГОРИТМ И ЕГО ПРИЛОЖЕНИЯ

Если нужно вычислить все или большую часть собственных значений матрицы A не слишком высокого порядка n, не принадлежащей ни к какому специальному классу и умещающейся — в виде квадратного $n \times n$-массива — в оперативной памяти ЭВМ, то в выборе метода не может быть никаких сомнений: следует пользоваться QR-алгоритмом (мы говорим о компьютерах последовательного действия; для машин новых типов архитектуры — параллельных, векторных и т. п. — ситуация в области спектральных задач еще не определилась с полной ясностью). Увеличивая требования к объему оперативной памяти — два $n \times n$-массива вместо одного, — мы можем найти не только собственные значения, но и все собственные векторы.

В первом приближении под QR-алгоритмом можно понимать итерационный процесс, протекающий при $k = 1, 2, \ldots$ по формулам

$$A_k = Q_k R_k, \tag{4.0.1}$$

$$R_k Q_k = A_{k+1}. \tag{4.0.2}$$

Первая формула означает, что очередная матрица A_k подвергается *ортогонально-треугольному* или — в комплексном случае — *унитарно-треугольному разложению* (более коротко, QR-*разложению*). Согласно (4.0.2), сомножители QR-разложения перемножаются в обратном порядке, порождая следующую матрицу A_{k+1}. В качестве начальной матрицы A_1 берут матрицу A.

QR-алгоритм в этой форме — так называемый *основной* QR-*алгоритм* — предложен в начале 60-х годов (независимо и почти одновременно) советским математиком Кублановской [26] и англичанином Фрэнсисом [109]. Для любой исходной матрицы A метод генерирует последовательность матриц A_k, которые унитарно (ортогонально) подобны матрице A и с ростом k почти всегда сходятся к матрице (блочной) верхней треугольной формы. В надлежащий момент можно прекратить QR-*итерации* (или QR-*шаги*) (4.0.1), (4.0.2), приняв диагональные элементы (и собственные значения диагональных блоков) последней матрицы процесса A_f за приближения к собственным значениям. При необходимости из A_f и промежуточных матриц $Q_1, Q_2, \ldots$ можно определить собственные векторы матрицы A; в этом случае нужно параллельно с (4.0.1), (4.0.2) накапливать произведение матриц Q_k.

Современный QR-алгоритм — метод, реализованный множеством библиотечных подпрограмм, среди которых в первую очередь нужно

назвать подпрограммы известного пакета EISPACK [180, 112],— значительно сложнее, чем описанная выше процедура. Дело в том, что основной QR-алгоритм сходится, если, конечно, сходится, очень медленно, ничем не отличаясь в этом отношении от степенного метода. Каждая QR-итерация для матрицы A общего вида очень дорога: нужно выполнить ортогонально-треугольное разложение, а затем перемножить две матрицы порядка n. При вычислении собственных векторов работа, затрачиваемая на QR-итерацию, еще увеличивается.

Трудоемкость QR-итераций можно существенно снизить, если до начала процесса (4.0.1), (4.0.2) привести матрицу A к хессенберговой форме. Эта форма сохраняется QR-итерациями. Сходимость метода значительно ускоряется, если использовать *сдвиги*, т. е. проводить QR-итерации для матриц вида $A_k - \mu_k I$ при подходящим образом выбранных числах μ_k. Применение сдвигов может, даже если матрица A вещественна, потребовать выполнения комплексных арифметических операций. Однако для вещественной A этого можно избежать, прибегая к так называемым *двойным* QR-*шагам*. Перечисленные приемы вместе с рядом других и составляют современный QR-алгоритм. Мы будем называть его *практическим* QR-*алгоритмом*.

Для практического QR-алгоритма в отличие от основного пока нет завершенной теории сходимости. Есть полное понимание симметричного (или комплексного эрмитова) случая, есть результаты относительно локальной сходимости для матриц общего вида, наконец, доказана глобальная сходимость метода для некоторых классов матриц, промежуточных между матрицами общего вида и симметричными. На практике же QR-алгоритм успешно работает всегда или почти всегда.

Обсуждение метода в этой главе построено следующим образом. Параграф 11 предназначен в первую очередь для читателя, который хотел бы воспользоваться библиотечными подпрограммами QR-алгоритма, а потому для компоновки своей программы в целом должен знать, из каких этапов состоит метод. Объяснение смысла и целей каждого этапа сопровождается подробным обсуждением средств, какими эти цели достигаются. Разбивка на этапы соответствует отдельным процедурам справочника [43] или пакета EISPACK [180]. Вопросам сходимости QR-алгоритма посвящен § 12. Читатель, склонный довериться тезису о практически безотказной сходимости метода, может пропустить этот параграф. Некоторые из многочисленных приложений QR-алгоритма, а именно вычисление функций от матриц и решение матричных уравнений, разбираются в § 13.

§ 11. Основные этапы QR-алгоритма, их вычислительные схемы

Вычисление собственных значений матрицы A посредством QR-алгоритма можно разбить на следующие этапы:

1. Уравновешивание (или масштабирование) матрицы:

$$A \to B = D^{-1} A D.$$

2. Приведение матрицы к форме Хессенберга:

$B \to H = P^{-1}BP.$

3. Итерационный QR-процесс:

$H_1 = H, \quad H_k = Q_k R_k, \quad R_k Q_k = H_{k+1}, \qquad k = 1, 2, \ldots$

Возможны, как увидим далее, и более сложные итерационные формулы.

Если нужны также и собственные векторы, то, во-первых, изменяется содержание этапа 3, где теперь, по существу, строится форма Шура T матрицы H и вычисляются ее собственные векторы, и, во-вторых, добавляется еще один этап:

4. Преобразование собственных векторов матрицы T (или матрицы H) в собственные векторы матрицы A.

Форма Шура данной матрицы определена неоднозначно, и в некоторых приложениях может потребоваться перейти от формы, вычисленной QR-алгоритмом, к более подходящей форме Шура. Этот переход, строго говоря, не является частью QR-алгоритма, но он используется в нескольких последующих процедурах, и нам удобно описать технику его выполнения именно здесь. Итак,

5. Переход к форме Шура с заданным расположением собственных значений на главной диагонали.

Далее мы подробно обсудим каждый из этапов.

1. Уравновешивание матрицы. Почти все арифметические операции над числами с плавающей точкой выполняются с ошибками округлений. Разумеется, возможны исключения, например сложение или вычитание чисел одинакового порядка или перемножение чисел с короткими мантиссами, но это именно исключения; ошибки же округления являются общим правилом. Поэтому мы не можем рассчитывать на то, что вычислим точные собственные значения заданной матрицы A. Лучшее, на что можно надеяться, это то, что каждое из полученных чисел будет *точным собственным значением матрицы* $\tilde{A} = A + F$, *близкой к* A. Близость $\tilde{A}$ к A означает малость матрицы возмущения F; согласно § 2, в алгебраической практике матрицу возмущения F считают малой, если для ее нормы удается получить оценку вида

$$\| F \| \leqslant f(n) \| A \| \beta^{-t}. \tag{11.1}$$

Возмущение F может, согласно § 6, привести к возмущению в собственном значении λ, достигающему величины $k(\lambda) \cdot \| F \|$, где $k(\lambda)$ — число обусловленности λ. Отсюда и из оценки (11.1) следует, что, даже если собственные значения матрицы обусловлены хорошо, т. е. $k(\lambda) \approx 1$ для каждого из них, мы все же не сможем определить самые младшие с *высокой относительной точностью*. Мешает то, что $\| F \|$ пропорциональна $\| A \|$, а последняя норма (см. (3.1)) не меньше, чем $|\lambda|$, для любого собственного значения λ, причем для младших из них отношение $\| A \| / |\lambda|$ может быть очень велико.

Однако для матриц, далеких от нормальных, отношение $\| A \| / |\lambda|$ велико и для старших собственных значений. Между тем от

спектрального метода естественно требовать, чтобы по крайней мере эти собственные значения определялись с малыми относительными ошибками. Уравновешивание есть процедура, помогающая удовлетворить такое требование. Именно до применения основного алгоритма проводится вспомогательный (и не требующий большой дополнительной работы) процесс *уменьшения нормы матрицы*, сближающий ее (норму) со спектральным радиусом.

Для матричных норм гельдерова семейства

$$\| A \|_{l_p} = \left[\sum_{i,j=1}^{n} |a_{ij}|^p \right]^{1/p}, \qquad p \geqslant 1,$$

удобно при уменьшении нормы пользоваться диагональными преобразованиями подобия. При переходе от матрицы A к матрице $G = \mathscr{D}_i^{-1}(\alpha) A \mathscr{D}_i(\alpha)$ $(\alpha > 0)$ изменятся только внедиагональные элементы i-й строки и i-го столбца, и

$$\| A \|_{l_p}^p - \| G \|_{l_p}^p = (1-\alpha^p) \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ji}|^p + \left(1 - \frac{1}{\alpha^p}\right) \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|^p \equiv$$

$$\equiv (1-\alpha^p) C_i^{(p)}(A) + (1 - 1/\alpha^p) R_i^{(p)}(A).$$

Наибольшее значение эта разность принимает для

$$\alpha = \left[R_i^{(p)}(A) / C_i^{(p)}(A) \right]^{1/(2p)}.$$

Циклически меняя индекс i и выполняя несколько циклов, мы добиваемся желаемого уменьшения нормы. Процедура balance (алгоритм II.11) из справочника [43] и подпрограммы BALANC, CBAL пакета EISPACK [180] реализуют именно этот процесс*) со следующими отклонениями и добавлениями:

1) чтобы сам процесс масштабирования не вносил ошибок в элементы матрицы, вместо числа $\alpha = \left[R_i^{(p)}(A) / C_i^{(p)}(A) \right]^{1/(2p)}$ берется число $\hat{\alpha} = \beta^{\sigma}$, где σ определяется неравенствами $\beta^{2\sigma-1} < \alpha^2 \leqslant \beta^{2\sigma+1}$. В этом случае подобное преобразование с матрицей $\mathscr{D}_i(\hat{\alpha})$ сведется к изменению (соответственно на σ и $-\sigma$) порядков элементов i-го столбца и i-й строки; мантиссы не изменятся и, следовательно, ошибок округлений не будет;

2) если уменьшение нормы, достигаемое для данного i, слишком незначительно, а именно

$$C_i^{(p)}(A) \hat{\alpha}^p + R_i^{(p)}(A) / \hat{\alpha}^p \geqslant \gamma \left(C_i^{(p)}(A) + R_i^{(p)}(A) \right), \tag{11.2}$$

где $\gamma \leqslant 1$ — задаваемое пользователем пороговое значение, то для индекса i подобное преобразование не выполняется. Если неравенство (11.2) справедливо для каждого i, то получена матрица с почти минимальным значением нормы, и процесс можно закончить;

3) если все внедиагональные элементы i-й строки или i-го столбца равны нулю (такие строки и столбцы будем называть *вырожда-*

*) Подпрограмма BALANC осуществляет уравновешивание для вещественной матрицы, а подпрограмма CBAL — для комплексной матрицы.

ющимися), то подобное преобразование невозможно, поскольку матрица $\mathscr{D}_i(\alpha)$ или вырождена, или не имеет смысла. Однако тогда и само уравновешивание i-й строки и i-го столбца излишне: ведь элемент $a_{ii}=\lambda$ есть собственное значение матрицы A. Разумно будет переставить строки и столбцы с номерами i и n так, чтобы матрица приобрела вид

$$\begin{bmatrix} A_{n-1} & a_{n-1} \\ 0 & \lambda \end{bmatrix} \tag{11.3}$$

(для определенности мы говорим о случае, когда нулевыми являются внедиагональные элементы *строки*), и последующие преобразования выполнять, ориентируясь на подматрицу A_{n-1}. Если в ней в свою очередь есть вырождающиеся строки и (или) столбцы, то возможна дальнейшая редукция порядка. Заметим, что если вырождающиеся строки переводятся вниз, как это было в (11.3), то вырождающиеся столбцы нужно, напротив, ставить в первые позиции. Так, матрицу

$$A=\begin{bmatrix} 6 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 4 & 0 & 3\times10^{-4} & 1\times10^{-2} & 2\times10^{-2} & 0.1 \\ 1 & 100 & 7 & 0 & 0 & -2 & 20 \\ 0 & 2\times10^{4} & 0 & 1 & -400 & 300 & -4\times10^{3} \\ -2 & -300 & 0 & 1\times10^{-2} & 2 & 2 & 40 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & 0 & 4\times10^{-3} & 0.1 & -0.2 & 3 \end{bmatrix}$$

процедура balance (минимизирующая норму с показателем $p=1$) преобразует в матрицу

$$B=\begin{bmatrix} 7 & 0.1 & 0.2 & 0 & 0 & 1 & -2 \\ 0 & 4 & 1 & 3 & 1 & 0 & 20 \\ 0 & 1 & 3 & 4 & 1 & 0 & -20 \\ 0 & 2 & -4 & 1 & -4 & 0 & 30 \\ 0 & -3 & 4 & 1 & 2 & -20 & 20 \\ 0 & 0 & 0 & 0 & 0 & 6 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(см. [43, с. 301]). В ходе этого преобразования устанавливается, что числа $\lambda_1=7$, $\lambda_6=6$ и $\lambda_7=0$ являются точными собственными значениями матрицы A. Остальные собственные значения составляют спектр подматрицы

$$B_4=\begin{bmatrix} 4 & 1 & 3 & 1 \\ 1 & 3 & 4 & 1 \\ 2 & -4 & 1 & -4 \\ -3 & 4 & 1 & 2 \end{bmatrix}.$$

Уравновешивание проводилось для $\beta=10$; то, что строчные и столбцевые суммы (модулей) матрицы B_4, имеющие одинаковый номер,

не совпадают, объясняется использованием в качестве чисел α целых степеней числа 10. О результативности процедуры уравновешивания в этом примере можно судить по следующим данным. Точные собственные значения матрицы B_4 (следовательно, и матрицы A) суть числа 4.5895..., $-0.077...$ и $2.7438... \pm i3.881...$ При применении QR-алгоритма к B_4 будут найдены приближения к этим числам с абсолютными погрешностями соответственно 4×10^{-10}, 0.63×10^{-10} и $(2\pm i2)\times10^{-10}$. Для матрицы A приближения к тем же числам, снова вычисленные посредством QR-алгоритма, имеют абсолютные погрешности 2.59×10^{-6}, -2.79×10^{-6} и $(0.88\pm i2.77)\times10^{-6}$. Более того, поскольку выделение вырождающихся строк и столбцов теперь не производится, приближения к собственным значениям $\lambda_1=7$ и $\lambda_6=6$ также будут получены с заметными погрешностями — соответственно -1.84×10^{-6} и 2.37×10^{-8} (для этих вычислений t в формуле (11.1) можно считать равным 10).

Заметим, что уравновешивание не является обязательным. Если у пользователя есть уверенность в том, что его матрица сама по себе масштабирована неплохо, то обращение к QR-алгоритму можно начинать с этапа 2. Если же уравновешивание проводится, то в случае, когда собственные векторы вычисляются, должна быть сохранена диагональная матрица D, равная произведению всех матриц $\mathscr{D}_i$. Она понадобится на этапе 4 при преобразовании собственных векторов матрицы H в собственные векторы матрицы A.

Что касается стоимости процесса уравновешивания относительно всей спектральной процедуры, сошлемся на [180, с. 59]; ни в одном случае уравновешивание не потребовало более 7% общего времени работы QR-алгоритма.

Уменьшая норму матрицы, уравновешивание тем самым приближает ее к множеству нормальных матриц, что особенно очевидно в случае евклидовой нормы ($p=2$). Поскольку собственные значения нормальных матриц обусловлены идеально, то уравновешивание обычно сопровождается уменьшением чисел обусловленности $k(\lambda_i)$. Да и сам способ уравновешивания в определенной степени подсказан известным свойством нормальной матрицы: евклидовы длины одноименных строки и столбца равны.

2. Приведение матрицы к форме Хессенберга. Существует ряд методов для подобного преобразования матрицы к форме Хессенберга. Чаще всего с этой целью используют методы отражений или исключения. Они и будут рассмотрены в данном разделе. Для определенности мы говорим о приведении к верхней хессенберговой матрице.

Метод отражений. Процесс приведения $n\times n$-матрицы B к форме Хессенберга состоит из $n-2$ шагов. Каждый шаг есть подобие с отражением в качестве трансформирующей матрицы. Реальные размерности отражений с каждый шагом уменьшаются на единицу — от $n-1$ в начале процесса до 2 на последнем шаге. После k шагов первые k столбцов преобразуемой матрицы приобретают и сохраняют до конца процесса нужную форму — форму Хессенберга.

Для первого шага находим отражение $\mathscr{H}_1$, такое, чтобы левое умножение B на $\mathscr{H}_1$ аннулировало элементы в позициях $(3,1)$, ..., $(n,1)$, не меняя первой строки в B. Способ построения матрицы $\mathscr{H}_1$ описан в § 2. Чтобы завершить подобие, нужно умножить матрицу $\tilde{B}=\mathscr{H}_1 B$ справа на $\mathscr{H}_1$ (напомним, что для всякого отражения $\mathscr{H}_1=\mathscr{H}_1^{-1}$). При этом умножении не меняется первый столбец и, следовательно, нули, полученные в $\tilde{B}$, сохранятся. Положим $B_2=\mathscr{H}_1 B\mathscr{H}_1$.

На втором шаге строим матрицу $\mathscr{H}_2$, которая аннулирует в B_2 элементы $(4,2)$, ..., $(n,2)$, не меняя первых двух строк. Нулевые элементы первого столбца останутся нулями и в $\tilde{B}_2=\mathscr{H}_2 B_2$. Правое умножение $\tilde{B}_2 \to B_3=\tilde{B}_2\mathscr{H}_2=\mathscr{H}_2 B_2\mathscr{H}_2$ не испортит первых двух столбцов, приобретших уже форму Хессенберга.

Продолжая таким образом, после $n-2$ шагов получим матрицу Хессенберга

$$H\equiv B_{n-1}=\mathscr{H}_{n-2}\ldots\mathscr{H}_2\mathscr{H}_1 B\mathscr{H}_1\mathscr{H}_2\ldots\mathscr{H}_{n-2}\equiv P^* BP, \tag{11.4}$$

где $P=\mathscr{H}_1\mathscr{H}_2\ldots\mathscr{H}_{n-2}$. Для $n=5$ процесс приведения иллюстрируется на рис. 2.

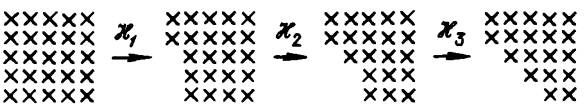

Рис. 2

Если вычисляются только собственные значения матрицы B, то хранить информацию об отражениях не обязательно. Однако стандартные подпрограммы, например процедура orthes (алгоритм II.13) из [43] и подпрограммы ORTHES, CORTH пакета EISPACK, обычно такую информацию сохраняют. Это позволяет пользоваться той же подпрограммой приведения к форме Хессенберга и тогда, когда собственные векторы нужны. В этом случае по завершении приведения надо обратиться к подпрограмме, формирующей в отдельном квадратном массиве произведение отражений $\mathscr{H}_i$.

Поскольку отражение $\mathscr{H}=I-\dfrac{1}{K^2}uu^*$ полностью описывается порождающим вектором u и числом K^2, то хранение отражений $\mathscr{H}_i$ сводится к запоминанию векторов u_i и чисел K_i^2 $(i=1, 2, \ldots, n-2)$. Формулы (2.20) показывают, что все, кроме главной, ненулевые компоненты вектора u_i совпадают с аннулируемыми элементами матрицы B_i (считаем, что $B_1\equiv B$). Оставляя последние на их месте, мы сохраняем тем самым соответствующие компоненты порождающих векторов. Дополнительные ячейки требуются только для записи главных компонент и чисел K_i^2. Впрочем, согласно (2.21), число K_i^2 можно найти по значению главной компоненты и новому значению матричного элемента в позиции $(i+1, i)$; поэтому при хранении отражений часто ограничиваются введением добавочного линейного массива длины n.

Приведение матрицы к форме Хессенберга посредством отражений является численно устойчивым процессом в смысле § 2. Именно реально вычисленная матрица Хессенберга H, хотя и не будет точно подобной исходной матрице B, подобна возмущенной матрице $\tilde{B}=B+F$; при этом (см., например, [43, с. 308]) выполняется оценка (11.1), где функция $f(n)$ имеет вид $f(n)=Cn^2$ (C — небольшая константа), а норму можно взять евклидову.

Метод исключения. В этом методе приведение к форме Хессенберга достигается с помощью цепочки элементарных преобразований матрицы. Как и в методе отражений, процесс состоит из $n-2$ шагов; на k-м шаге форму Хессенберга приобретает k-й столбец, а форма предыдущих столбцов сохраняется.

Суть метода поясним, разбирая его первый шаг. Считая элемент a_{21} ненулевым, вычтем из 3-й, ..., n-й строк матрицы B ее 2-ю строку, умноженную соответственно на числа l_{32}, ..., l_{n2}, где

$$l_{i2}=b_{i1}/b_{21}, \qquad i=3, \ldots, n. \tag{11.5}$$

В совокупности выполненные преобразования равносильны умножению матрицы B слева на элементарную матрицу вида

$$\mathscr{L}_2=\begin{bmatrix} 1 & & & & & \\ 0 & 1 & & & 0 & \\ 0 & -l_{32} & 1 & & & \\ 0 & -l_{42} & 0 & 1 & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & -l_{n2} & 0 & \ldots & 0 & 1 \end{bmatrix}. \tag{11.6}$$

Завершая подобное преобразование, умножим матрицу $\tilde{B}=\mathscr{L}_2 B$ справа на $\mathscr{L}_2^{-1}$. Это означает, что ко 2-му столбцу $\tilde{B}$ нужно прибавить ее 3-й ..., n-й столбцы, умноженные соответственно на $l_{32}, l_{42}, \ldots, l_{n2}$. Поскольку 1-й столбец на этом полушаге не меняется, то нули, полученные за счет левого умножения, сохранятся. В целом 1-й шаг описывается матричной формулой

$$B_2=\mathscr{L}_2 B \mathscr{L}_2^{-1}. \tag{11.7}$$

В изложенной форме 1-й шаг не удастся провести, если элемент b_{21} нулевой. Даже если он не равен нулю, но много меньше, чем какой-то из нижележащих элементов 1-го столбца, некоторые из чисел l_{i2}, называемых *множителями 1-го шага*, будут велики. Это приведет к росту абсолютных величин элементов матрицы B_2 по сравнению с уровнем элементов в B. Такой рост нежелателен с точки зрения численной устойчивости процесса (подробнее об этом см. в [19, § 8]). Чтобы избежать или по крайней мере ограничить его, операции по исключению в 1-м столбце предваряются подходящей

перестановкой строк и столбцов матрицы B. Именно выбирается максимальный по модулю среди поддиагональных элементов 1-го столбца; пусть это будет элемент $b_{k_1 1}$. Если $k_1 = 2$, перестановка не нужна; в противном случае переставляем строки с номерами 2 и k_1, а затем — чтобы преобразование было подобием — одноименные столбцы. Для полученной матрицы проводим исключение так, как это описано выше. Теперь 1-му шагу вместо (11.7) соответствует матричная формула

$$B_2 = \mathscr{L}_2 \mathscr{P}_2 B \mathscr{P}_2 \mathscr{L}_2^{-1}. \tag{11.8}$$

Здесь

$$\mathscr{P}_2 = \begin{cases} I, & \text{если} \quad k_1 = 2, \\ \mathscr{P}_{2,k_1}, & \text{если} \quad k_1 > 2. \end{cases} \tag{11.9}$$

Для матрицы исключения мы сохранили прежнее обозначение $\mathscr{L}_2$, а ее элементы по-прежнему вычисляются по формулам (11.5), применяемым, однако, к переупорядоченной матрице B. Поскольку элемент (2, 1) теперь имеет максимальный модуль, то множители в (11.5) по модулю ограничены единицей. Если положить $B^{(k)} = [b_{ij}^{(k)}]$,

$$b = \max_{i,j} |b_{ij}|, \qquad b_k = \max_{i,j} |b_{ij}^{(k)}|,$$

то из простых формул, связывающих элементы матриц B и B_2, легко выводится оценка

$$b_2 \leqslant 4b. \tag{11.10}$$

Аналогично протекают последующие шаги. Так, на втором шаге вначале просматривают поддиагональные элементы 2-го столбца, с тем чтобы найти элемент $b_{k_2 2}$ с максимальным модулем. Если $k_2 > 3$, то производится перестановка строк и столбцов с номерами 3 и k_2; при $k_2 = 3$ перестановка не выполняется. Вычисляя *множители 2-го шага*

$$l_{i3} = b_{i2}^{(2)} / b_{32}^{(2)}, \qquad i = 4, \ldots, n$$

(где для элементов матрицы $B^{(2)}$, несмотря на возможную перестановку, используются прежние обозначения $b_{ij}^{(2)}$), проводят исключение во 2-м столбце, вычитая 3-ю строку, умноженную на $l_{43}, \ldots, l_{n3}$, из 4-й, ..., n-й строк. Нули 1-го столбца при этом сохраняются, благо элемент (3,1) нулевой. Соответствующее правостороннее преобразование состоит в прибавлении к 3-му столбцу 4-го, ..., n-го столбцов, умноженных соответственно на $l_{43}, \ldots, l_{n3}$. Матричным описанием второго шага будет равенство

$$B_3 = \mathscr{L}_3 \mathscr{P}_3 B_2 \mathscr{P}_3 \mathscr{L}_3^{-1}, \tag{11.11}$$

где $\mathscr{P}_3$ определяется по аналогии с (11.9), а $\mathscr{L}_3$ — матрица вида

$$\mathscr{L}_3 = \begin{bmatrix} 1 & & & & & & \\ 0 & 1 & & & 0 & & \\ 0 & 0 & 1 & & & & \\ 0 & 0 & -l_{43} & 1 & & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & -l_{n3} & 0 & \ldots & 0 & 1 \end{bmatrix}.$$

Матричная формула всего процесса приведения получается объединением равенств (11.8), (11.11) и аналогичных равенств для последующих шагов:

$$H \equiv B_{n-1} = \mathscr{L}_{n-2}\mathscr{P}_{n-2} \ldots \mathscr{L}_3\mathscr{P}_3\mathscr{L}_2\mathscr{P}_2 B \mathscr{P}_2 \mathscr{L}_2^{-1}\mathscr{P}_3\mathscr{L}_3^{-1} \ldots$$
$$\ldots \mathscr{P}_{n-2}\mathscr{L}_{n-2}^{-1} \equiv Q^{-1}BQ.$$

Множители каждого шага могут быть записаны в освобождающиеся позиции аннулируемых элементов соответствующего столбца. Это и делается в процедурах elmhes, comhes (алгоритм II.13) из [43] и одноименных подпрограммах пакета EISPACK (первая из каждой пары подпрограмм предназначена для вещественных, а вторая — для комплексных матриц). В дополнительном целочисленном массиве хранятся строчные индексы $k_1, k_2, \ldots, k_{n-2}$ максимальных элементов. Если для матрицы B нужно вычислить собственные векторы, то вслед за приведением к форме Хессенберга производится обращение к подпрограмме, формирующей в отдельном квадратном массиве произведение матриц $\mathscr{L}_i$, разумеется, с учетом матриц-перестановок $\mathscr{P}_i$.

Что касается численной устойчивости метода исключения, то и здесь реально вычисленную хессенбергову матрицу H можно представить как точное подобное преобразование возмущенной матрицы $B+F$. Оценка для нормы матрицы эквивалентного возмущения имеет в общем тот же вид, что и в случае метода отражений, с точностью, однако, до дополнительного множителя $g(B)$. Этот множитель, определяемый как

$$g(B) = \max_{1 \leqslant k \leqslant n-1} b_k/b, \qquad b_1 \equiv b,$$

учитывает рост элементов матрицы на протяжении всего процесса приведения к форме Хессенберга. Отсюда его название — *коэффициент роста*. С помощью перестановок мы ограничиваем рост на каждом шаге: действует оценка (11.10), как и аналогичные оценки для последующих шагов

$$b_k \leqslant 4b_{k-1}.$$

Это означает тем не менее, что за $n-2$ шагов приведения уровень элементов в матрице может вырасти в 4^{n-2} раз, что существенно

(исключая разве совсем малые порядки n) ухудшит или вовсе обесценит вычисленные результаты. Правда, примеры матриц, для которых подобный рост имеет место, не известны, но теоретическая возможность экспоненциального возрастания существует. Если говорить не о матрицах $B^{(k)}$, а о матрице Q, трансформирующей B к форме Хессенберга, то здесь есть и примеры, на которых максимум возможного роста достигается [72]. Действительно, при любых числах β_1, β_2, ..., β_n приведение матрицы

$$B=\begin{bmatrix} \beta_1 & \beta_2 & \beta_3 & \dots & \beta_{n-1} & \beta_n \\ 1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 \\ -1 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ -1 & 0 & 0 & \dots & 1 & 0 \end{bmatrix}$$

к форме Хессенберга не требует перестановок. Как легко видеть, множители всех шагов равны -1 и в произведении $Q^{-1}= \mathscr{L}_{n-2}\dots\mathscr{L}_3\mathscr{L}_2$ последняя строка имеет вид

$$(0\ 2^{n-3}\ 2^{n-4}\dots 4\ 2\ 1\ 1),$$

а первые две строки те же, что и у единичной матрицы. Матрица Q с такой обратной при больших n очень плохо обусловлена, что неизбежно скажется на точности вычислений при преобразовании собственных векторов хессенберговой матрицы H в собственные векторы матрицы B.

Несмотря на сказанное, в численной практике сильный рост элементов матриц $B^{(k)}$ наблюдается в методе исключения редко. Поэтому пользование методом лишь немногим более рискованно, чем применение отражений. В то же время число арифметических операций в методе исключения примерно вдвое меньше.

3. Итерационный QR-*процесс.* Как уже говорилось во введении к главе, в простейшем случае QR-процесс осуществляется в форме основного QR-алгоритма:

$$H_k=Q_kR_k, \tag{11.12}$$

$$R_kQ_k=H_{k+1} \tag{11.13}$$

при $k=1, 2, \dots$

Здесь считается, что начальная матрица H_1 — это хессенбергова матрица H, вычисленная на предыдущем этапе.

В вещественной и комплексной версиях алгоритма выполняются одинаковые по смыслу операции, хотя некоторые детали и различаются. Для определенности мы начинаем обсуждение с вещественного случая, а потом даем краткий комментарий относительно комплексного.

Убедимся прежде всего, что процесс (11.12), (11.13) сохраняет хессенбергову форму матрицы. Заметим, что QR-разложение

хессенберговой матрицы H_k обычно реализуется посредством умножения H_k слева на последовательность из $n-1$ вращений:

$$\mathscr{R}_{n-1,n} \dots \mathscr{R}_{23} \mathscr{R}_{12} H_k = R_k. \tag{11.14}$$

Вращение $\mathscr{R}_{12}$ подбирается так, чтобы аннулировать в H_k элемент (2,1); в полученной матрице элемент (3,2) уничтожается за счет умножения на вращение $\mathscr{R}_{23}$, при этом поддиагональные элементы 1-го столбца остаются нулевыми; затем аннулируется элемент (4,3) и т. д. Ход процесса при $n=5$ показан на рис. 3.

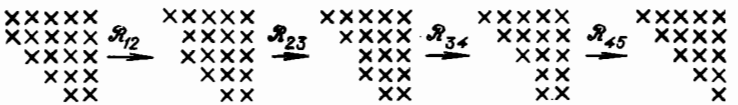

Рис. 3

Сопоставляя (11.14) и (11.12), заключаем, что

$$Q_k = \mathscr{R}^{\mathrm{T}}_{12} \mathscr{R}^{\mathrm{T}}_{23} \dots \mathscr{R}^{\mathrm{T}}_{n-1,n}.$$

Поэтому на втором полушаге нужно вычислить произведение

$$H_{k+1} = R_k \mathscr{R}^{\mathrm{T}}_{12} \mathscr{R}^{\mathrm{T}}_{23} \dots \mathscr{R}^{\mathrm{T}}_{n-1,n}. \tag{11.15}$$

То, что такое произведение имеет форму Хессенберга, очевидно из рис. 4.

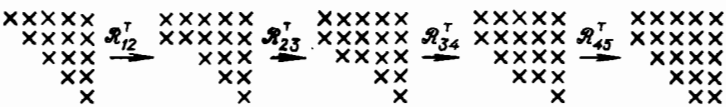

Рис. 4

Приведение матрицы H_k к треугольной форме R_k, описываемое равенством (11.14), требует приблизительно $2n^2$ операций умножения и примерно столько же операций сложения. Если бы H_k была матрицей общего вида, ее QR-разложение обошлось приблизительно в $\frac{4}{3}n^3$ операций как того, так и другого типа. Отсюда понятно, насколько ценной является *инвариантность формы Хессенберга* относительно QR-алгоритма. То же соотношение между трудоемкостью реализации в хессенберговом и общем случае имеет место для формулы (11.15).

Необходимо заметить следующее: процесс (11.12), (11.13) основан на неявном предположении, что в матрице H_1 все элементы $h_{i+1,i}$ ($i=1, 2, \dots, n-1$) отличны от нуля; другими словами, матрица H_1 *неразложима*. Действительно, если $h_{i_0+1,i_0}=0$ для некоторого i_0, то матрицу H_1 можно представить в блочно треугольной форме

$$H_1 = \begin{bmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{bmatrix}$$

с диагональными блоками порядка i_0 и $n-i_0$ соответственно. Спектр матрицы H_1 есть объединение спектров диагональных блоков, и итерирование полной матрицы H_1 имеет мало смысла; гораздо разумней проводить QR-процесс раздельно для H_{11} и H_{22}.

Пусть матрица H_1 неразложима и не вырождена, тогда неразложимы и не вырождены все матрицы H_k основного QR-алгоритма. В самом деле, неразложимость матрицы H_k обеспечивает, что диагональные элементы (1,1), (2,2), ..., $(n-1, n-1)$ треугольной матрицы R_k будут положительны (см. формулы преобразования элементов матрицы в результате вращения из § 2), а синусы углов вращений $\mathscr{R}_{12}$, ..., $\mathscr{R}_{n-1,n}$ не равны нулю. Из невырожденности H_k следует, что и последний диагональный элемент $r_{nn}^{(k)}$ не может быть нулем. Нетрудно проверить, что

$$h_{i+1,i}^{(k+1)} = -s_i r_{i+1,i+1}^{(k)},$$

где s_i — синус угла вращения $\mathscr{R}_{i,i+1}$. Таким образом, и в матрице $H^{(k+1)}$ все элементы кодиагонали (2,1), (3,2), ..., $(n, n-1)$ ненулевые, т. е. $H^{(k+1)}$ неразложима.

Если матрица H_1 неразложима, но вырождена, то первый же шаг QR-процесса приводит к расщеплению матрицы. Действительно, повторяя приведенное рассуждение, видим, что элементы $r_{11}^{(1)}$, .., $r_{n-1,n-1}^{(1)}$ по-прежнему положительны, однако элемент $r_{nn}^{(1)}$ теперь обязан быть нулем. Последняя строка матрицы R_k тем самым нулевая, и это же справедливо в отношении матрицы H_2. Таким образом, $h_{n,n-1}^{(2)}=0$, а нулевое собственное значение выделилось в явном виде на диагонали матрицы. Рассуждая «по непрерывности», заключаем, что результатом QR-шага для неразложимой и почти вырожденной матрицы H_1 будет сильное убывание модуля элемента в позиции $(n, n-1)$.

Это наблюдение лежит в основе QR-*алгоритма со сдвигами* — значительно более эффективной, чем основной алгоритм, разновидности QR-процесса. Если известно приближение τ к некоторому собственному значению матрицы H_1, то, применяя формулы (11.12), (11.13) к $H_1-\tau I$, мы оказываемся в ситуации почти вырожденности. Прибавляя к полученной матрице $R_k Q_k$ матрицу τI — *восстанавливая сдвиг*, получаем матрицу H_2, подобную H_1, но с уменьшившимся значением элемента в позиции $(n, n-1)$. Итак, рабочие формулы QR-алгоритма со сдвигами имеют вид

$$H_k - \tau_k I = Q_k R_k, \tag{11.16}$$

$$H_{k+1} = R_k Q_k + \tau_k I \tag{11.17}$$

при $k=1, 2, ...$

Как и основной алгоритм, QR-процесс со сдвигами порождает последовательность ортогонально (или унитарно) подобных матриц:

$$H_{k+1} = Q_k^* H_k Q_k, \qquad k=1, 2, ... \tag{11.18}$$

Разумеется, для различных *стратегий сдвигов*, т. е. правил выбора сдвигов τ_k, последовательности $H_2, H_3, ...$ и $Q_1, Q_2, Q_3, ...$ будут разными.

В библиотечных программах QR-алгоритма всегда реализуется тот или иной вариант процесса со сдвигами. Имеется несколько вариантов, что объясняется незавершенностью теории QR-алгоритма

со сдвигами в случае матриц общего вида. Например, не найдено ни одной единообразной стратегии сдвигов, для которой была бы доказана глобальная сходимость при выполнении некоторых обозримых условий на матрицу (не требующих ее принадлежности к какому-нибудь специальному классу, скажем классу эрмитовых матриц). Для основного QR-алгоритма такая обозримая система условий глобальной сходимости хорошо известна (см. § 12).

Чаще всего в библиотечных подпрограммах используется одна из следующих двух стратегий:

1. *Сдвиги по Рэлею*: $\tau_k = h_{nn}^{(k)}$.

2. *Сдвиги по Уилкинсону*: сдвиг τ_k определяется как собственное значение подматрицы

$$\begin{bmatrix} h_{n-1,n-1}^{(k)} & h_{n-1,n}^{(k)} \\ h_{n,n-1}^{(k)} & h_{nn}^{(k)} \end{bmatrix}. \tag{11.19}$$

Предпочтение, отдаваемое этим стратегиям, объясняется, с одной стороны, тем, что они почти всегда обеспечивают сходимость итерационного процесса — факт, установленный многолетней практической эксплуатацией программ QR-алгоритма; с другой стороны, их глобальная сходимость строго доказана в случае вещественных симметричных трехдиагональных матриц (см. [35, гл. 8]).

Поскольку в программы справочника [43] и пакета EISPACK, предназначенные для матриц общего вида, встроена стратегия Уилкинсона, мы остановимся на ней более подробно.

Сдвиг матрицы перед выполнением QR-разложения, требуемый формулой (11.16), имеет, с точки зрения вычислителя, следующий недостаток: если значение τ_k велико по сравнению с некоторыми диагональными элементами $h_{ii}^{(k)}$, то информация, заключенная в их младших разрядах, будет после сдвига утрачена. Так всегда происходит при машинном сложении большого числа с малым, и этот недостаток свойствен, разумеется, любой стратегии, а не только сдвигам по Уилкинсону. Однако в случае вещественных матриц у стратегии Уилкинсона есть и собственное, присущее ей одной неудобство. Даже если спектр матрицы H_1 чисто вещественный, собственные значения подматрицы (11.19) могут оказаться комплексными, что вызовет необходимость проводить дальнейшие операции в комплексной арифметике, которая (если говорить об операциях умножения) вчетверо более трудоемка, чем вещественная. Между тем при работе с вещественными матрицами естественно желание оставаться в области вещественных вычислений. Это желание с принципиальной стороны подкрепляется теоремой Шура, согласно которой вещественная матрица с вещественным спектром может быть ортогонально преобразована в вещественную же верхнюю треугольную матрицу. Если у вещественной матрицы имеются комплексные собственные значения, ее приведение к вещественному треугольному виду не осуществимо, однако сохраняется возможность ортогонального преобразования к вещественной форме Шура (см. § 1). Хотелось бы, чтобы в вещественном случае QR-алгоритм как раз и был практическим средством такого преобразования.

Выход из обрисованных трудностей указан Фрэнсисом [109], предложившим остроумный прием *сдваивания сдвигов*. Вкратце суть этого приема заключается в следующем. Вычислив собственные значения $\lambda_1^{(k)}$, $\lambda_2^{(k)}$ подматрицы (11.19), принимают их в качестве сдвигов для двух последовательных QR-шагов: $\tau_k=\lambda_1^{(k)}$, $\tau_{k+1}=\lambda_2^{(k)}$. Однако матрицу H_{k+1} не строят, вычисляя сразу матрицу H_{k+2}. Это удается сделать целиком в вещественной арифметике, притом не вычитая сдвига из диагональных элементов матрицы.

Перейдем теперь к более детальному разбору приема. Чтобы упростить обозначения, опустим индекс k — номер QR-итерации — и будем считать, что от исходной вещественной матрицы H_1 $(=H)$ совершается переход к матрице H_2, а затем к H_3. Согласно (11.16), (11.17), следовало бы проделать такие вычисления:

$$H_1-\tau_1 I=Q_1 R_1, \qquad H_2=R_1 Q_1+\tau_1 I, \tag{11.20}$$

$$H_2-\tau_2 I=Q_2 R_2, \qquad H_3=R_2 Q_2+\tau_2 I. \tag{11.21}$$

Разрешая соотношения (11.20), (11.21) относительно R_1 и R_2, получаем

$$R_1=Q_1^*(H_1-\tau_1 I), \tag{11.22}$$

$$R_2=(H_3-\tau_2 I)Q_2^*=(Q_2^*Q_1^*H_1Q_1Q_2-\tau_2 I)Q_2^*=$$
$$=Q_2^*Q_1^*(H_1-\tau_2 I)Q_1. \tag{11.23}$$

В последней цепочке равенств использована вытекающая из (11.18) формула

$$H_3=Q_2^*Q_1^*H_1Q_1Q_2. \tag{11.24}$$

Перемножая (11.22) и (11.23), находим

$$Q_1Q_2R_2R_1=(H_1-\tau_2 I)(H_1-\tau_1 I). \tag{11.25}$$

Положим

$$Q=Q_1Q_2, \qquad R=R_2R_1,$$

$$M=(H_1-\tau_2 I)(H_1-\tau_1 I)=H_1^2-(\tau_1+\tau_2)H_1+\tau_1\tau_2 I=$$
$$=H_1^2-(h_{n-1,n-1}+h_{nn})H_1+(h_{n-1,n-1}h_{nn}-h_{n-1,n}h_{n,n-1})I. \tag{11.26}$$

Здесь учтено, что τ_1, τ_2 — корни характеристического уравнения подматрицы (11.19) в матрице H.

Согласно (11.25), матрицы Q и R суть сомножители унитарно-треугольного разложения матрицы M. Поскольку M вещественна даже при комплексных τ_1, τ_2 (см. (11.26)), то эти сомножители могут быть выбраны вещественными. Но тогда, как следует из (11.24), матрица $H_3=Q^TH_1Q$ также вещественна, несмотря на возможную комплексность матрицы H_2. Теперь нужно найти способ прямого перехода от H_1 к H_3.

Искомый способ основан на следующем утверждении (которое мы будем называть *теоремой единственности*):

Пусть Q и P — вещественные ортогональные матрицы, трансформирующие вещественную матрицу A к неразложимым матрицам Хессенберга:

$$H_Q=Q^TAQ, \qquad H_P=P^TAP.$$

Если первые столбцы матриц Q и P совпадают, то найдется диагональная ортогональная матрица D (т. е. диагональная матрица с ± 1 на главной диагонали) такая, что

$$P=QD, \qquad H_P=DH_QD.$$

Иными словами, матрицы Q и P могут различаться разве лишь знаками столбцов, а хессенберговы матрицы H_Q и H_P — только знаками внедиагональных элементов.

Доказательство этого утверждения можно найти, например, в [42, гл. 6, § 7; 7, § 46]. В действительности это доказательство проходит и в том случае, когда предположена неразложимость только одной из хессенберговых матриц H_Q, H_P; неразложимость второй тогда является следствием остальных условий теоремы. Для себя мы делаем из нее такой вывод: если удастся построить ортогональную матрицу P, которая трансформирует H_1 в неразложимую хессенбергову матрицу $\hat{H}$ и при этом имеет такой же первый столбец, что и искомая матрица Q, то и в остальных своих столбцах P, по существу, совпадает с Q; с точностью до знаков внедиагональных элементов совпадают и хессенберговы матрицы $\hat{H}$ и H_3. Матрица $\hat{H}$ и будет конечным результатом *двойного QR-шага.*

Возвращаясь к равенству (11.25) или — в более компактной записи — к равенству

$$QR=M, \tag{11.27}$$

мы можем выяснить, как выглядит первый столбец матрицы Q. Дело в том, что произвол в ортогонально-треугольном разложении невырожденной матрицы невелик: в любых двух разложениях ортогональные матрицы могут различаться разве что знаками столбцов, а треугольные матрицы — знаками строк. Если, как это обычно делается в QR-алгоритме, дополнительно потребовать, чтобы диагональные элементы треугольной матрицы были положительны, то QR-разложение определено однозначно.

В наших обстоятельствах матрицу M можно считать невырожденной. Действительно, вырождение матрицы $M=(H_1-\tau_2 I)(H_1-\tau_1 I)$ происходит только в случае, когда хотя бы одно из чисел τ_1, τ_2 является собственным значением самой матрицы H_1 (а не только ее угловой подматрицы). Разумеется, это маловероятно. Отметим, что матрица H_3 в этом случае обязательно разложима.

Рассмотрим конкретный способ вычисления QR-разложения матрицы M, откуда будет ясен вид первого столбца матрицы Q. Прежде всего заметим, что в первом столбце матрицы M только первые три элемента не равны нулю:

$$\begin{aligned} m_{11}&=h_{11}^2-(\tau_1+\tau_2)h_{11}+\tau_1\tau_2+h_{12}h_{21},\\ m_{21}&=h_{21}(h_{11}+h_{22}-\tau_1-\tau_2),\\ m_{31}&=h_{32}h_{21}. \end{aligned} \tag{11.28}$$

Если построить отражение $\hat{\mathscr{H}}_1=I-\frac{1}{K^2}uu^{\mathrm{T}}$, аннулирующее в M поддиагональные элементы первого столбца, то порождающий вектор

u имеет лишь три ненулевые компоненты (см. формулы (2.20) при $l=m=1$, $a_{i1}=m_{i1}$):

$$u_1=m_{11}-\alpha, \quad u_2=m_{21}, \quad u_3=m_{31},$$
$$\alpha=\pm(m_{11}^2+m_{21}^2+m_{31}^2)^{1/2}, \quad u_i=0, \quad i=4, \ldots, n. \tag{11.29}$$

Уничтожив (за счет левого умножения на $\hat{\mathscr{H}}_1$) элементы m_{21} и m_{31}, мы должны в дальнейшем аннулировать в матрице $\hat{M}=\hat{\mathscr{H}}_1 M$ поддиагональные элементы второго и последующих столбцов. Каким бы образом это ни делалось, преобразования не должны затрагивать первую строку, иначе в поддиагональных позициях первого столбца снова появятся ненулевые элементы. Отсюда следует, что если $\tilde{P}$ — пронзвольная ортогональная матрица, для которой матрица

$$\tilde{P}\hat{M}=\tilde{P}\hat{\mathscr{H}}_1 M=R$$

будет верхней треугольной, то *первые строки* матриц $\tilde{P}\hat{\mathscr{H}}_1$ и $\hat{\mathscr{H}}_1$ должны совпадать. Но $\hat{\mathscr{H}}_1$ — симметричная матрица, а $\tilde{P}\hat{\mathscr{H}}_1$ транспонирована по отношению к матрице Q в ортогонально-треугольном разложении матрицы M. Конечный вывод таков: *первый столбец искомой матрицы Q совпадает с первым столбцом отражения $\hat{\mathscr{H}}_1$, описываемого формулами* (11.28), (11.29).

Выяснив вид первого столбца матрицы Q, построим ортогональную матрицу P с таким же первым столбцом, которая преобразовывала бы H_1 в хессенбергову матрицу. Для начала выполним подобие с трансформирующей матрицей $\hat{\mathscr{H}}_1$: $H_1 \to G=\hat{\mathscr{H}}_1 H_1 \hat{\mathscr{H}}_1$. Вид матрицы G для $n=7$ показан на первом фрагменте рис. 5. Она отличается от хессенберговой матрицы элементами, указанными крестиками с подчеркиванием.

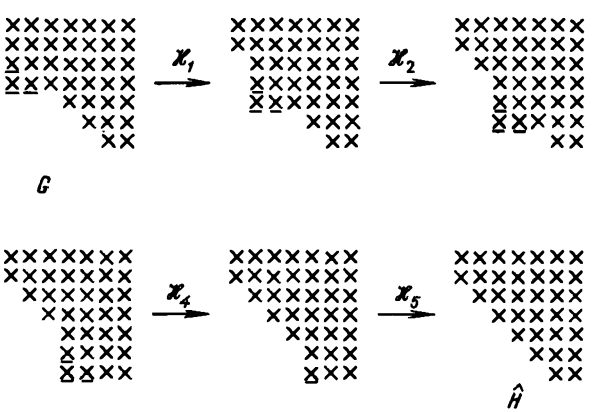

Рис. 5

Теперь приведем матрицу G к форме Хессенберга с помощью метода отражений:

$$\hat{H}=\mathscr{H}_{n-2}\ldots\mathscr{H}_2\mathscr{H}_1 G\mathscr{H}_1\mathscr{H}_2\ldots\mathscr{H}_{n-2}. \tag{11.30}$$

Рисунок иллюстрирует процесс приведения. Из него видно, что на каждом шаге нужно аннулировать лишь два поддиагональных

элемента (а на последнем шаге — только один); следовательно, в порождающих векторах u_i не более трех ненулевых компонент и преобразованиям данного шага подвергаются три строки и три столбца матрицы. Это означает, что вычисление матрицы $\hat{H}$ потребует $O(n^2)$ арифметических операций, а не $O(n^3)$, как было бы в случае матрицы общего вида. Точный подсчет показывает, что вычислительная работа при переходе от H_1 к $\hat{H}$ лишь немногим превосходит работу при выполнении *одного* вещественного QR-шага.

Остается показать, что $\hat{H}$ и есть искомая матрица H_3. Подставляя в (11.30) выражение для G, имеем

$$\hat{H}=(\mathscr{H}_{n-2}\dots\mathscr{H}_2\mathscr{H}_1\hat{\mathscr{H}}_1)H_1(\hat{\mathscr{H}}_1\mathscr{H}_1\mathscr{H}_2\dots\mathscr{H}_{n-2})=P^{\mathrm{T}}H_1P. \qquad (11.31)$$

Матрица $P=\hat{\mathscr{H}}_1\mathscr{H}_1\mathscr{H}_2\dots\mathscr{H}_{n-2}$ ортогональная, и, так как все матрицы $\mathscr{H}_i$ метода отражений имеют единичный первый столбец, у P первый столбец совпадает с первым столбцом $\hat{\mathscr{H}}_1$, т. е. в конечном счете с первым столбцом матрицы Q. При этом P, как и Q, преобразует H_1 в хессенбергову матрицу $\hat{H}$. Поскольку матрица H_3 по предположению неразложима, то, учитывая замечание к теореме единственности, заключаем: матрицы $\hat{H}$ и H_3 могут различаться разве лишь знаками внедиагональных элементов.

Подводя итог, можно сказать, что в случае вещественной матрицы H итерационный QR-процесс, основанный на стратегии Уилкинсона, представляет собой последовательность двойных QR-шагов. Двойные шаги выполняются независимо от того, вещественны или комплексны собственные значения очередной подматрицы (11.19). Действительно, в приведенном выше описании комплексность собственных значений нигде не проверялась и не использовалась.

В случае комплексной матрицы A имеются два способа применения QR-подпрограмм. Первый способ состоит в замене комплексной спектральной задачи вещественной. Он опирается на следующий хорошо известный факт. Если $\lambda_1, \dots, \lambda_n$ суть собственные значения комплексной матрицы $A=F+iG$, где F и G — вещественные матрицы, то числа $\lambda_1, \dots, \lambda_n, \bar{\lambda}_1, \dots, \bar{\lambda}_n$ составляют спектр вещественной матрицы удвоенного порядка

$$A_R=\begin{bmatrix} F & -G \\ G & F \end{bmatrix}.$$

С каждым собственным вектором z матрицы A ассоциироваио двумерное инвариантное подпространство матрицы A_R. В самом деле, представим z в виде $z=x+iy$, где $x, y\in\mathbf{R}^n$, и пусть $\lambda=\mu+i\nu$ — соответствующее собственное значение. Комплексное равенство $Az=\lambda z$ эквивалентно каждой из систем вещественных равенств:

$$Fx-Gy=\mu x-\nu y, \qquad F(-y)-Gx=\mu(-y)-\nu x,$$
$$Gx+Fy=\nu x+\mu y; \qquad G(-y)+Fx=\nu(-y)+\mu x,$$

означающих, что подпространство, натянутое на векторы

$$z_R^1=\begin{bmatrix} x \\ y \end{bmatrix}, \qquad z_R^2=\begin{bmatrix} -y \\ x \end{bmatrix},$$

инвариантно относительно матрицы A_R. Нетрудно проверить, что векторы z_R^1 и z_R^2 линейно независимы.

Итак, первый способ вычисления собственных значений и векторов комплексной матрицы A — это вычисление посредством программ вещественного QR-алгоритма собственных значений и векторов (или инвариантных подпространств) вещественной матрицы A_R. Из них несложно восстановить спектральную информацию, относящуюся к A. Недостаток этого подхода — удвоение требований к памяти: хранение комплексной $n \times n$-матрицы A обходится в $2n^2$ машинных слов, а вещественной матрицы A_R порядка $2n$ — в $4n^2$ слов. Увеличится и вычислительная работа: бóльшая скорость вещественной арифметики не компенсирует удвоения порядка — ведь число арифметических операций в QR-алгоритме пропорционально n^3.

Таким образом, более оправданным является второй способ — использование комплексного QR-алгоритма. Здесь исчезает основная мотивация техники двойных шагов: незачем избегать комплексной арифметики. Вероятно, по этой причине стратегия Уилкинсона в программах QR-алгоритма для комплексных матриц реализуется в форме одинарных шагов (11.16)—(11.17). Матрица (11.19) по-прежнему участвует в выборе сдвига (что и характеризует стратегию); в качестве τ_k берется то ее собственное значение, которое расположено ближе к диагональному элементу $h_{nn}^{(k)}$; более удаленное собственное значение не используется совсем. В этой форме QR-алгоритм воплощен в подпрограммах COMQR, COMQR2 пакета EISPACK.

Сдвиги в качестве средства ускорения QR-процесса были введены нами как приближения к собственным значениям матрицы H. Однако в QR-программах сдвиги по Уилкинсону начинают применяться с первой же итерации, когда нет никаких оснований считать их близкими к собственным значениям. Тем не менее, как правило, стратегия Уилкинсона работает успешно, хотя полноценного объяснения этому факту нет до сих пор (исключением является случай вещественной симметричной трехдиагональной матрицы H, для которого показано, что произведение $|h_{n-1,n-2}^{(k)}|(h_{n,n-1}^{(k)})^2$ в стратегии Уилкинсона монотонно убывает с ростом k; см. [35, § 8.10]). Успешность QR-процесса — это сходимость к нулю поддиагональных элементов последовательных хессенберговых матриц: в первую очередь элемента в позиции $(n, n-1)$ — в комплексном QR-алгоритме и наряду с ним или вместо него элемента в позиции $(n-1, n-2)$ — в вещественном QR-алгоритме. Когда достаточная степень малости поддиагонального элемента достигнута, его заменяют нулем. В результате матрица приобретает блочно-диагональный вид:

$$H^{(k)} = \begin{bmatrix} H_{11}^{(k)} & H_{12}^{(k)} \\ 0 & H_{22}^{(k)} \end{bmatrix}, \tag{11.32}$$

и в дальнейшем можно вычислять собственные значения подматрицы $H_{22}^{(k)}$, а затем подматрицы $H_{11}^{(k)}$. В типичном случае порядок матрицы $H_{22}^{(k)}$ равен 1 либо 2, и ее собственные значения определяются тривиально. Однако возможны и бóльшие значения порядка; тогда,

начиная с k-й итерации, QR-процесс ведется с $H_{22}^{(k)}$. Это в свою очередь может привести к расщеплению, и последующие итерации выполняются вначале с самой нижней из образовавшихся подматриц. Лишь после вычисления всех ее собственных значений процесс возвращается к более «высокопоставленным» подматрицам.

Какой же поддиагональный элемент можно заменить нулем? Обычно руководствуются таким рассуждением. Каждый шаг QR-процесса вносит в матрицу ошибки, суммарно измеряемые величиной порядка $O(n\|A\|\beta^{-t})$. Поэтому, устраняя поддиагональный элемент с модулем, не превосходящим $\|A\|\beta^{-t}$, мы как бы совершаем дополнительную ошибку, соизмеримую с ошибками, уже присутствующими в элементах матрицы. Однако в программах QR-алгоритма из справочника [43] и пакета EISPACK принят более осторожный критерий: поддиагональный элемент $h_{i+1,i}^{(k)}$ аннулируется, если

$$|h_{i+1,i}^{(k)}| \leqslant \left(|h_{ii}^{(k)}| + |h_{i+1,i+1}^{(k)}|\right)\beta^{-t}. \tag{11.33}$$

Таким образом, важна малость элемента $h_{i+1,i}^{(k)}$ относительно ближайшего окружения, а не только относительно общего уровня элементов матрицы.

Различие между двумя критериями проиллюстрируем с помощью матрицы [43, с. 325]

$$H = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 10^{-3} & 3\times 10^{-3} & 2\times 10^{-3} & 10^{-3} \\ 0 & 10^{-6} & 3\times 10^{-6} & 2\times 10^{-6} \\ 0 & 0 & 10^{-9} & 2\times 10^{-9} \end{bmatrix}.$$

Если $t=24$, то для элемента h_{43} выполнено неравенство

$$|h_{43}| < \|H\| \cdot 2^{-t},$$

но не выполнено (11.33). Если заменить этот элемент нулем, диагональный элемент 2×10^{-9} следует считать собственным значением. Между тем собственные значения матрицы H равны $1.002\ldots$, $9.9699\ldots \times 10^{-4}$, $4.005\ldots \times 10^{-6}$, $7.497\ldots \times 10^{-10}$. Среди них нет ни одного значения, имеющего порядок 10^{-9}. Если рассматривать 2×10^{-9} как возмущение собственного значения $7.497\ldots \times 10^{-10}$, то абсолютная погрешность возмущения невелика, а относительная превышает 1!

В программах QR-алгоритма, как правило, используется еще один любопытный прием, также придуманный Фрэнсисом. Мы изложим его, следуя [42, гл. 8, § 38—40].

Рассматриваемый прием относится к ситуации, когда ни один из поддиагональных элементов не удовлетворяет тесту (11.33), зато имеется пара подряд идущих элементов $h_{r+1,r}$, $h_{r+2,r+1}$ с достаточно малым произведением. Положим $h_{r+1,r}=\varepsilon_1$, $h_{r+2,r+1}=\varepsilon_2$ и представим матрицу H_k в блочном виде, опуская для простоты индекс k:

$$H = \begin{bmatrix} X & Y \\ E & W \end{bmatrix}. \tag{11.34}$$

Разбиение на блоки проиллюстрируем для случая $n=6$, $r=2$:

$$\left[\begin{array}{cc|cccc} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \hline & \varepsilon_1 & \times & \times & \times & \times \\ & & \varepsilon_2 & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \times & \times \end{array}\right]. \tag{11.35}$$

Оказывается, что всякое собственное значение μ подматрицы W будет в то же время собственным значением матрицы $H'(\mu)$, отличающейся от H единственным элементом, а именно элементом в позиции $(r+2,\ r)$ со значением $\varepsilon_1\varepsilon_2/(h_{r+1,r+1}-\mu)$. В самом деле, пусть $\bar{v}=(\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_{n-r})^T$ — левый собственный вектор матрицы W для собственного значения μ, т. е.

$$v^T(W-\mu I)=0. \tag{11.36}$$

Тогда, в частности,

$$(h_{r+1,r+1}-\mu)v_1+\varepsilon_2 v_2=0,$$

откуда

$$\varepsilon_1 v_1+[\varepsilon_1\varepsilon_2/(h_{r+1,r+1}-\mu)]v_2=0. \tag{11.37}$$

Соотношения (11.36) и (11.37) вместе означают, что в матрице $H'(\mu)-\mu I$ линейная комбинация последних $n-r$ строк с коэффициентами соответственно $v_1, v_2, \ldots, v_{n-r}$ будет нулевой строкой. Следовательно, матрица $H'(\mu)-\mu I$ вырождена, и μ есть собственное значение матрицы $H'(\mu)$.

Если величина $\gamma(r, \mu)\equiv\varepsilon_1\varepsilon_2/(h_{r+1,r+1}-\mu)$ удовлетворяет неравенству

$$|\gamma(r, \mu)|\leqslant\|A\|\beta^{-t} \tag{11.38}$$

или более строгому условию типа

$$|\gamma(r, \mu)|\leqslant(|h_{rr}|+|h_{r+1,r+1}|+|h_{r+2,r+2}|)\beta^{-t}, \tag{11.39}$$

то различием между H и $H'(\mu)$ можно пренебречь и считать, что μ — собственное значение подматрицы W — есть в то же время собственное значение самой матрицы H.

В реальном QR-процессе значение μ не известно. Более того, не известно и r. В общем случае все или многие поддиагональные элементы матрицы сходятся к нулю (разве что с разной скоростью), а потому может быть несколько пар соседних элементов с малым произведением. Но одной малости произведения недостаточно: необходимо еще иметь представление о величине разности $h_{r+1,r+1}-\mu$. Поэтому программа QR-алгоритма определяет значение r, проверяя неравенство (11.38) или (11.39) (или, как, например, в программе вещественного QR-алгоритма из [43], несколько более сложное неравенство того же типа) сначала для элементов $h_{n,n-1}$ и $h_{n-1,n-2}$,

затем для элементов $h_{n-1,n-2}$ и $h_{n-2,n-3}$ и т. д., продвигаясь вверх, причем в качестве приближения к μ берется сдвиг Уилкинсона, найденный из подматрицы (11.19).

Пусть r найдено и тем самым выявлено представление (11.34). Очередной QR-шаг проводится так, как если бы вычислялось собственное значение подматрицы W. Если для определенности говорить об одинарном шаге, это означает, что при помощи вращений матрицу $W-\tau I$ приводят к треугольному виду, а затем выполняют правосторонние вращения и восстановление сдвига. Однако эти преобразования, определяемые только матрицей W, применяют к строкам и столбцам *всей матрицы* H. Так как на первом полушаге обработке подвергаются лишь строки $r+1, \ldots, n$, то единственное изменение по сравнению с разложением матрицы $W-\tau I$ состоит в необходимости пересчитать значение ε_1 элемента $(r+1, r)$. Правда, появится еще ненулевой элемент в позиции $(r+2, r)$, но, как наперед известно из теста (11.38), этим элементом можно пренебречь, сохранив тем самым форму (11.35) (в случае двойного QR-шага возникнет также ненулевой элемент в позиции $(r+3, r)$, но и тест типа (11.39) в программе сдваивания ориентирован на возможность отбрасывания обоих появившихся элементов). Правосторонние вращения затрагивают лишь столбцы $r+1, \ldots, n$ матрицы H. Из сказанного ясно, что прием Фрэнсиса дает значительную экономию арифметической работы, если порядок подматрицы W много меньше порядка H.

Отметим, что в библиотечных программах алгоритма проверка матрицы на возможность расщепления (тест типа (11.33)), а затем поиск пары малых поддиагональных соседей выполняются перед каждым QR-шагом, одинарным или двойным.

Уже говорилось, что QR-алгоритм сходится не всегда. Имеются даже матрицы, которые метод не меняет. Так, матрица

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{11.40}$$

инвариантна относительно как основного QR-алгоритма, так и стратегии Уилкинсона: оба сдвига, определяемые нижней угловой подматрицей, равны нулю, и метод со сдвигами не отличается от основного алгоритма. Заметим, однако, что почти любой ненулевой сдвиг выводит матрицу из стационарного состояния, после чего процесс быстро сходится. Примеры подобного типа побудили составителей библиотечных подпрограмм включить в алгоритм ограничения на число итераций. Так, в процедуре hqr из справочника [43] после десяти подряд итераций, в течение которых не выделилось ни собственного значения, ни блока второго порядка, производится двойной шаг со сдвигами τ_1, τ_2, определяемыми формулами

$$\tau_1+\tau_2 = 1.5\left(|h^{(k)}_{n,n-1}|+|h^{(k)}_{n-1,n-2}|\right),$$
$$\tau_1\tau_2 = \left(|h^{(k)}_{n,n-1}|+|h^{(k)}_{n-1,n-2}|\right)^2.$$

Вместо коэффициента 1.5 могло бы стоять почти любое другое число. Если после вспомогательного шага следующие десять (обычных двойных) итераций снова безуспешны, то вспомогательный шаг повторяется. Если и на этот раз очередные десять итераций не приносят результата, совершается выход из процедуры.

Повторим, однако, что несходимость QR-итераций наблюдается чрезвычайно редко. Напротив, для программ, реализующих стратегии Рэлея и Уилкинсона, характерна быстрая сходимость; принято считать, что в среднем на вычисление собственного значения затрачиваются две итерации. Типичный ход QR-процесса демонстрирует следующий пример, заимствованный из [116, с. 236]. Для вычисления собственных значений матрицы

$$A = H = \begin{bmatrix} 2 & 3 & 4 & 5 & 6 \\ 4 & 4 & 5 & 6 & 7 \\ 0 & 3 & 6 & 7 & 8 \\ 0 & 0 & 2 & 8 & 9 \\ 0 & 0 & 0 & 1 & 10 \end{bmatrix}$$

понадобилось 12 итераций. Поведение поддиагональных элементов видно из табл. 5. После семи итераций $h_{43}^{(7)}$ и $h_{54}^{(7)}$ признаны пренебрежимо малыми, а $h_{44}^{(7)}$ и $h_{55}^{(7)}$ — приближенными собственными значениями. Дальнейшие итерации идут с верхней 3×3-подматрицей. Элемент $h_{32}^{(k)}$ не сходится к нулю, но после пяти шагов выделяется диагональный блок порядка 2. Дополнительный к этому блоку элемент $h_{11}^{(12)}$ также принимается за приближенное собственное значение, и процесс заканчивается.

Таблица 5

Номер итерации	$O(\lvert h_{21}\rvert)$	$O(\lvert h_{32}\rvert)$	$O(\lvert h_{43}\rvert)$	$O(\lvert h_{54}\rvert)$
1	10^0	10^0	10^0	10^0
2	10^0	10^0	10^0	10^0
3	10^0	10^0	10^{-1}	10^0
4	10^0	10^0	10^{-3}	10^{-3}
5	10^0	10^0	10^{-6}	10^{-5}
6	10^{-1}	10^0	10^{-13}	10^{-13}
7	10^{-1}	10^0	10^{-28}	10^{-13}
8	10^{-4}	10^0	Условие выхода выполнено	Условие выхода выполнено
9	10^{-8}	10^0		
10	10^{-8}	10^0		
11	10^{-16}	10^0		
12	10^{-32}	10^0		
	Условие выхода выполнено			

Оценивая эффективность QR-алгоритма, нужно и в общем случае учитывать, что реальная вычислительная работа не пропорциональна числу итераций: ведь выделение каждого собственного значения или диагонального 2×2-блока приводит к уменьшению порядка об-

рабатываемой матрицы, а вместе с тем и числа операций, затрачиваемых на QR-шаг.

Итерационный QR-процесс численно устойчив в том смысле, как это определено в § 2: вычисляемые собственные значения будут точными для матрицы, близкой к исходной матрице H. Основа устойчивости и QR-процесса, и метода отражений одна и та же: используемые в них преобразования являются ортогональными и уровень элементов матрицы на любом этапе обработки остается неизменным.

Теперь обсудим, какие изменения вносит в QR-процесс необходимость вычислять собственные векторы матрицы A. Если есть возможность хранить начальную хессенбергову матрицу H QR-итераций, то в самом QR-процессе ничего менять не нужно. Закончив вычисление собственных значений, мы можем по ним методом обратных итераций определить собственные векторы матрицы H. На этапе 4 QR-алгоритма они будут преобразованы в собственные векторы матрицы A.

Если хранение матрицы H невозможно, приходится действовать иначе. Теоретически матрицы H_k в QR-процессе подобны матрице H. Однако после первого же расщепления внедиагональный блок $H_{12}^{(4)}$ (см. (11.32)) перестает учитываться, и о подобии в дальнейшем говорить не приходится. Между тем если бы мы могли рассматривать QR-алгоритм как метод унитарно (ортогонально) подобного приведения матрицы H к матрице T верхнего треугольного или блочно треугольного вида

$$T = S^* HS, \tag{11.41}$$

то вычисление собственных векторов для H свелось бы к вычислению собственных векторов для T (что тривиально достигается посредством обратной подстановки или ее блочного обобщения) и их преобразованию посредством матрицы S. Отсюда вытекают два вывода: во-первых, преобразования, выполняемые в QR-процессе, должны быть действительно подобиями, т. е. они должны применяться и к внедиагональным блокам; во-вторых, должна быть накоплена матрица S произведения всех этих преобразований. Чтобы не заводить для S отдельный массив, вспомним, что в случае вычисления собственных векторов между этапами 2 и 3 QR-алгоритма может быть сформировано в явном виде произведение P отражений $\mathscr{H}_i$, построенных при приведении уравновешенной матрицы B к форме Хессенберга, т. е. трансформирующая матрица преобразования $H = P^* BP$. Подставляя это выражение для H в (11.41), получаем

$$T = (PS)^* B(PS).$$

Тем самым матрица PS трансформирует собственные векторы матрицы T в собственные векторы матрицы B. Поскольку собственные векторы для H сами по себе нам не нужны, мы можем отказаться от их вычисления и вместо хранения матрицы P накапливать произведение PS. Это значит, что на каждом QR-шаге к массиву, первоначально — на входе в итерационный процесс — хранившему

матрицу P, применяются все правосторонние преобразования этого шага.

Итак, на выходе модифицированного указанным образом QR-процесса мы получаем треугольную или блочно-треугольную матрицу T и унитарную (ортогональную) матрицу PS. После этого, как уже сказано, вычисляются обратной подстановкой собственные векторы для T.

4. Преобразование собственных векторов матрицы $H(T)$ в собственные векторы матрицы A. Если хессенбергова матрица H сохранена, то, после вычисления ее собственных векторов, каждый из них преобразуется по формуле

$$y = Pz = \mathscr{H}_1 \mathscr{H}_2 \dots \mathscr{H}_{n-2} z$$

в собственный вектор y матрицы B. Для построения этого произведения матрица P в явном виде не нужна: вектор z последовательно умножается на отражения $\mathscr{H}_{n-2}, \dots, \mathscr{H}_2, \mathscr{H}_1$, порождающие векторы которых сохранены подпрограммой этапа 2.

Если матрица H не хранится, то вычисленные собственные векторы треугольной (или блочно-треугольной) матрицы T умножением на матрицу PS, построенную в QR-процессе, переводятся в собственные векторы матрицы B.

И в том, и в другом случае из собственных векторов матрицы B получаются по формуле $x = Dy$ собственные векторы исходной матрицы A. Так как матрица D диагональная и ее диагональные элементы суть целые степени числа β — основания машинной арифметики, то умножение на D сводится к изменению порядков компонент вектора y.

Нужно отметить, что собственные векторы верхней треугольной матрицы T можно расположить на месте, занимаемом самой этой матрицей, если вычисление их вести в обратном порядке: от вектора, отвечающего собственному значению t_{nn}, до вектора, ассоциированного с t_{11}. То же справедливо в отношении блочно-треугольной матрицы T, если вместо каждой пары сопряженных комплексных собственных векторов $w = u \pm iv$ хранить соответствующую пару вещественных векторов u, v.

5. Переход к форме Шура с заданным расположением собственных значений на главной диагонали. Пусть T — форма Шура, вычисленная на этапе 3. В комплексном случае это верхняя треугольная матрица, в вещественном — блочно-треугольная матрица с диагональными блоками порядка 1 либо 2. При этом 2×2-блоки обычно соответствуют парам комплексно-сопряженных собственных значений.

Если расположение собственных значений на диагонали важно для пользователя и в полученной матрице T оно отличается от желаемого, то можно сравнительно дешевым способом переупорядочить диагональные элементы или блоки, не теряя блочно-треугольной формы. Этот процесс переупорядочения состоит из последовательности шагов, каждый из которых меняет местами два рядом стоящих элемента (блока). Понятно, что, умея выполнять такие *транспозиции*, мы сможем перевести любой элемент или блок в указанную позицию главной диагонали.

Рассмотрим вначале наиболее простой случай — транспозицию двух диагональных элементов t_{ii} и $t_{i+1,i+1}$. Если переставить строки

и столбцы с номерами i и $i+1$, то стоящий на их пересечении диагональный блок изменится так:

$$\begin{bmatrix} t_{ii} & t_{i,i+1} \\ 0 & t_{i+1,i+1} \end{bmatrix} \to \begin{bmatrix} t_{i+1,i+1} & 0 \\ t_{i,i+1} & t_{ii} \end{bmatrix}. \tag{11.42}$$

Понятно, что, за исключением позиции $(i+1,\ i)$, форма Шура не нарушается. При $t_{i,i+1}=0$ перестановка завершена. Если же $t_{i,i+1}\neq 0$, то подберем элементарную унитарную матрицу $\mathscr{R}_{i,i+1}$ так, чтобы аннулировать элемент $(i+1,\ i)$ в матрице $\tilde{T}=\mathscr{R}^*_{i,i+1}\mathscr{P}_{i,i+1}T\mathscr{P}_{i,i+1}\mathscr{R}_{i,i+1}$. Принцип подбора одинаков и в вещественном, и в комплексном случае, но формулы проще для вещественного. Если c и s — параметры вращения, то для элементов $\tilde{t}_{ii}$ и $\tilde{t}_{i+1,i}$ справедливы выражения

$$\tilde{t}_{ii}=t_{i+1,i+1}+\left[t_{i,i+1}\,c+(t_{ii}-t_{i+1,i+1})\,s\right]s,$$

$$\tilde{t}_{i+1,i}=\left[t_{i,i+1}\,c+(t_{ii}-t_{i+1,i+1})\,s\right]c.$$

Всегда можно найти c и s, обращающие в нуль содержимое квадратных скобок. Тогда элемент $\tilde{t}_{i+1,i}$ будет аннулирован, а элемент в позиции $(i,\ i)$ сохранит свое значение $t_{i+1,i+1}$. Не изменится и элемент $(i+1,\ i+1)$.

Для комплексной матрицы T при построении $\mathscr{R}_{i,i+1}$ нужно дополнительно задать подходящие значения аргументов φ_1, φ_2, φ_3, φ_4 (см. (2.18)).

Разберем теперь процедуру транспозиции 2×2-блоков в вещественной форме Шура. Если вместо обычных говорить о блочных строках и столбцах, то начинаем так же, как в предыдущем случае. По аналогии с (11.42) имеем

$$\begin{bmatrix} T_{ii} & T_{i,i+1} \\ 0 & T_{i+1,i+1} \end{bmatrix} \to \begin{bmatrix} T_{i+1,i+1} & 0 \\ T_{i,i+1} & T_{ii} \end{bmatrix} \equiv T_4. \tag{11.43}$$

Других отклонений от формы Шура не будет.

Приведем 4×4-матрицу T_4 к форме Хессенберга, а затем выполним для полученной хессенберговой матрицы двойной QR-шаг, взяв в качестве сдвигов собственные значения λ_1 и λ_2 блока T_{ii}. В результате поддиагональный блок станет нулевым, а нижний диагональный блок сохранит собственные значения λ_1, λ_2.

Разумеется, указанные преобразования необходимо применять ко всей матрице T. По их завершении форма Шура будет восстановлена.

Таким же образом производится перестановка 1×1- и 2×2-блоков. В зависимости от их порядка для 3×3-матрицы, аналогичной (11.43), совершается либо двойной, либо одинарный QR-шаг со сдвигом. В последнем случае в качестве сдвига берется значение диагонального элемента.

ДОПОЛНЕНИЯ К § 11.

1. Построение формы Шура в QR-алгоритме может быть оправданным и в том случае, если собственные векторы сами по себе не нужны, зато желательно получить оценки погрешностей в вычисленных собственных значениях. Приближенное собственное значение $\tilde{\lambda}$, найденное QR-алгоритмом,

является точным собственным значением для возмущения $A+F$ исходной матрицы A. При этом оценка нормы матрицы F известна из развитой Уилкинсоном теории обратного аиализа ошибок. Если собствеиое значение λ матрицы A, приближением которого служит $\tilde{\lambda}$, простое, то оценкой для ошибки $|\lambda-\tilde{\lambda}|$ с точностью до величин 2-го порядка может быть произведение $k(\lambda)\|F\|$ (см. § 6). Отсюда следует, что для *оценивания ошибки* в $\tilde{\lambda}$ вычисление соответствующего собственного вектора (причем не только правого, но и левого) все же необходимо. Другое дело, что эти векторы незачем хранить после того, как $k(\lambda)$ найдено. Исходя из сказанного, действуют следующим образом. Прежде всего используют такую подпрограмму QR-алгоритма, которая наряду с вычислением собственных значений матрицы строит ее форму Шура T. От подпрограммы, применяемой при вычислении собственных векторов, данная отличается тем, что не хранит произведение унитарных матриц. Число обусловленности

$$k(\lambda)=\|x\|_2\,\|y\|_2/|(x, y)|$$

не меняется при унитарном преобразовании матрицы: если $B=Q^*AQ$, то собственные векторы x и y матрицы A перейдут в собственные векторы Q^*x и Q^*y матрицы B. Поэтому на заключительном этапе для каждого простого собственного значения λ определяются правый и левый собственные векторы треугольной (или блочно-треугольной) матрицы T, а по ним число $k(\lambda)$. Эта методика оценивания ошибок предложена в работе [77]; там же приведен (с некоторыми купюрами) текст фортраиной подпрограммы, ее воплощающей. Подпрограмма предназначена для обработки вещественных матриц и использует вещественную арифметику (даже тогда, когда матрица имеет комплексные собственные значения). Для матриц порядка от 20 до 60 процедура оценивания требует в среднем 50% времени, затрачиваемого на вычисление собственных чисел.

Другой способ оценивания погрешностей в собственных значениях предложен в [196]. Он рассчитан на случай, когда форма Шура не строится, но зато хранится форма Хессенберга. С помощью последней вычисляются нужные собственные векторы, а также некоторая численная характеристика обусловленности.

2. В [188] дано описание подпрограммы, вычисляющей для вещественной хессенберговой матрицы H ее вещественную форму Шура, в которой диагональные элементы и блоки 2×2 упорядочены по убыванию модулей собственных значений; вычисляется и ортогональная трансформирующая матрица. В подпрограмме используется техника перестановки диагональных блоков, близкая к рассмотренной в основном тексте (см. этап 5). Отличия состоят в следующем. Во-первых, до выполнения двойного шага со сдвигами τ_1, τ_2, почерпнутыми из верхнего блока, производится предварительный шаг с произвольными сдвигами. Его цель — вывести подматрицу, натянутую на переставляемые блоки, из (нижней) блочно-треугольной формы, с тем чтобы получить неразложимую хессенбергову матрицу. Во-вторых, если применение двойного шага со сдвигами τ_1, τ_2 не привело к выделению нужного блока в нижнем правом углу, этот шаг повторяется, возможно, неоднократно. Наконец, на случай очень плохо обусловленных собственных значений установлено ограничение на число двойных шагов. Если оно превышено, происходит аварийный выход из подпрограммы.

Упорядочение по убыванию модулей используется в методах одновременной итерации (см. гл. 5) при вычислении группы старших собственных значений и соответствующих собственных векторов. Подпрограмма в [188] и была составлена как вспомогательный модуль программы одновремеиной итерации, реализующей алгоритм Стьюарта из § 14. Некоторые опечатки в тексте подпрограммы указаны в [108].

3. В связи с существованием матриц, инвариантных относительно основного QR-алгоритма (см. (11.40)), в [155] поставлена и решена задача описания всех таких матриц.

Теорема (Парлетт). *Неразложимая хессенбергова матрица H тогда и только тогда инвариантна относительно основного* QR*-алгоритма, когда она с точностью до скалярного множителя унитарная.*

Доказательство. Пусть $H=\alpha U$, где U — унитарная матрица; число α без ограничения общности можно считать положительным. Унитарно-треугольное разложение матрицы H имеет вид $H=U(\alpha I)$. Второй полушаг QR-итерации снова дает H.

Предположим теперь, что H инвариантна относительно основного QR-алгоритма; тогда $H=QR=RQ$. Отсюда получаем

$$HQ=QRQ=QH, \tag{11.44}$$

т. е. H и Q перестановочны. Поскольку H — неразложимая хессенбергова матрица, в ее жордановой форме каждому собственному значению λ отвечает ровно одна жорданова клетка. Другими словами, H — недефектная матрица. Действительно, ранг матрицы $H-\lambda I$ равен $n-1$, как показывает рассмотрение подматрицы, получаемой вычеркиванием первой строки и последнего столбца, а потому для λ имеется лишь один (с точностью до пропорциональности) собственный вектор. Установленное свойство матрицы H обеспечивает, что всякая перестановочная с ней матрица Q представима в виде многочлена от H (см. [12, гл. VIII, § 2]). Итак, $Q=\Phi(H)$, и степень m многочлена Φ в силу теоремы Гамильтона — Кэли можно считать не превосходящей $n-1$. Нетрудно проверить, что элемент матрицы $\Phi(H)$ в позиции $(m+1, 1)$ равен произведению $\varphi_m h_{21} h_{32} \dots h_{m+1,m}$, где φ_m — коэффициент при старшем члене многочлена Φ; этот элемент не равен нулю, так как $h_{i+1,i}\neq 0$ при всех i. В то же время матрица Q, которую можно представлять себе как произведение элементарных унитарных матриц $\mathscr{R}_{12}^{*}\mathscr{R}_{23}^{*}\dots\mathscr{R}_{n-1,n}^{*}$, будет, как и H, верхней хессенберговой матрицей. Из равенства $Q=\Phi(H)$ теперь следует, что $m\leqslant 1$, и многочлен $\Phi(x)$ можно записать в виде $\Phi(x)=\varphi_0+\varphi_1 x$. Сравнивая элементы первого столбца в матричном равенстве $H=(\varphi_0 I+\varphi_1 H)R$, получаем $r_{11}^{-1}h_{11}=\varphi_0+\varphi_1 h_{11}$, $r_{11}^{-1}h_{21}=\varphi_1 h_{21}$, откуда $\varphi_0=0$, $\varphi_1=r_{11}^{-1}$. Итак, $H=r_{11}Q$, что и требовалось доказать.

Там же, в [155], дано описание матриц, инвариантных относительно двойного шага. Положим $\sigma=h_{n-1,n-1}+h_{nn}$, $\rho=h_{n-1,n-1}h_{nn}-h_{n,n-1}h_{n-1,n}$.

Теорема (Парлетт). *Неразложимая хессенбергова матрица H тогда и только тогда инвариантна относительно двойного* QR*-шага, когда матрица $H^2-\sigma H+\rho I$ с точностью до скалярного множителя унитарная.*

§ 12*. О сходимости QR-алгоритма

В этом параграфе будет дан обзор известных фактов относительно сходимости QR-алгоритма. Для основного варианта метода, описываемого формулами (4.0.1) — (4.0.2), создана*), по существу, полная теория, главные положения которой излагаются в начале параграфа. Однако библиотечные подпрограммы QR-алгоритма всегда реализуют ту или иную стратегию сдвигов. Это приводит на практике к ускорению сходимости, но теория метода значительно усложняется. В результате мы до сих пор не имеем доказательства глобальной сходимости

*) В первую очередь усилиями В. Н. Кублановской, В. В. Воеводина, Дж. Уилкинсона и Б. Парлетта.

в случае матриц общего вида для таких наиболее часто используемых стратегий, как сдвиги по Рэлею или Уилкинсону. Правда, для некоторых, хотя и специальных, но важных матричных классов, а именно эрмитовых и унитарных матриц, глобальная сходимость этих стратегий — в чистом или комбинированном варианте — установлена, равно как и ее асимптотическая скорость. Приведя соответствующие утверждения, мы заканчиваем параграф обсуждением связей QR-алгоритма со степенным методом и обратными итерациями. Не будучи строгим доказательством, эти связи все же помогают понять, почему QR-алгоритм сходится и в случае произвольных матриц.

Сделаем два предварительных замечания. Прежде всего нам будет удобно считать, что в QR-разложении на первом полушаге QR-итерации матрица R имеет неотрицательные диагональные элементы. Это требование мы относим ко всем вариантам QR-алгоритма, и общности оно нисколько не ограничивает. Более того, при QR-разложении хессенберговой матрицы посредством вращений (или элементарных унитарных матриц) положительность всех, кроме, может быть, последнего, диагональных элементов r_{ii} обеспечивается автоматически. Если при этом оказалось, что условие $r_{nn} \geqslant 0$ не выполнено, то от первоначально полученного разложения $A = QR$ достаточно перейти к разложению $A = (QD^*)(DR)$, где $D = \operatorname{diag}(1, \ldots, 1, \exp\{-i \arg r_{nn}\})$. Отметим, что QR-разложение с положительными r_{ii} $(i = 1, \ldots, n)$ для невырожденной матрицы A определено единственным образом.

Второе замечание относится к толкованию термина «сходимость». Обычной поэлементной сходимости QR-алгоритм не обеспечивает. Пусть, например, A — верхняя треугольная матрица, среди диагональных элементов которой есть неположительные. Хотя A уже имеет форму Шура и самым естественным было бы не менять ее, так в действительности не будет. Согласно сказанному выше, QR-разложение для A описывается равенством $A = (D^*)(DA)$, где диагональная матрица D учитывает аргументы элементов a_{ii} $(1 \leqslant i \leqslant n)$. Завершая QR-шаг, мы получим матрицу $B = DAD^*$, отличающуюся от A аргументами *внедиагональных* элементов.

Таким образом, сходимость QR-алгоритма — это *сходимость по форме* к треугольной или блочно-треугольной матрице. Она характеризуется сходимостью к нулю поддиагональных элементов или элементов поддиагональных блоков, тогда как наддиагональные элементы от итерации к итерации могут изменяться.

Основной QR-*алгоритм* мы рассмотрим вначале в применении к нормальным матрицам. Для этого случая в [97] дано изящное доказательство глобальной сходимости к блочно-диагональной форме, которое сейчас будет изложено.

Поскольку все матрицы A_k, генерируемые QR-алгоритмом, унитарно подобны (см. (11.18)), то при нормальной матрице A матрицы A_k $(k = 2, 3, \ldots)$ также будут нормальными.

Теорема 12.1. *Если A — нормальная матрица, то треугольные матрицы R_k основного* QR-*алгоритма сходятся к диагональной матрице.*

Доказательство. Рассмотрим две соседние матрицы QR-процесса, обозначая их для простоты через A и B. Переход от A к B описывается формулами

$$A=QR, \tag{12.1}$$

$$B=RQ. \tag{12.2}$$

Пусть b^i, a^i, r^i — столбцы, а b_i, a_i, r_i — строки соответственно матриц B, A и R $(i=1,\ldots,n)$. В силу (12.1) $\|a^i\|_2=\|r^i\|_2$, а в силу (12.2) $\|r_i\|_2=\|b_i\|_2$ $(i=1,\ldots,n)$. Нормальность матриц A и B означает, в частности, что $\|a^i\|_2=\|a_i\|_2$, $\|b^i\|_2=\|b_i\|_2$ $(i=1,\ldots,n)$. Положим

$$\Delta_m=\sum_{i=1}^{m}(\|b_i\|_2^2-\|a_i\|_2^2), \qquad m=1,\ldots,n-1.$$

Тогда

$$\Delta_1=\|b_1\|_2^2-\|a_1\|_2^2=\|r_1\|_2^2-\|r^1\|_2^2=|r_{12}|^2+\ldots+|r_{1n}|^2, \tag{12.3}$$

$$\Delta_m=\sum_{i=1}^{m}\|r_i\|_2^2-\sum_{i=1}^{m}\|r^i\|_2^2=\sum_{i=1}^{m}\sum_{j=m+1}^{n}|r_{ij}|^2, \qquad 2\leqslant m\leqslant n-1.$$

Если теперь составить для каждого k величины $\sigma_m^{(k)}=\sum_{i=1}^{m}\|a_i^{(k)}\|_2^2$, где $a_i^{(k)}$ — строки матрицы A_k и $1\leqslant m\leqslant n-1$, то получим $n-1$ последовательностей $\{\sigma_m^{(k)}\}$. Согласно (12.3), каждая из этих последовательностей монотонно возрастает и каждая ограничена (например, общим значением квадрата евклидовой нормы матриц A_k). Следовательно, последовательности $\{\sigma_m^{(k)}\}$ сходятся. Соответствующие последовательности $\{\Delta_m^{(k)}\}$, где $\Delta_m^{(k)}=\sigma_m^{(k+1)}-\sigma_m^{(k)}$, сходятся к нулю, а вместе с тем сходятся к нулю и все наддиагональные элементы матриц R_k.

Поскольку $\|r_1^{(k)}\|_2^2=\|a_1^{(k+1)}\|_2^2$, то

$$|r_{11}^{(k)}|^2=\sigma_1^{(k+1)}-|r_{12}^{(k)}|^2-\ldots-|r_{1n}^{(k)}|^2.$$

По соглашению $r_{11}^{(k)}\geqslant 0$ для всех k, а потому последовательность $\{r_{11}^{(k)}\}$ имеет предел. Точно так же из равенств

$$|r_{22}^{(k)}|^2=\|r_2^{(k)}\|_2^2-|r_{23}^{(k)}|^2-\ldots-|r_{2n}^{(k)}|^2=\|a_2^{(k+1)}\|_2^2-|r_{23}^{(k)}|^2-\ldots-|r_{2n}^{(k)}|^2=$$
$$=\sigma_2^{(k+1)}-\sigma_1^{(k+1)}-|r_{23}^{(k)}|^2-\ldots-|r_{2n}^{(k)}|^2$$

выводим, что и последовательность $\{r_{22}^{(k)}\}$ сходится. Продолжая аналогичным образом, установим существование предела матричной последовательности $\{R_k\}$.

Из доказанной теоремы вытекает, что с ростом k матрица A_k все более приближается к произведению унитарной и диагональной матриц. Остается установить, как могут выглядеть такие произведения.

Теорема 12.2. *Нормальная матрица A вида $A=QD$, где Q — унитарная, а D — диагональная матрицы с неотрицательными диагональными элементами, с точностью до симметричной перестановки строк и столбцов является блочно-диагональной. Каждый диагональный блок лишь скалярным множителем отличается от унитарной матрицы соответствующего порядка.*

Доказательство. Из равенства $AA^*=A^*A$ следует, что $D^2=QD^2Q^*$, или $D^2Q=QD^2$. Отсюда

$$q_{ij}(d_{ii}^2-d_{jj}^2)=0 \quad \forall i, j$$

и $q_{ij}=0$, если $d_{ii}\neq d_{jj}$. Матрица-перестановка $\mathscr{P}$, группирующая равные диагональные элементы матрицы $D: D\to\tilde{D}=\mathscr{P}^{\mathrm{T}}D\mathscr{P}$, приводит Q к блочно-диагональному виду: $Q\to\tilde{Q}=\mathscr{P}^{\mathrm{T}}Q\mathscr{P}$. Но тогда и матрица $\tilde{A}=\mathscr{P}^{\mathrm{T}}A\mathscr{P}=(\mathscr{P}^{\mathrm{T}}Q\mathscr{P})(\mathscr{P}^{\mathrm{T}}D\mathscr{P})=\tilde{Q}\tilde{D}$ блочно-диагональная, причем каждый диагональный блок есть произведение одноименного блока унитарной матрицы Q на число d, отвечающее этому блоку.

Мы переходим теперь к рассмотрению матриц общего вида. Хотя всегда можно считать исходную матрицу QR-процесса хессенберговой, поговорим сперва об условиях сходимости без такого предположения. Условия, одновременно необходимые и достаточные для сходимости, в этом случае, вероятно, были бы очень сложными. Автору не известна какая-либо работа, где подобные условия сформулированы.

На практике обычно пользуются достаточными условиями большей или меньшей общности. Один из чаще всего цитируемых критериев принадлежит Уилкинсону. Требования к матрице A в нем следующие. Во-первых, модули всех собственных значений должны быть различными:

$$|\lambda_1|>|\lambda_2|>\ldots>|\lambda_n|. \tag{12.4}$$

Отсюда вытекает, что матрица A диагонализуема:

$$A=X\Lambda X^{-1}, \quad \Lambda=\operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

Во-вторых, матрица $Y=X^{-1}$ должна допускать треугольное разложение. Другими словами, все ведущие главные миноры ее обязаны быть ненулевыми.

При выполнении этих условий матрицы A_k, порождаемые QR-алгоритмом, сходятся к верхней треугольной матрице, на диагонали которой стоят числа λ_i, причем именно в порядке $\lambda_1, \lambda_2, \ldots, \lambda_n$. Поддиагональный элемент $a_{ij}^{(k)}$ $(i>j)$ сходится к нулю со скоростью геометрической прогрессии со знаменателем $|\lambda_i|/|\lambda_j|$.

Мы не приводим доказательства критерия Уилкинсона, которое можно найти в [42, гл. 8, § 30]. Однако построенное на тех же идеях обобщение этого доказательства, пригодное для всех *методов декомпозиционного типа*, дано в п. 1 дополнений к § 12.

Два условия, составляющие критерий Уилкинсона, играют в нем существенно разные роли. Упорядоченность собственных значений на диагонали предельной треугольной формы обеспечивается именно вторым условием. От него можно отказаться, не теряя сходимости к треугольному виду, только диагональные элементы в последнем уже, вообще говоря, не будут расположены по убыванию модулей.

Отказ от условия (12.4) может иметь разные последствия. Если матрица A диагонализуема и модули ее *разных* собственных значений различны, то даже при наличии кратных корней сходимость к треугольной форме сохраняется (см. [42, гл. 8, § 32]). С другой стороны,

присутствие несовпадающих собственных значений одинакового модуля, вообще говоря, препятствует получению треугольного предельного вида. Примером может служить матрица (11.40).

В общем случае предельная форма основного QR-алгоритма — это блочно-треугольная матрица (или матрица, полученная из блочно-треугольной симметричной перестановкой строк и столбцов), диагональные блоки которой соответствуют группам собственных значений матрицы A с равными модулями. Исследование сходимости метода в самой общей ситуации проведено в [7, § 45].

Пусть теперь A — хессенбергова матрица. Без ограничения общности ее можно считать неразложимой: сходимость QR-алгоритма для разложимой матрицы эквивалентна его сходимости для каждого — уже неразложимого — диагонального блока.

Для неразложимой хессенберговой матрицы A Парлеттом получены необходимые и достаточные условия сходимости основного QR-алгоритма. Сформулируем основную теорему из его работы [156].

Предположим, что среди *модулей* собственных значений $\lambda_1, \ldots, \lambda_n$ матрицы A только r не равны нулю и различны; обозначим их через $\omega_1 > \omega_2 > \ldots > \omega_r > 0$. Среди корней с модулем ω_i пусть $p(i)$ имеют четные кратности

$$m_1^i \geqslant m_2^i \geqslant \ldots \geqslant m_{p(i)}^i > m_{p(i)+1}^i \equiv 0$$

и $q(i)$ имеют нечетные кратности

$$n_1^i \geqslant n_2^i \geqslant \ldots \geqslant n_{q(i)}^i > n_{q(i)+1}^i \equiv 0.$$

Если кратность нулевого собственного значения матрицы A равна m, то после m шагов QR-процесса нижние m строк матрицы A_{m+1} станут нулевыми, а последующие итерации можно вести с верхней главной подматрицей порядка $n-m$. Поэтому в дальнейшем A считается невырожденной матрицей.

Теорема 12.3 (см. [156]). *Матрицы A_k основного* QR*-процесса, начатого из матрицы A, сходятся к верхней блочно-треугольной форме. Диагональные блоки этой формы соответствуют группам собственных значений с одинаковым модулем. Если $A_{ii}^{(k)}$ $(i=1, \ldots, r)$ — диагональный блок в матрице A_k, то он в свою очередь сходится к верхней блочно-треугольной форме, далее уже неразложимой. Число диагональных блоков последней формы определяется правилом: имеется $m_j^i - m_{j+1}^i$ блоков порядка j $(j=1, \ldots, p(i))$, в совокупности дающих собственные значения модуля ω_i четной кратности, и $n_j^i - n_{j+1}^i$ блоков порядка j $(j=1, \ldots, q(i))$, отвечающих собственным значениям модуля ω_i нечетной кратности. Относительное расположение диагональных блоков зависит от положения чисел m_j^i, n_l^i в линейно упорядоченном множестве $\{m_j^i, n_l^i\}$ $(j=1, \ldots, p(i),\ l=1, \ldots, q(i))$. Скорость сходимости в общем случае медленная: для элементов поддиагональных блоков A_{ij} $(i>j)$ убывание модулей происходит со скоростью $O([\omega_i/\omega_j+\varepsilon]^k)$, ε — как угодно малое положительное число; внутри диагональных блоков — в их собственной поддиагональной части — скорость может падать до скорости гармонической последовательности $\{1/k\}$.*

Пример 12.4 (см. [156]). Предположим, что матрица A порядка 10 имеет четыре собственных значения модуля 1 с кратностями соответственно $m_1=4$, $m_2=2$, $n_1=3$, $n_2=1$. Тогда в предельной блочно-треугольной форме будет два диагональных блока порядка 2 и два диагональных элемента, которые в совокупности отвечают первым двум собственным значениям матрицы A; остальным двум собственным значениям соответствуют один диагональный блок порядка 2 и два диагональных элемента.

Из теоремы 12.3 вытекает

Следствие 12.5. *Для того чтобы при каждом значении* i $(2\leqslant i\leqslant n-1)$ *выполнялось предельное соотношение* $a^{(k)}_{i,i-1}a^{(k)}_{i+1,i}\to 0$, *необходимо и достаточно, чтобы в каждой группе собственных значений одинакового модуля было не более двух чисел четной и не более двух нечетной кратности.*

Если условия следствия 12.5 выполнены, то в предельной блочно-треугольной форме порядки диагональных блоков не превосходят 2. Будучи примененным к вещественным матрицам, следствие устанавливает критерий сходимости к вещественной форме Шура. Кратность собственных значений при этом оказывается несущественной. Важно лишь, чтобы число других собственных значений с тем же модулем было невелико.

Переходя к QR-алгоритму со сдвигами, мы начнем со сводки результатов для вещественного симметричного трехдиагонального случая. Хотя симметричные матрицы не являются предметом рассмотрения этой книги, сводка пригодится нам для сравнительной оценки фактов, относящихся к сходимости метода в случае матриц общего вида. Трехдиагональность матрицы не является ограничением: метод отражений для приведения к форме Хессенберга, будучи применен к симметричной матрице, сохранит ее симметрию, и, следовательно, форма Хессенберга будет трехдиагональной. То же самое справедливо в отношении эрмитовых матриц; при этом полученная из эрмитовой матрицы A трехдиагональная матрица может быть дополнительным диагональным унитарным преобразованием превращена в вещественную симметричную и по-прежнему трехдиагональную матрицу.

Так как все собственные значения трехдиагональной матрицы

$$T=\begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot\ \beta_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{bmatrix} \tag{12.5}$$

вещественны, то необходимости в двойных шагах нет. Поэтому будут рассматриваться только две разновидности процесса (11.16)—(11.17), а именно QR-алгоритм со сдвигами Рэлея и стратегия

Уилкинсона, понимаемая так: в качестве сдвига берется то собственное значение текущей подматрицы

$$\begin{bmatrix} \alpha_{n-1} & \beta_{n-1} \\ \beta_{n-1} & \alpha_n \end{bmatrix},$$

которое ближе к α_n.

Оказывается, что в некотором (см. ниже) смысле стратегия Рэлея для матрицы T эквивалентна методу RQI обратных итераций со сдвигами Рэлея. В силу известных свойств метода RQI в эрмитовом случае (см. п. 2 дополнений к § 10) отсюда следует, что с ростом k модуль элемента $\beta_{n-1}^{(k)}$ монотонно не возрастает; как правило, $\beta_{n-1}^{(k)}$ сходится к нулю, причем асимптотически скорость сходимости кубическая. Если сходимость имеет место, то после вычеркивания последних строки и столбца итерации продолжаются с подматрицей порядка $n-1$, и тогда аналогичные утверждения справедливы в отношении последовательности $\{\beta_{n-2}^{(k)}\}$, затем последовательности $\{\beta_{n-3}^{(k)}\}$ и т. д. Ситуации, когда какой-то внедиагональный элемент не стремится к нулю, неустойчивы: сколь угодно малым возмущением матрицу можно перевести в режим сходимости.

То, что несходимость матриц T_k к диагональному виду возможна, показывает пример матрицы

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

инвариантной относительно стратегии Рэлея. Однако не известно ни одного примера матрицы T, для которой последовательность $\{T_k\}$ не была бы стационарной и в то же время не сходилась к диагональной форме [35, § 8.7].

Заметим еще, что всегда справедливо предельное соотношение $\beta_{n-1}^{(k)}\beta_{n-2}^{(k)} \to 0$. Поэтому если $\beta_{n-1}^{(k)}$ не стремится к нулю, то $\beta_{n-2}^{(k)} \to 0$. Правда, сходимость последовательности $\{\beta_{n-2}^{(k)}\}$ может быть очень медленной [208].

Что касается стратегии Уилкинсона, то здесь сходимость к диагональной матрице имеет место всегда. Доказательство этого важного утверждения можно найти в [30, приложение B] или в [35, § 8.10].

На начальных шагах сходимость в стратегии Уилкинсона может быть линейной. Более точно, выполняются оценки [35, § 8.10]

$$|\beta_{n-1}^{(k+1)}|^3 < (\beta_{n-1}^{(k)})^2 |\beta_{n-2}^{(k)}| < \beta_{n-1}^2 |\beta_{n-2}| / (\sqrt{2})^{k-1}.$$

Асимптотически же скорость сходимости не хуже квадратичной (см. [30, соотношения (B.48)—(B.50)]). Пример, когда сходимость именно квадратичная, приведен в [35, § 8.11]. Однако более характерна для стратегии Уилкинсона сходимость кубическая или даже более быстрая. Дадим точную формулировку этого утверждения.

В общем случае порядок, в каком QR-алгоритм со сдвигами вычисляет собственные значения, может быть произвольным. Предположим, однако, что $\alpha_n^{(k)}$ сходятся к наименьшему собственному значению λ_1 матрицы T. Одновременно со стремлением к нулю чисел $\beta_{n-1}^{(k)}$ к нулю сходятся (разве что более медленно) и после-

довательности $\{\beta_{n-2}^{(k)}\}$, $\{\beta_{n-3}^{(k)}\}$; стало быть, диагональные элементы $\alpha_{n-1}^{(k)}$ и $\alpha_{n-2}^{(k)}$ сходятся к некоторым собственным значениям λ_2 и λ_3.

Теорема 12.6 (см. [35, § 8.11]). *Если λ_2 и λ_3 — соответственно второе и третье среди собственных значений матрицы T при упорядочении их по возрастанию, то*

$$|\beta_{n-1}^{(k+1)}/[(\beta_{n-1}^{(k)})^3(\beta_{n-2}^{(k)})^2]| \to 1/[(\lambda_2-\lambda_1)^3(\lambda_3-\lambda_1)].$$

Знаменатель правой части отличен от нуля, что следует из известного факта: неразложимая трехдиагональная симметричная матрица T не может иметь кратных собственных значений.

Другой специальный класс матриц, для которого также имеются глобальные результаты относительно сходимости QR-алгоритма со сдвигами,— это класс унитарных хессенберговых матриц. Предположение относительно формы Хессенберга снова не ограничивает общности: унитарность сохраняется при приведении к хессенбергову виду.

Сами по себе ни основной QR-алгоритм, ни стратегия Рэлея, ни стратегия Уилкинсона не гарантируют сходимости в унитарном случае. Подходящим примером остается матрица (11.40). Однако сходимости можно добиться, применяя следующую комбинированную стратегию [97] (члены QR-последовательности, порождаемой унитарной матрицей U, обозначаются через U_k):

1. Если $|u_{n,n-1}^{(k)}|=1$ (что в силу унитарности эквивалентно равенству $u_{nn}^{(k)}=0$), то в качестве τ_k можно взять любое ненулевое число.

2. Если $|u_{n-1,n-2}^{(k)}| \leqslant |u_{n,n-1}^{(k)}|/\sqrt{2}$, то выполняется сдвиг по Уилкинсону (т. е. за τ_k берется собственное значение подматрицы

$$\begin{bmatrix} u_{n-1,n-1}^{(k)} & u_{n-1,n}^{(k)} \\ u_{n,n-1}^{(k)} & u_{nn}^{(k)} \end{bmatrix},$$

ближайшее к $u_{nn}^{(k)}$).

3. Если условия пп. 1 и 2 не выполнены, производится сдвиг Рэлея $\tau_k = u_{nn}^{(k)}$.

Теорема 12.7 (см. [97, теорема 4.1]). *Последовательность $\{|u_{n,n-1}^{(k)}|\}$ монотонно сходится к нулю, причем асимптотически скорость сходимости не хуже квадратичной.*

Мы не даем доказательства, поскольку, с одной стороны, оно идейно очень схоже с доказательством для симметричного случая, приведенным в [30], с другой — диагонализация унитарных матриц не представляет большого интереса для практики. Вместе с тем скромная сама по себе теорема 12.7 приобретает известное значение на фоне отсутствия глобальной теории для матриц общего вида.

Заметим, что в ходе доказательства теоремы в [97] получено утверждение, также заслуживающее внимания: если U_k — последовательность, генерируемая из U посредством стратегии Рэлея, то при $k \to \infty$ либо $u_{n,n-1}^{(k)} \to 0$, либо $u_{n-1,n-2}^{(k)} \to 0$. Поскольку последовательность $\{|u_{n,n-1}^{(k)}|\}$ всегда монотонна, то в первом случае сходимость к нулю монотонная (в смысле расстояний).

Поведение QR-алгоритма для нормальных матриц исследовалось в работах Буурема [70, 71]. К сожалению, эти работы были не

доступны для автора, и судить о них приходится по кратким упоминаниям в [97, 124, 157]. В частности, в [157] отмечается, что Буурема доказал монотонное убывание последовательности $\{|h_{n,n-1}^{(k)}|\}$, где H_k — хессенберговы матрицы, получаемые в QR-процессе со сдвигами Рэлея из хессенберговой нормальной матрицы H. Как мы знаем из симметричного и унитарного случаев, доказать сходимость этой последовательности к нулю нельзя даже для этих двух матричных классов.

Рассмотрим теперь вопрос о связи QR-алгоритма с методом обратных итераций, причем никаких ограничений на матрицу или стратегию сдвигов накладывать не будем. Перепишем соотношение (11.16) в виде

$$(A_k-\tau_k I)^* Q_k = R_k^*. \tag{12.6}$$

Если обозначить через $q_n^{(k)}$ последний столбец матрицы Q_k и учесть, что в последнем столбце матрицы R_k^* только диагональный элемент может быть отличен от нуля, то получим выражение

$$(A_k-\tau_k I)^* q_n^{(k)} = \bar{r}_{nn}^{(k)} e_n. \tag{12.7}$$

Если τ_k — собственное значение матрицы A_k (т. е. A), то, как мы знаем из § 11, $r_{nn}^{(k)}=0$ и $q_n^{(k)}$ есть левый собственный вектор матрицы A_k, отвечающий числу τ_k. В противном случае формула (12.7) может интерпретироваться как шаг обратной итерации со сдвигом $\bar{\tau}_k$, проведенной для матрицы A_k^*, исходя из координатного вектора e_n. Число $|r_{nn}^{(k)}|$ с этой точки зрения обратно евклидовой длине вектора y, решающего систему $(A_k-\tau_k I)^* y = e_n$.

Поскольку $a_{nn}^{(k)} = (A_k e_n, e_n) = (A_k^* e_n, e_n)$, то стратегия Рэлея соответствует обратной итерации со сдвигом, равным отношению Рэлея. Если τ_k — хорошее приближение к некоторому собственному значению λ, то метод обратных итераций (см. § 10) сходится очень быстро, и уже один шаг может дать хорошее приближение $q_n^{(k)}$ к соответствующему собственному вектору.

Совокупный эффект ряда QR-шагов можно оценить с помощью следующих соображений. Так как, согласно (11.18), $A_{k+1} = Q_k^* A_k Q_k$, то

$$A_{k+1} = \tilde{Q}_k^* A \tilde{Q}_k, \tag{12.8}$$

где

$$\tilde{Q}_k = Q_1 Q_2 \dots Q_k. \tag{12.9}$$

Подставляя (12.8) (с заменой k на $k-1$) в (12.7), имеем

$$(A-\tau_k I)^* \tilde{Q}_{k-1}(Q_k e_n) = \bar{r}_{nn}^{(k)} \tilde{Q}_{k-1} e_n.$$

Мы видим, что последние столбцы соседних унитарных матриц $\tilde{Q}_{k-1}$ и $\tilde{Q}_k = \tilde{Q}_{k-1} Q_k$ связаны соотношением

$$(A-\tau_k I)^* \tilde{q}_n^{(k)} = \bar{r}_{nn}^{(k)} \tilde{q}_n^{(k-1)};$$

при $k=1$ полагаем $Q_0 = I$, $\tilde{q}_n^{(0)} = e_n$. Таким образом, последовательность $\{\tilde{q}_n^{(k)}\}$ порождается обратными итерациями матрицы A^*, начатыми

с вектора e_n и использующими сдвиги $\bar{\tau}_k$ ($k=1, 2, \ldots$). Эти сдвиги в силу равенств

$$\tau_k=(A_k e_n, e_n)=(A\tilde{Q}_{k-1}e_n, \tilde{Q}_{k-1}e_n)=(Aq_n^{(k-1)}, q_n^{(k-1)}),$$
$$\bar{\tau}_k=(A^* q_n^{(k-1)}, q_n^{(k-1)})$$

суть отношения Рэлея.

Известно, что обратные итерации со сдвигами Рэлея, если они сходятся к простому собственному значению и соответствующему собственному вектору, асимптотически сходятся очень быстро: с квадратичной скоростью для матрицы общего вида и кубической, если матрица эрмитова. Отсюда можно вывести локальную асимптотически квадратичную сходимость QR-алгоритма со стратегией Рэлея к простому собственному значению матрицы. Впрочем, это несложно доказать и непосредственно. Для простоты рассмотрим вещественный случай. Вид матрицы $A_k-\tau_k I$ можно, следуя [184, с. 366], проиллюстрировать следующим рисунком для $n=6$:

$$\begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & a & b \\ 0 & 0 & 0 & 0 & \varepsilon & 0 \end{bmatrix}.$$

Согласно нашему предположению, ε — малое число; наоборот, a не мало, поскольку $a_{nn}^{(k)}$ сходится к простому собственному значению. Можно проверить, что в позиции $(n, n-1)$ матрицы A_{k+1} стоит число $-\varepsilon^2 b/(a^2+\varepsilon^2)$, что и доказывает квадратичную сходимость.

Чтобы установить связь QR-алгоритма — на этот раз в основном варианте — со степенным методом, выведем одно вспомогательное соотношение.

Наряду с унитарной матрицей (12.9) определим треугольную матрицу

$$\tilde{R}_k=R_k \ldots R_2 R_1. \tag{12.10}$$

Покажем, что при всех k справедливо равенство

$$A^k=\tilde{Q}_k\tilde{R}_k. \tag{12.11}$$

При $k=1$ эта формула просто описывает первый полушаг алгоритма: $A=Q_1R_1=\tilde{Q}_1\tilde{R}_1$. Пусть (12.11) установлено для всех $k<m$. Используя (11.13) и (12.9), имеем

$$R_m=A_{m+1}Q_m^*=\tilde{Q}_m^*A\tilde{Q}_mQ_m^*=\tilde{Q}_m^*A\tilde{Q}_{m-1}.$$

Умножая эти равенства слева на $\tilde{Q}_m$, а справа на $\tilde{R}_{m-1}$, получаем

$$\tilde{Q}_m\tilde{R}_m=A\tilde{Q}_{m-1}\tilde{R}_{m-1}.$$

Поскольку при $k=m-1$ формула (12.11) верна, то она верна и при $k=m$.

Фиксируем теперь целое число l $(1 \leqslant l < n)$ и представим матрицу $\tilde{R}_k$ в виде

$$\tilde{R}_k = \begin{bmatrix} \tilde{R}_{11}^{(k)} & \tilde{R}_{12}^{(k)} \\ 0 & \tilde{R}_{22}^{(k)} \end{bmatrix},$$

где $\tilde{R}_{11}^{(k)}$ — подматрица порядка l. Приравнивая столбцы в (12.11), находим

$$\left[A^k e_1 \mid A^k e_2 \mid \ldots \mid A^k e_l\right] = \left[\tilde{q}_1^{(k)} \mid \tilde{q}_2^{(k)} \mid \ldots \mid \tilde{q}_l^{(k)}\right] \tilde{R}_{11}^{(k)}. \tag{12.12}$$

Если матрица $\tilde{R}_{11}^{(k)}$ не вырождена, то смысл соотношения (12.12) такой: векторы $\tilde{q}_1^{(k)}, \ldots, \tilde{q}_l^{(k)}$ составляют ортонормированный базис линейного подпространства, натянутого на $A^k e_1, \ldots, A^k e_l$. Это замечание позволяет вывести сходимость основного QR-алгоритма из сходимости одновременных итераций (см. п. 2 дополнений к § 14).

При $l = 1$ имеем равенство

$$A^k e_1 = \tilde{r}_{11}^{(k)} \tilde{q}_1^{(k)}.$$

Следовательно, *первые* столбцы матриц $\tilde{Q}_k$ основного QR-алгоритма порождаются степенным методом, начинаемым из вектора e_1. Метод применяется к самой матрице A, а не к A^* (ср. с (12.7)).

ДОПОЛНЕНИЯ К § 12

1. Исторически QR-алгоритм возник как унитарный аналог более раннего метода — так называемого LR-алгоритма (вспомним название статьи [109] Фрэнсиса). Основная операция LR-алгоритма (в отличие от QR-разложения не всегда выполнимая) — разложение матрицы в произведение двух треугольных: нижней L и верхней R. Итерационный LR-процесс описывается формулами

$$A_k = L_k R_k, \qquad A_{k+1} = R_k L_k.$$

Очевидная параллельность теории обоих методов породила ряд попыток описания их с общеалгебраических позиций. Расскажем об одной из них, следуя статье [92] и книге [67, § 4.5].

Примем следующие обозначения: $GL_n(C)$ — мультипликативная группа невырожденных комплексных матриц порядка n; $T_n(C)$ — группа невырожденных верхних треугольных комплексных матриц того же порядка. Пусть V и U — подгруппы соответственно в $GL_n(C)$ и $T_n(C)$, удовлетворяющие четырем условиям:

1) множество $V \cdot U = \{X \in GL_n(C) \mid X = GR,\ G \in V,\ R \in U\}$ открыто и плотно в $GL_n(C)$;

2) пересечение V и U содержит только единичную матрицу I;

3) пересечение V с $T_n(C)$ состоит только из диагональных матриц с равномерно ограниченными нормами;

4) отображение, ставящее в соответствие каждой матрице $X \in V \cdot U$ первый множитель G ее GR-*разложения* $X = GR$, $G \in V$, $R \in U$, непрерывно.

Отметим, что условие 2 обеспечивает единственность GR-разложения для каждой матрицы $X \in V \cdot U$. Действительно, из $X = GR = \tilde{G}\tilde{R}$ следовало бы $\tilde{G}^{-1} G = \tilde{R} R^{-1} \in V \cap U$; отсюда $G = \tilde{G}$ и $R = \tilde{R}$. Отметим еще, что как следствие условия 4 отображение $X \to R$ также непрерывно.

Конкретными примерами подгрупп V и U, удовлетворяющих условиям 1—4, могут служить подгруппы нижних унитреугольных и верхних треугольных

матриц или подгруппа унитарных матриц вместе с подгруппой матриц из $T_n(C)$, имеющих положительную диагональ.

Располагая парой подгрупп V и U, мы можем определить GR-*алгоритм* формулами

$$A_1 = A, \qquad A_k = G_k R_k, \quad G_k \in V, \ R_k \in U, \tag{12.13}$$

$$A_{k+1} = R_k G_k. \tag{12.14}$$

Предполагается, что начальная матрица A и порождаемые ею матрицы A_k допускают GR-разложение, т. е. принадлежат множеству $V \cdot U$.

Как и в QR-алгоритме, следствием соотношений (12.13)—(12.14) является подобие матриц GR-последовательности:

$$A_{k+1} = G_k^{-1} A_k G_k, \tag{12.15}$$

откуда

$$A_{k+1} = \tilde{G}_k^{-1} A \tilde{G}_k, \tag{12.16}$$

где

$$\tilde{G}_k = G_1 G_2 \ldots G_k. \tag{12.17}$$

Рассуждая таким же образом, как в основном тексте параграфа (см. (12.11)), можно вывести равенство

$$A^k = \tilde{G}_k \tilde{R}_k, \tag{12.18}$$

в котором $\tilde{R}_k = R_k \ldots R_2 R_1$.

Теорема (см. [92; 67, теорема 4.5.3]). *Пусть модули всех собственных значений матрицы A различны: $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$, и пусть в представлении*

$$A = XDX^{-1}, \qquad D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

матрица X допускает GR-*разложение $X = GR$, а матрица $Y = X^{-1}$ имеет обычное треугольное разложение $Y = L_1 U_1$, в котором матрицу L_1 без ограничения общности можно считать унитреугольной. В таком случае матрицы A_k, порождаемые* GR-*процессом* (12.13), (12.14), *сходятся к треугольной форме, причем* $\lim\limits_{k\to\infty} a_{ii}^{(k)} = \lambda_i$, *т. е. собственные значения упорядочиваются на диагонали предельной матрицы по убыванию модулей.*

Доказательство. Представим матрицу A^k в виде

$$A^k = XD^k X^{-1} = GRD^k L_1 U_1 = GR\left(D^k L_1 D^{-k}\right) D^k U_1. \tag{12.19}$$

Матрица $D^k L_1 D^{-k}$ нижняя унитреугольная, и ее элемент в позиции (i, j), $i > j$, равен $l_{ij}(\lambda_i/\lambda_j)^k$; следовательно, $\lim D^k L_1 D^{-k} = I$. Положим $E_k = I - D^k L_1 D^{-k}$, $F_k = RE_k R^{-1}$; тогда $E_k \to 0$, $F_k \overset{k\to\infty}{\to} 0$ и

$$A^k = GR(I - E_k) D^k U_1 = G(I - F_k) RD^k U_1. \tag{12.20}$$

Так как множество $V \cdot U$ открыто и содержит единичную матрицу I (поскольку $I \in V$ и $I \in U$), то при всех достаточно больших k матрица $I - F_k$ допускает GR-разложение $I - F_k = \hat{G}_k \hat{R}_k$, где $\hat{G}_k \in V$, $\hat{R}_k \in U$. Непрерывная зависимость сомножителей разложения от матрицы означает, что $\hat{G}_k \to I$, $\hat{R}_k \to I$ при $k \to \infty$. Еще раз перепишем представление для A^k:

$$A^k = G\left(\hat{G}_k \hat{R}_k\right) RD^k U_1. \tag{12.21}$$

Вместе с (12.18) это дает

$$\left(G\hat{G}_k\right)^{-1} \tilde{G}_k = \hat{R}_k RD^k U_1 \tilde{R}_k^{-1}.$$

Левая часть есть матрица из V, правая — матрица из $T_n(C)$. В силу условия 3 обе матрицы равны некоторой диагональной матрице D_k. Таким образом,

для всех достаточно больших k справедливо представление $\tilde{G}_k = G\hat{G}_k D_k$, откуда, используя (12.16), имеем $A_{k+1} = D_k^{-1}\hat{G}_k^{-1}G^{-1}AG\hat{G}_k D_k$. Но $A = XDX^{-1} = GRDR^{-1}G^{-1}$; поэтому $A_{k+1} = D_k^{-1}\hat{G}_k^{-1}RDR^{-1}\hat{G}_k D_k$, или, иначе, $D_k A_{k+1} D_k^{-1} = \hat{G}_k^{-1}RDR^{-1}\hat{G}_k$. Вследствие того что $\hat{G}_k \to I$, получаем равенство $\lim_{k\to\infty} D_k A_{k+1} D_k^{-1} = RDR^{-1}$. Матрицы D_k и D_k^{-1} принадлежат миожеству $V \cap T_n(C)$ и, согласно условию 3, их нормы равномерно ограничены. Поэтому матрицы A_k должны сходиться к верхней треугольной форме, в которой, как и в матрице RDR^{-1}, диагональные элементы упорядочены по убыванию модулей.

Замечание. Ничто не изменится в этом доказательстве, если A — вещественная матрица с вещественными и различными по абсолютной величине собственными значениями, а V и U определяются условиями 1—4 по отношению к группе $GL_n(R)$ вещественных невырожденных матриц.

В случае QR-алгоритма предположение теоремы относительно матрицы X излишне, поскольку QR-разложение существует для всякой невырожденной матрицы. Упорядоченность собственных значений на диагонали предельной формы обеспечивается наличием треугольного разложения у матрицы Y. Для LR-алгоритма нужны оба предположения (см. [42, гл. 8, § 6]).

К тематике, связанной с теоретико-групповым подходом к методам декомпозиционного типа, относятся также статьи [100, 103].

2. Упомянутый в п. 1 LR-алгоритм при всех своих недостатках (численная устойчивость его не гарантирована, разложение матрицы на треугольные множители не всегда существует) имеет одно немаловажное преимущество перед QR-алгоритмом — он сохраняет трехдиагональную форму матрицы. Для QR-алгоритма это верно лишь в вещественном симметричном или комплексном эрмитовом случае. LR-итерация трехдиагональной матрицы обходится в $O(n)$ операций, тогда как один шаг (LR или QR) для хессенберговой матрицы — в $O(n^2)$ операций. Из этого соотношения исходили авторы статьи [90], предлагая такую стратегию решения спектральных задач не очень высокого порядка. Вначале, как и в QR-алгоритме, матрица приводится к (верхней) форме Хессенберга H. Где-то запоминается копия матрицы H, а затем делается попытка преобразовать H к трехдиагональному виду. Для этого используется известный алгоритм Стрэйчи — Фрэнсиса [190], описанный, например, в [42, гл. 6, § 44]. Именно в предположении, что $h_{12} \neq 0$, составляются отношения

$$l_{23} = h_{13}/h_{12}, \quad l_{24} = h_{14}/h_{12}, \ \ldots, \ l_{2n} = h_{1n}/h_{12},$$

после чего 2-й столбец матрицы H, умноженный соответственно на l_{23}, l_{24}, ..., l_{2n}, вычитается из 3-го, 4-го, ..., n-го столбцов, а затем, чтобы завершить подобие, ко 2-й строке прибавляется 3-я, умноженная на l_{23}, 4-я, умноженная на l_{24}, и т. д. Теперь матрица приобрела трехдиагональную форму в первой строке и первом столбце. Если элемент (2, 3) не равен нулю, можно перейти ко второму шагу процесса: посредством операций со столбцами исключаются элементы в позициях (2, 4), ..., (2, n), а затем к 3-й строке прибавляется соответствующая линейная комбинация нижележащих строк. Если в дальнейшем элементы в *главных* позициях (3, 4), ..., ($n-2$, $n-1$) в нужные моменты оказываются ненулевыми, то после $n-2$ шагов процесса будет получена трехдиагональная матрица, собственные значения и векторы которой вычисляются посредством LR-алгоритма. При обнаружении нулевого или очень малого главного элемента предлагаемая стратегия предусматривает возвращение к хессенберговой форме H: приведение к трехдиагональному виду оказалось невозможным или численно рискованным. Чтобы повысить устойчивость приведения, рекомендуется выполнять его в арифметике удвоенной точности, благо число операций в этом процессе сравнительно невелико — приблизительно $n^3/6$.

Если приведение к трехдиагональной форме завершено благополучно, время решения спектральной задачи, по данным из [90], составляет всего 17% времени работы QR-алгоритма. В случае неудачи затраты времени возрастают не более чем на 13%. Поэтому в среднем стратегия комбинирования двух методов окупает себя.

3. В связи с теоремой 12.6 заметим, что, как показано в [129], условие $\beta_{n-3}^{(k)} \to 0$ при $k \to \infty$ является излишним. В [130] предложена смешанная стратегия, в которой в зависимости от относительной величины элементов $\beta_{n-1}^{(k)}$ и $\beta_{n-2}^{(k)}$ выбирается либо сдвиг Рэлея, либо сдвиг Уилкинсона. Доказано, что QR-алгоритм с такой стратегией сдвигов сходится глобально, и асимптотически скорость его сходимости *всегда* кубическая. Другая стратегия, также со сходимостью 3-го порядка, но требующая решения кубического уравнения на каждом шаге, указана в [217].

4. В недавней публикации Грэга [117] показано, что для унитарной хессенберговой матрицы U шаг QR-алгоритма со сдвигом можно выполнить за $O(n)$ операций. Основывается этот результат на том, что всякая унитарная хессенбергова матрица может быть разложена в произведенне $n-1$ элементарных унитарных матриц с опорными индексами соответственно $(1, 2)$, $(2, 3)$, ..., $(n-1, n)$ и диагональной унитарной матрицы, отличающейся от единичной разве что элементом (n, n). Параметры, определяющие сомножители такого разложения, называются параметрами Шура матрицы U; этих параметров ровно n. QR-процесс для матрицы U можно вести, пересчитывая от шага к шагу значения параметров Шура матриц U_k. Фактически Грэг указал способ пересчета, требующий на каждом шаге лишь $O(n)$ операций.

Значение методики Грэга состоит в том, что в комбинации с описанной в основном тексте параграфа стратегией Эберляйн — Хуанга она превращает вычисление собственных значений унитарных хессенберговых матриц в $O(n^2)$-процесс (если, как это принято, считать, что на все n значений достаточно $O(n)$ итераций QR-алгоритма).

Известно [12, гл. IX, § 12], что преобразование Кэли

$$B=(I+A)^{-1}(I-A)$$

переводит неисключительные унитарные матрицы (т. е. матрицы, не содержащие -1 в спектре) в косоэрмитовы и, наоборот, косоэрмитовы — в унитарные. Исходя из этого, в [22] описаны экономичные алгоритмы восстановления оригиналов по кэли-образам трехдиагональных косоэрмитовых и хессенберговых унитарных матриц, т. е. двух классов матриц, допускающих быстрое вычисление спектра.

5. В связи с упоминавшимся в основном тексте результатом Буурема [71] Парлетт [157] исследовал вопрос о том, насколько богато многообразие неразложимых нормальных хессенберговых матриц порядка n. Показано, что оно управляется n комплексными и $n-1$ вещественными положительными параметрами, тогда как линейное пространство всех хессенберговых $n \times n$-матриц имеет размерность $(n^2+3n-2)/2$.

6. В [170] для случая трехдиагональной симметричной матрицы T установлена связь одинарного QR-шага с методом Ньютона. Положим

$$\psi(\tau)=\frac{\det(T-\tau I)}{\det(T_{n-1}-\tau I_{n-1})}, \tag{12.22}$$

где T_{n-1} — ведущая главная подматрица порядка $n-1$. Если τ не является собственным значением для T_{n-1}, то в матрице $\tilde{T}$, полученной из T посредством QR-шага со сдвигом τ, элемент (n, n) выражается формулой

$$\tilde{t}_{nn}=\tau-\psi(\tau)/\psi'(\tau).$$

В частности, при пользовании стратегией Рэлея диагональные элементы $t_{nn}^{(k)}$ образуют ньютонову последовательность.

Этот результат обобщается [9] следующим образом. Если H — неразложимая хессенбергова матрица, а функция $\psi(\tau)$ определена по аналогии с (12.22), то для матрицы $\tilde{H}$, полученной после QR-шага со сдвигом t, выполнены соотношения

$$\tilde{h}_{n,n-1}=\alpha_t\psi(t), \qquad |\alpha_t|\leqslant 1,$$
$$\tilde{h}_{nn}=t+\beta_t\psi(t), \qquad 0\leqslant\beta_t\leqslant 1.$$

Как и выше, предполагается, что t не является собственным значением подматрицы H_{n-1}.

Таким образом, поведение диагонального элемента (n, n) соответствует методу последовательных приближений (с нижней релаксацией) для уравнения $\psi(t)=0$. Отсюда выводится

Теорема (см. [9]). *Если матрица H не имеет кратных собственных значений и последовательность сдвигов $\{\tau_k\}$ сходится к какому-либо ее собственному числу, то* QR*-алгоритм с такой стратегией сдвигов сходится.*

Исходя из этого результата, предлагается выбирать сдвиги в соответствии с тем или иным глобально сходящимся методом вычисления корней характеристического многочлена. Из численной практики известно (хотя теоретического доказательства нет до сих пор), что глобальной сходимостью обладает метод парабол (или метод Мюллера). Этот метод и рекомендован в [9] как основа стратегии сдвигов.

Параболический сдвиг используется и в версии QR-алгоритма из [181], при этом метод трактуется как средство уточнения уже имеющихся приближений к собственным значениям.

В [145] QR-алгоритм приспосабливается к задаче вычисления собственных значений, лежащих в заданном круге $\mathscr{K}(z_0, r)\equiv|z-z_0|<r$. Применяются двойные шаги, причем если из сдвигов, вычисленных по правилу Уилкинсона, хотя бы один попадает в $\mathscr{K}(z_0, r)$, то берутся именно эти сдвиги; в противном случае в качестве сдвигов берутся числа z_0 и $\bar{z}_0$.

Комбинированная стратегия, в которой, начиная с некоторого момента, происходит переключение с основного QR-алгоритма на двойные шаги со сдвигами по Уилкинсону, изучается в [138]. Указаны достаточные условия, обеспечивающие существование успешной стратегии такого типа. Однако момент переключения четко не охарактеризован.

7. В ряде работ последних лет [47, 80—82, 91, 146, 191] обсуждается любопытная связь QR-алгоритма с некоторым классом динамических систем, называемых *обобщенными потоками Тода*. Это нелинейные матричные дифференциальные уравнения (для определеиности — комплексные) вида

$$\frac{dL}{dt}=BL-LB$$

с косоэрмитовой матрицей B, зависящей специальным образом от L. В приводимом ниже изложении мы следуем статье [146].

Лемма 1. *Пусть $B(t)$ — косоэрмитова матричная функция, определенная для всех $t\in\mathbf{R}$. Матрица $U(t)$, решающая задачу Коши*

$$\frac{dU}{dt}=BU, \qquad U(0)=I, \tag{12.23}$$

также определена на всей вещественной оси и при этом унитарна.

Доказательство. Решение задачи (12.23) может быть распространено на все значения t, что следует из общей теории линейных обыкновенных дифференциальных уравиеиий. Дифференцируя матрицу U^*U, имеем

$$\frac{d}{dt}(U^*U)=U^*B^*U+U^*BU=U^*(B^*+B)U=0.$$

Так как $U^*(0)U(0)=I$, то $U^*(t)U(t)=\text{const}=I$.

Лемма 2. *Решение задачи Коши*

$$\frac{dL}{dt}=BL-LB, \qquad L(0)=L_0 \tag{12.24}$$

выражается формулой

$$L(t)=U(t)\,L_0\,U^*(t), \qquad -\infty<t<\infty. \tag{12.25}$$

Матрицы B и U те же, что и в лемме 1.

Доказательство дает следующая выкладка:

$$\frac{d}{dt}\left(U^*(t)\,L(t)\,U(t)\right)=U^*B^*LU+U^*(BL-LB)\,U+U^*LBU=0.$$

Так как $U(0)=I$ и $L(0)=L_0$, то (12.25) верно при $t=0$, а потому и при любом t.

Следствие. *Матрица $L(t)$ при всех t имеет одни и те же собственные значения. По-другому это выражают словами: поток* (12.25) *изоспектральный.*

До сих пор связь между матрицами B и L никак не использовалась. Теперь укажем, какого рода связи допускаются.

Пусть $G(\lambda)$ — аналитическая функция, определенная на спектре матрицы L, так что имеет смысл матрица $G(L)$. Обозначим через $[G(L)]_-$ и $[G(L)]_+$ соответственно нижнюю и верхнюю строго треугольные части матрицы $G(L)$ и положим

$$B=([G(L)]_-)^*-[G(L)]_-. \tag{12.26}$$

Теорема (см. [146]). *Решением задачи* (12.24), *где B указана в* (12.26), *является матрица*

$$L(t)=Q^*(t)\,L_0\,Q(t). \tag{12.27}$$

Здесь $Q(t)$ — унитарный сомножитель в том (единственном) QR*-разложении матрицы $e^{tG(L_0)}$, для которого треугольная матрица R имеет положительные диагональные элементы.*

Доказательство. Достаточно показать, что U^* в формуле (12.25) и есть матрица Q. Согласно лемме 1,

$$\frac{dU^*}{dt}=-U^*B=U^*(G(L)-\operatorname{diag}G(L)-[G(L)]_+)-U^*([G(L)]_-)^*. \tag{12.28}$$

Как следствие формулы (12.25) $G(L)=UG(L_0)\,U^*$. Полагая $R_1=-\operatorname{diag}G(L)-[G(L)]_+-([G(L)]_-)^*$, перепишем (12.28) в виде

$$\frac{dU^*}{dt}=G(L_0)\,U^*+U^*R_1. \tag{12.29}$$

Матрица R_1 верхняя треугольная. Решение задачи (12.29) с начальным условием $U^*(0)=I$ дает формула

$$U^*(t)=e^{tG(L_0)}R_2(t), \tag{12.30}$$

где R_2 определяется задачей

$$\frac{dR_2}{dt}=R_2R_1, \qquad R_2(0)=I.$$

Покажем, что R_2 — верхняя треугольная матрица с положительными диагональными элементами. Действительно, $d(R_2)_-/dt=[(R_2)_-R_1]_-$, и так как $(R_2)_-(0)=0$, то $(R_2)_-(t)=0$ при всех t. Для диагональных элементов имеем $d(R_2)_{jj}/dt=(R_2)_{jj}(R_1)_{jj}$; поскольку $(R_2)_{jj}(0)=1>0$, то $(R_2)_{jj}(t)>0$ $\forall t$.

Теперь формула (12.30), записанная как $e^{tG(L_0)} = U^*(t) R_2^{-1}(t)$, доказывает утверждение теоремы.

Следствие. Матрица $e^{G(L(k))}$ совпадает с $(k+1)$-й матрицей QR-алгоритма, примененного к $e^{G(L_0)}$. В частности, полагая $G(\lambda) = \ln \lambda$ (что возможно, если матрица L_0 не вырождена), приходим к QR-алгоритму для самой матрицы L_0.

Обнаружившееся родство между QR-алгоритмом и потоками, хотя и вызвало «мини-бум» среди специалистов по динамическим системам, тем не менее не привело пока к новым теоремам сходимости алгебраического метода.

§ 13. Некоторые приложения QR-алгоритма

В этом параграфе будут рассмотрены два из приложений QR-алгоритма к задачам с несимметричными матрицами — численное решение матричных уравнений некоторых типов и вычисление функций от матриц.

1. Численное решение матричных уравнений некоторых типов.

Линейные уравнения. Общий вид линейного матричного уравнения таков:

$$A_1 X B_1 + A_2 X B_2 + \ldots + A_k X B_k = C. \tag{13.1}$$

Здесь $m \times m$-матрицы $A_1, \ldots, A_k$, $n \times n$-матрицы $B_1, \ldots, B_k$ и $m \times n$-матрица C заданы, а X — искомая $m \times n$-матрица.

На основе QR-алгоритма можно построить эффективные численные методы для решения двух частных типов уравнения (13.1):

$$AX + XB = C \tag{13.2}$$

и

$$AXB - X = C. \tag{13.3}$$

Уравнения (13.2), (13.3) называются соответственно *непрерывным* и *дискретным уравнениями Сильвестра*. Среди линейных матричных уравнений в приложениях чаще всего встречаются уравнения именно этих двух типов. В особенности это относится к их частному случаю — *непрерывному* и *дискретному уравнениям Ляпунова*:

$$\begin{aligned} A^*X + XA &= C, \\ A^*XA - X &= C, \end{aligned} \tag{13.4}$$

где $m = n$, а матрица C эрмитова и разыскивается эрмитово решение X.

Встречающиеся на практике матричные уравнения обычно имеют вещественные коэффициенты, поэтому вплоть до конца данного раздела мы считаем матрицы A, B, C, X вещественными.

С принципиальной точки зрения каждое уравнение Сильвестра представляет собой компактную запись системы mn линейных уравнений с mn неизвестными — элементами искомой матрицы X. Но хотя традиционно решение линейных систем считается значительно более простым делом, чем решение спектральных задач, оказывается,

что наиболее просто уравнения Сильвестра решаются как раз с использованием спектральных методов.

Пусть V — ортогональная матрица, трансформирующая B к форме Шура:

$$\tilde{B}=V^{\tau}BV=\begin{bmatrix}\tilde{B}_{11} & \tilde{B}_{12} & \dots & \tilde{B}_{1l}\\ 0 & \tilde{B}_{22} & \dots & \tilde{B}_{2l}\\ \dots & \dots & \dots & \dots\\ 0 & 0 & \dots & \tilde{B}_{ll}\end{bmatrix}. \qquad (13.5)$$

Диагональные блоки $\tilde{B}_{11}, \dots, \tilde{B}_{ll}$ матрицы $\tilde{B}$ имеют порядок 1 либо 2. Матрицы $\tilde{B}$ и V строятся в QR-алгоритме, работающем в режиме вычисления собственных векторов.

Матрицу A также преобразуем к форме Шура, но для удобства не к верхней, а к нижней: .

$$\tilde{A}=U^{\tau}AU=\begin{bmatrix}\tilde{A}_{11} & 0 & \dots & 0\\ \tilde{A}_{21} & \tilde{A}_{22} & \dots & 0\\ \dots & \dots & \dots & \dots\\ \tilde{A}_{k1} & \tilde{A}_{k2} & \dots & \tilde{A}_{kk}\end{bmatrix}. \qquad (13.5)$$

Вычислить матрицу $\tilde{A}$ и ортогональную матрицу U можно посредством той же программы QR-алгоритма, что и в первом случае, только задать ей следует транспонированную матрицу A^{τ}.

Умножим обе части уравнений Сильвестра слева на U^{τ}, а справа на V:

$$(U^{\tau}AU)(U^{\tau}XV)+(U^{\tau}XV)(V^{\tau}BV)=U^{\tau}CV,$$

$$(U^{\tau}AU)(U^{\tau}XV)(V^{\tau}BV)-(U^{\tau}XV)=U^{\tau}CV.$$

Полагая $\tilde{C}=U^{\tau}CV$ и вводя замену неизвестных $Y=U^{\tau}XV$, приходим к уравнениям тех же типов:

$$\tilde{A}Y+Y\tilde{B}=\tilde{C}, \qquad (13.6)$$

$$\tilde{A}Y\tilde{B}-Y=\tilde{C}, \qquad (13.7)$$

но с блочно-треугольными матрицами коэффициентов $\tilde{A}$ и $\tilde{B}$. Системы линейных уравнений, эквивалентные таким уравнениям Сильвестра, распадаются на ряд подсистем первого, второго или четвертого порядков.

Поясним это утверждение с помощью примера. Пусть в уравнении (13.7) все матрицы имеют порядок 3, причем

$$\tilde{A}=\begin{bmatrix}\lambda_1 & 0\\ \tilde{A}_{21} & \tilde{A}_{22}\end{bmatrix},\quad \tilde{B}=\begin{bmatrix}\tilde{B}_{11} & \tilde{B}_{12}\\ 0 & \mu_3\end{bmatrix},$$

где $\tilde{A}_{22}$ и $\tilde{B}_{11}$ — блоки порядка 2, а $\tilde{A}_{21}$ и $\tilde{B}_{12}$ — 2×1-матрицы. Представим матрицу $\tilde{C}$ и искомую матрицу Y в виде

$$\tilde{C}=\begin{bmatrix}\tilde{C}_{11} & \tilde{C}_{12}\\ \tilde{C}_{21} & \tilde{C}_{22}\end{bmatrix},\quad Y=\begin{bmatrix}Y_{11} & Y_{12}\\ Y_{21} & Y_{22}\end{bmatrix}. \qquad (13.8)$$

В этом разбиении блоки верхнего ряда имеют размеры 1×2 и 1×1, а нижнего ряда — размеры 2×2 и 2×1. Подставляя представления (13.8) в уравнение (13.7), получаем

$$\lambda_1 Y_{11}\tilde{B}_{11}-Y_{11}=\tilde{C}_{11}, \tag{13.9}$$

$$(\lambda_1\mu_3-1)\,Y_{12}=\tilde{C}_{12}-\lambda_1 Y_{11}\tilde{B}_{12}, \tag{13.10}$$

$$\tilde{A}_{22}Y_{21}\tilde{B}_{11}-Y_{21}=\tilde{C}_{21}-\tilde{A}_{21}Y_{11}\tilde{B}_{11}, \tag{13.11}$$

$$\mu_3\tilde{A}_{22}Y_{22}-Y_{22}=\tilde{C}_{22}-\tilde{A}_{21}Y_{11}\tilde{B}_{12}-\tilde{A}_{22}Y_{21}\tilde{B}_{12}-\mu_3\tilde{A}_{21}Y_{12}. \tag{13.12}$$

Эти простейшие матричные уравнения Сильвестра решаются последовательно как линейные системы с двумя, одним, четырьмя и снова двумя неизвестными.

Заметим, что для однозначной разрешимости уравнений (13.10), (13.9), (13.12) должны выполняться соответственно условия $\lambda_1\mu_3\neq 1$, $\det(\lambda_1\tilde{B}_{11}-I_2)\neq 0$, $\det(\mu_3\tilde{A}_{22}-I_2)\neq 0$. Так как спектр матрицы A составляют λ_1 и собственные значения блока $\tilde{A}_{22}$, а спектр матрицы B — собственные значения блока $\tilde{B}_{11}$ и число μ_3, то эти неравенства означают, что λ_1 не может быть обратным числом ни для какого собственного значения матрицы B, а μ_3 не является обратным ни для одного собственного значения матрицы A. Можно показать, что и однозначная разрешимость уравнения (13.11) равносильна следующему требованию: $\lambda_i\mu_j\neq 1$ для собственных значений блоков $\tilde{A}_{22}$ и $\tilde{B}_{11}$.

В общем случае произвольных m и n условие однозначной разрешимости дискретного уравнения Сильвестра по-прежнему имеет вид

$$\lambda_i\mu_j\neq 1 \quad \forall\lambda_i\in\sigma(A), \qquad \forall\mu_j\in\sigma(B).$$

Непрерывное уравнение Сильвестра однозначно разрешимо тогда и только тогда, когда

$$\lambda_i+\mu_j\neq 0 \qquad \forall\lambda_i\in\sigma(A), \qquad \forall\mu_j\in\sigma(B).$$

Возвращаясь к уравнениям (13.6), (13.7), заметим, что после того, как матрица Y полностью вычислена, матрица X восстанавливается из нее по формуле $X=UYV^{\mathrm{T}}$.

В случае уравнений Ляпунова вычисления упрощаются: достаточно одного обращения (с матрицей A) к программе QR-алгоритма и может быть с выгодой учтена симметрия матриц C и X.

Изложенная методика решения уравнений Сильвестра называется *алгоритмом Бартелса — Стьюарта* и впервые была описана в 1972 г. [55]. Существует модификация алгоритма [115], позволяющая ограничиться приведением к форме Шура только одной из матриц A, B в уравнениях (13.2), (13.3); другая при этом преобразуется в хессенбергову матрицу, что, как мы знаем, может быть достигнуто посредством конечного процесса. Если порядки m и n матриц A и B существенно различаются, модифицированный алгоритм дает ощутимую экономию арифметической работы, правда, за счет некоторого увеличения требований к памяти.

Подробней об алгоритмах решения линейных матричных уравнений можно прочесть в книге [19].

Квадратичные уравнения. В этом разделе мы рассмотрим матричные уравнения следующего вида:

$$AX+XB+C+XFX=0. \tag{13.13}$$

Матрицы A, B, F, C заданы, причем A и B — квадратные матрицы порядка соответственно m и n; C — $m\times n$- и F — $n\times m$-матрица. Искомая матрица X имеет размеры $m\times n$. Заметим, что при $F=0$ уравнение (13.13) переходит в непрерывное уравнение Сильвестра. Однако, в то время как для линейных уравнений типична ситуация однозначной разрешимости, уравнения типа (13.13) обычно имеют много решений. Перечислить все комплексные решения можно, сопоставляя уравнению *вспомогательную* квадратную матрицу M порядка $n+m$:

$$M=\begin{bmatrix} -B & -F \\ C & A \end{bmatrix}. \tag{13.14}$$

Можно показать, что:

1) всякому решению X_0 уравнения (13.13) соответствует инвариантное подпространство матрицы M, натянутое на столбцы матрицы

$$\begin{bmatrix} I_n \\ X_0 \end{bmatrix}$$

и имеющее размерность n;

2) наоборот, если

$$W=\begin{bmatrix} Y \\ Z \end{bmatrix} \tag{13.15}$$

есть базисная матрица произвольного n-мерного инвариантного подпространства матрицы M и при этом $n\times n$-матрица Y не вырождена, то решение X уравнения (13.13) можно получить по формуле

$$X=ZY^{-1}. \tag{13.16}$$

Невырожденность блока Y не зависит от выбора конкретной базисной матрицы данного инвариантного подпространства. Доказательство этих утверждений можно найти, например, в [19, § 17].

В типичном случае матрица M имеет $n+m$ различных собственных значений. Тогда каждое инвариантное подпространство этой матрицы представляет собой линейную оболочку некоторой системы собственных векторов. Число различных n-мерных инвариантных подпространств оказывается равным числу различных выборов по n из $n+m$ собственных значений. Таким образом, если уравнение (13.13) имеет конечное множество решений, число их может доходить до C_{n+m}^{n}.

Матрица M с хотя бы одним кратным собственным значением может иметь в $\mathbf{C}^{n+m}$ бесконечное множество различных инвариантных подпространств, в том числе размерности n. В этом случае и множество решений уравнения (13.13), вообще говоря, бесконечно.

составлять те и только те числа из $\sigma(M)$, которые имеют отрицательные вещественные части. Будем для краткости называть такие собственные значения *устойчивыми*. Мы приходим к следующему выводу: чтобы найти по формуле $P=U_{21}U_{11}^{-1}$ положительно определенное решение уравнения Риккати, нужно построить ортогональную матрицу $U^{(P)}$, трансформирующую M к вполне определенной форме Шура, а именно к блочно-треугольной матрице $S^{(P)}$, сосредоточивающей все устойчивые свои собственные значения в верхнем $n\times n$-блоке $S_{11}^{(P)}$.

Строить матрицу $U^{(P)}$ можно таким образом: вначале посредством QR-алгоритма привести M к *какой-либо* форме Шура, а затем выполнить дополнительные ортогональные преобразования, переводящие эту форму в форму Шура нужного вида (см. последний раздел § 11). Если первый этап этого процесса по-прежнему описывать равенством (13.17), а второй этап — формулой $V^*SV=S^{(P)}$, то произведение UV и будет матрицей $U^{(P)}$:

$$(UV)^*M(UV)=S^{(P)}.$$

В отношении алгоритмической стороны построения матрицы $U^{(P)}$ заметим, что последняя формируется в том же массиве, какой на выходе программы QR-алгоритма содержал матрицу U. Правосторонние преобразования, совершаемые при перестройке формы Шура, применяются и к массиву, хранящему U. Конечный результат есть матрица $U^{(P)}$.

Положительная определенность симметричных матриц C и F — это простейшее из условий, обеспечивающих существование положительно определенного решения уравнения Риккати. Существуют и более тонкие условия такого типа; по этому вопросу отсылаем читателя к монографиям, посвященным линейной теории оптимального управления, например к [44].

Сходным образом может быть вычислено симметричное, положительно определенное решение дискретного уравнения Риккати

$$A^{\mathrm{T}}XA-X-A^{\mathrm{T}}XB(R+B^{\mathrm{T}}XB)^{-1}B^{\mathrm{T}}XA+Q=0, \tag{13.22}$$

которое, хотя и не является квадратичным, тесно связано с квадратичными уравнениями. Алгоритмы, решающие уравнения вида (13.13) и (13.22), описаны в § 17—19 книги [19].

2. Вычисление функций от матриц

Аналитические функции. Всякая аналитическая функция от матрицы A представима в виде многочлена от A и, значит, коммутирует с A. Основываясь на этом, Парлетт предложил остроумный способ вычисления функций от матрицы [159].

Пусть вначале A — комплексная $n\times n$-матрица с *различными* собственными значениями $\lambda_1, \lambda_2, \ldots, \lambda_n$, а $\Phi=\varphi(A)$ — искомая функция от нее. Если выполнить с A подобное преобразование: $A\to B=P^{-1}AP$, то таким же образом преобразуется функция φ: $\Phi=\varphi(A)\to\varphi(B)=P^{-1}\Phi P$. Если, в частности, взять унитарную матрицу, преобразующую A к форме Шура, то дело сведется к вычислению функции φ от треугольной матрицы T.

Итак, достаточно указать метод вычисления функций от верхней треугольной матрицы T. Заметим прежде всего, что матрица $F=\varphi(T)$, будучи многочленом от T, сама является верхней треугольной; при этом диагональные ее элементы суть не что иное, как числа $\varphi(\lambda_1), \ldots, \varphi(\lambda_n)$, т. е. значения скалярной функции φ на спектре матрицы T. Тем самым элементы $f_{11}, \ldots, f_{nn}$ можно считать известными. Чтобы вычислить наддиагональные элементы, воспользуемся свойством перестановочности

$$FT=TF. \tag{13.23}$$

Приравнивая здесь элементы в позиции $(i, i+1)$; получаем

$$f_{ii}t_{i,i+1}+f_{i,i+1}\lambda_{i+1}=\lambda_i f_{i,i+1}+t_{i,i+1}f_{i+1,i+1}, \tag{13.24}$$

откуда

$$f_{i,i+1}=t_{i,i+1}\frac{f_{i+1,i+1}-f_{ii}}{\lambda_{i+1}-\lambda_i}, \qquad i=1, \ldots, n-1.$$

Эти формулы определяют значения элементов первой наддиагонали матрицы F. Переходя к позициям $(i, i+2)$, из (13.23) выводим уравнения

$$f_{i,i+2}=t_{i,i+2}\frac{f_{i+2,i+2}-f_{ii}}{\lambda_{i+2}-\lambda_i}+\frac{t_{i,i+1}f_{i+1,i+2}-f_{i,i+1}t_{i+1,i+2}}{\lambda_{i+2}-\lambda_i}, \qquad i=1, \ldots, n-2. \tag{13.25}$$

Поскольку в правой части все величины известны, можно найти элементы второй наддиагонали. Продолжая подобным образом, в конечном счете построим всю матрицу F. Ее вычисление потребует приблизительно $n^3/3$ операций умножения (и примерно столько же сложений — вычитаний). Чтобы получить матрицу $\Phi=\varphi(A)$, остается проделать обратное преобразование: $F \to \Phi=PFP^*$.

В методе Парлетта QR-алгоритм используется как средство вычисления формы Шура T и трансформирующей матрицы P.

Понятно, что все сказанное до сих пор справедливо и для вещественных матриц с вещественным спектром, так как формы Шура в этом случае суть вещественные треугольные матрицы.

Если комплексная матрица имеет *кратные* собственные значения, то удобно взять такую ее форму Шура, в которой равные собственные числа занимают последовательные позиции диагонали. Это естественным образом приводит к блочному представлению:

$$T=\begin{bmatrix} T_{11} & T_{12} & \ldots & T_{1k} \\ 0 & T_{22} & \ldots & T_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & T_{kk} \end{bmatrix} \tag{13.26}$$

Диагональные блоки соответствуют группам равных собственных значений.

Аналогичную форму Шура можно выбрать для вещественной матрицы с кратными точками в вещественном спектре. Наконец, для произвольной вещественной матрицы будем считать, что блоки

Если в задаче, приводящей к квадратичному матричному уравнению, достаточно определить любое его решение, то, как вытекает из нашего обсуждения, можно действовать следующим образом: найти произвольное n-мерное инвариантное подпространство матрицы (13.14), выбрать любую базисную матрицу этого подпространства и, проверив невырожденность верхнего $n \times n$-блока в ней, вычислить решение по формуле (13.16). Эффективным средством поиска инвариантных подпространств как раз и оказывается QR-алгоритм.

Пусть, в самом деле, унитарная матрица U трансформирует матрицу M к (комплексной) верхней форме Шура S:

$$U^* M U = S. \tag{13.17}$$

Представим матрицы U и S в блочной форме:

$$U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}, \quad S = \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{22} \end{bmatrix},$$

где U_{11} и S_{11} — подматрицы порядка n. Переписывая (13.17) в виде

$$MU = US,$$

имеем, в частности,

$$M \begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix} = \begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix} S_{11}. \tag{13.18}$$

Тем самым подпространство, натянутое на столбцы матрицы

$$W = \begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix}, \tag{13.19}$$

является инвариантным подпространством матрицы M. Следовательно, при невырожденной подматрице U_{11} решение матричного уравнения дает формула $X = U_{21} U_{11}^{-1}$.

Как видно, главную работу в этом методе вычисления решений квадратичного матричного уравнения выполняет QR-алгоритм, строящий матрицу U.

До сих пор коэффициенты уравнения (13.13) считались комплексными матрицами, хотя уравнения, возникающие в приложениях, как правило, бывают вещественными. Мы умышленно начали с комплексного случая, поскольку здесь не возникает проблем с существованием инвариантных подпространств нужной размерности. Как следует из теоремы Шура, над комплексным полем всякая $N \times N$-матрица имеет инвариантное подпространство любой размерности k $(0 \leqslant k \leqslant N)$. Иначе обстоит дело в вещественном случае. Например у матрицы, все собственные значения которой комплексны, не может быть в $\mathbf{R}^N$ инвариантных подпространств нечетной размерности.

Однако если наперед известно, что уравнение (13.13) с вещественными коэффициентами имеет вещественные решения, то существование инвариантных подпространств размерности n обеспечено, и для их вычисления можно пользоваться вещественным QR-алгоритмом.

Более сложен случай, когда требуется найти не произвольное, а некоторое конкретное решение квадратичного уравнения. Если искать такое решение методом данного раздела, то нужно иметь информацию, какой части спектра соответствует инвариантное подпространство, определяемое этим решением.

В качестве примера рассмотрим важное для линейной теории оптимального управления *непрерывное уравнение Риккати*

$$XFX - A^{\mathrm{T}}X - XA - C = 0. \tag{13.20}$$

В нем все матрицы квадратные одинакового порядка n. При $F=0$ (13.20) есть (непрерывное) уравнение Ляпунова.

Будем считать, что матрицы C и F симметричные и положительно определенные. Хорошо известно, что в этом случае имеется ровно одна симметричная, положительно определенная матрица P, решающая уравнение Риккати. Прочие решения, если они существуют, не могут быть положительно определены. Известно также, что *матрица замкнутого контура* $A-FP$, отвечающая решению P, устойчива (см. § 1).

Вспомогательная матрица уравнения Риккати

$$M = \begin{bmatrix} A & -F \\ -C & -A^{\mathrm{T}} \end{bmatrix} \tag{13.21}$$

имеет четный порядок $2n$ и принадлежит классу *гамильтоновых матриц*. Подробней о матрицах этого типа рассказано в обзоре автора «Спектральные особенности специальных классов матриц» (Вычислительные процессы и системы. Вып. 8.— М.: Наука, 1990.). Сейчас упомянем только следующее: гамильтоновы матрицы характеризуются тем, что в их блочном 2×2-разбиении блоки (1,2) и (2,1)— симметричные матрицы, а блоки (1,1) и (2,2) связаны соотношением $M_{22} = -M_{11}^{\mathrm{T}}$. Спектр всякой гамильтоновой матрицы обладает замечательным свойством: наряду с каждым числом λ в него входит (причем с той же кратностью) число $-\lambda$. Дополнительное условие положительной определенности блоков C и F гарантирует отсутствие у M чисто мнимых собственных значений. В результате ровно половина спектра вспомогательной матрицы M расположена в левой комплексной полуплоскости.

Если вернуться к соотношению (13.18), то из него следует, в частности,

$$AU_{11} - FU_{21} = U_{11}S_{11}$$

или

$$A - FX = U_{11}S_{11}U_{11}^{-1},$$

где $X = U_{21}U_{11}^{-1}$ — решение уравнения (13.20). Таким образом, матрица замкнутого контура, определяемая решением X, подобна сужению S_{11} матрицы M на инвариантное подпространство с базисной матрицей (13.19).

Для разыскиваемого положительно определенного решения P матрица $A-FP$ устойчива, поэтому спектр матрицы $S_{11}^{(P)}$ должны

порождаются сопряженными парами комплексных собственных значений и вещественными собственными числами. Блоки первого типа обязательно имеют четный порядок; для блоков второго типа порядок может быть произвольным. Важно, что во всех перечисленных случаях у разных блоков T_{ii} и T_{jj} нет общих собственных значений.

Матрица $F=\varphi(T)$ тоже будет блочно-треугольной с таким же, как в (13.26), блочным разбиением:

$$F=\begin{bmatrix} F_{11} & F_{12} & \dots & F_{1k} \\ 0 & F_{22} & \dots & F_{2k} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & F_{kk} \end{bmatrix} \tag{13.27}$$

Методика Парлетта может быть перенесена на блочный случай, если удается (например, непосредственно из определения функции от матрицы) вычислить диагональные блоки $F_{ii}=\varphi(T_{ii})$ $(i=1, \dots, k)$. Формулам (13.24) теперь будут соответствовать равенства

$$T_{ii}F_{i,i+1}-F_{i,i+1}T_{i+1,i+1}=F_{ii}T_{i,i+1}-T_{i,i+1}F_{i+1,i+1}, \quad i=1, \dots, k-1, \tag{13.28}$$

представляющие собой уравнения Сильвестра относительно неизвестных матриц $F_{i,i+1}$. Эти уравнения однозначно разрешимы, поскольку спектры матриц T_{ii} и $T_{i+1,i+1}$ не пересекаются. Матрицы T_{ii} уже имеют максимально простой — для данной формы Шура — вид. Поэтому решать уравнения (13.28) следует, как системы линейных уравнений. Обычно порядки диагональных блоков в (13.26) невелики, поэтому и линейные системы не будут большими.

Определив блоки первой блочной наддиагонали, перейдем к вычислению блоков вида $F_{i,i+2}$. Они тоже являются решениями уравнений Сильвестра

$$\begin{aligned} T_{ii}F_{i,i+2}-F_{i,i+2}T_{i+2,i+2}=F_{ii}T_{i,i+2}- \\ -T_{i,i+2}F_{i+2,i+2}+F_{i,i+1}T_{i+1,i+2}- \\ -T_{i,i+1}F_{i+1,i+2}, \qquad i=1, \dots, k-2 \end{aligned}$$

(ср. с (13.25)), а к самим уравнениям приложимо все сказанное относительно (13.28). Продвигаясь от диагонали к диагонали, мы в конце концов построим матрицу (13.27).

Отметим, что метод вычисления функций от треугольной матрицы, рассмотренный в этом разделе, независимо от Парлетта (хотя и позднее, чем он) обнаружил Филиппони [107].

Корни целой степени. Начнем данный раздел с наиболее важного частного случая — *вычисления квадратных корней*. Сама по себе эта задача имеет ряд существенных отличий от вычисления аналитических функций матрицы. Такие функции, как экспонента или тригонометрические, существуют для любой квадратной матрицы. В то же время корень можно извлечь не из всякой матрицы, хотя функция $\sqrt{z}$ определена на всей комплексной плоскости. Эта функция двузначна, что влечет за собой многозначность матричной функции

$A^{1/2}$ в области ее определения. Наконец, корень из матрицы может не быть многочленом от нее. Так, матрица

$$X=\begin{bmatrix}0&0&1\\0&0&0\\0&1&0\end{bmatrix}$$

является квадратным корнем из матрицы

$$A=\begin{bmatrix}0&1&0\\0&0&0\\0&0&0\end{bmatrix}, \tag{13.29}$$

но в отличие от A не имеет треугольного вида.

По поводу существования квадратного корня отметим следующее. Из всякой невырожденной матрицы корень извлечь можно. Число различных корней в общем случае очень велико и может даже быть бесконечным. Так, всякая диагонализуемая матрица со спектром, состоящим из двух точек 1 и -1, представляет собой квадратный корень из единичной матрицы. Для матрицы порядка n, все собственные значения которой различны, число корней равно 2^n.

Вырожденная матрица может иметь (но может и не иметь) квадратные корни. Например, для матрицы (13.29) квадратный корень существует, а для матрицы

$$\begin{bmatrix}0&1\\0&0\end{bmatrix}$$

такого корня нет.

Как ни странно, даже в монографиях, глубоко трактующих тему функций от матрицы (см., например, [12, гл. VIII]), отсутствуют удобные критерии существования квадратных корней в вырожденном случае. Такой критерий приведен в статье [64]. Он заключается в следующем.

Пусть $n_1, n_2, \ldots, n_{2m-1}, n_{2m}$ — упорядоченные по убыванию размерности жордановых клеток матрицы A, относящихся к нулевому собственному значению. Число таких клеток мы считаем четным, полагая при необходимости $n_{2m}=0$. Квадратный корень из матрицы A существует тогда и только тогда, когда

$$n_{2i-1}-n_{2i}\leqslant 1, \qquad i=1, \ldots, m.$$

Хотя квадратный корень X из матрицы A не обязан быть многочленом от нее, всегда верно обратное отношение: A есть многочлен от любого своего корня (а именно $A=X^2$). Впрочем, это отношение справедливо для корней любой целой степени. Следовательно, A коммутирует со всеми своими корнями.

Перестановочность A и X позволяет применить к вычислению матрицы X метод Парлетта из предыдущего раздела. Совершая унитарный переход

$$A\to T=P^*AP, \qquad X\to U=P^*XP,$$

сводим задачу к вычислению корня из треугольной матрицы T. В выборе формы Шура T у нас, как всегда, есть свобода, и мы используем ее для того, чтобы сгруппировать нулевые диагональные элементы в нижних позициях диагонали $(r+1, r+1)$, ..., (n, n). В дальнейшем мы *будем разыскивать квадратный корень U треугольного, как и сама матрица T, вида*. Достаточным условием его существования, как легко проверить, является выполнение равенств $t_{ij}=0$ $(r<i<j)$.

В методе Парлетта элементы матрицы U вычислялись бы с помощью соотношения перестановочности $UT=TU$. В данном случае можно исходить непосредственно из определения квадратного корня

$$U^2=T. \tag{13.30}$$

Полагая $u_{ii}=t_{ii}^{1/2}$ $(i=1, ..., n)$ и приравнивая в (13.30) элементы в позиции (i, j) $(i<j)$, получаем выражение

$$u_{ij}=\frac{t_i^j-\sum\limits_{k=i+1}^{j-1} u_i^k u_k^j}{u_{ii}+u_{jj}}, \qquad i<j. \tag{13.31}$$

Формула (13.31) дает возможность, двигаясь вдоль диагоналей матрицы U, последовательно находить все ее элементы: вначале элементы в позициях $(i, i+1)$ как частные $t_i^j/(u_{ii}+u_{jj})$ $(i=1, ..., n-1)$, затем элементы

$$u_{i,\,i+2}=\frac{t_{i,\,i+2}-u_{i,\,i+1}\,u_{i+1,\,i+2}}{u_{ii}+u_{i+2,\,i+2}}, \qquad i=1, ..., n-2,$$

и т. д. Нужно сказать только о случае, когда $t_{ii}=t_{jj}$ для разных i, j. Если для таких i, j и значения квадратных корней брать равными, т. е. $u_{ii}=u_{jj}$, то знаменатель в (13.31) будет отличен от нуля при $t_{ii}\neq 0$; для $t_i^i=t_{jj}=0$ полагаем $u_{ij}=0$.

Теперь приведем некоторые соображения относительно численной устойчивости изложенной методики, как и вообще задачи отыскания квадратного корня. В этом, как и в данном выше описании численной процедуры, мы следуем материалу статьи [64].

Анализ ошибок округлений в вычислениях по формуле (13.31) приводит к такому выводу: *вычисленная* матрица $\tilde{U}$ удовлетворяет соотношению

$$\tilde{U}^2=T+F,$$

где для нормы матрицы эквивалентного возмущения F справедлива оценка

$$\|F\| \leqslant f(n)\left(\|T\|+\|\tilde{U}\|^2\right)\beta^{-t}.$$

Напомним (см. § 2), что для метода, численно устойчивого по Уилкинсону, эквивалентное возмущение должно быть ограничено величиной типа $f(n)\|T\|\beta^{-t}$. Поэтому формулы (13.31) можно считать численно устойчивой процедурой извлечения квадратного корня лишь

в том случае, если $\| \tilde{U} \|^2$ и $\| T \|$ имеют одинаковый порядок. Так будет не всегда. Например, для матрицы

$$T=\begin{bmatrix} \varepsilon & 1 \\ 0 & \varepsilon \end{bmatrix}, \qquad \varepsilon > 0,$$

существует только два квадратных корня:

$$U_1=\begin{bmatrix} \varepsilon^{1/2} & 1/(2\varepsilon^{1/2}) \\ 0 & \varepsilon^{1/2} \end{bmatrix}, \quad U_2=-U_1=\begin{bmatrix} -\varepsilon^{1/2} & -1/(2\varepsilon^{1/2}) \\ 0 & -\varepsilon^{1/2} \end{bmatrix}$$

и нормы обоих бесконечно растут при $\varepsilon \to 0$, тогда как норма $\| T \|$ ограничена.

Однако и обусловленность самой задачи извлечения квадратного корня тоже связана с относительной величиной корней. Действительно, пусть U — *точный* квадратный корень из матрицы T. Округление элементов матрицы U до машинно-представимых чисел ведет к замене ее матрицей $\tilde{U}$ такой, что

$$\tilde{U}=U+E, \quad \| E \| \leqslant \beta^{-t} \| U \|.$$

Так как $\tilde{U}^2=T+UE+EU+E^2$, то, полагая $F=UE+EU+E^2$, можно рассматривать $\tilde{U}$ как точный квадратный корень из возмущенной матрицы $T+F$. При этом оценка нормы матрицы F, вообще говоря, неизбежно должна включать в себя слагаемое, зависящее от $\| U \|^2$.

Если матрица T не вырождена, то плохая обусловленность задачи вычисления корня U свидетельствует о плохой обусловленности самого этого корня. В самом деле, из тождества $U=U^2U^{-1}=TU^{-1}$ следует $\| U \| \leqslant \| T \| \| U^{-1} \|$, или

$$\| U \|^2 / \| T \| \leqslant \| U \| \| U^{-1} \| = \operatorname{cond} U.$$

Резюмируя, можно сказать, что формулы (13.31) дают численно устойчивый метод извлечения корня из треугольной матрицы, а вся двухступенчатая процедура, основанная на приведении к форме Шура,— численно устойчивый метод извлечения корня из матрицы общего вида, если сама задача извлечения хорошо обусловлена.

Однако это заключение неполно. Дело в том, что мы ограничили себя отысканием *корней треугольного вида* из треугольной матрицы T. Между тем такой корень может не существовать или быть плохо обусловлен в то же время, когда имеются хорошо обусловленные *нетреугольные* корни из T. Например, все верхние треугольные квадратные корни из матрицы

$$T=\begin{bmatrix} \varepsilon & 1 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{bmatrix}, \qquad \varepsilon > 0,$$

содержат элемент $u_{12}=\pm 1/(2\varepsilon^{1/2})$ и потому плохо обусловлены при $\varepsilon \to 0$. Но у матрицы T есть и хорошо обусловленный корень

$$X=\begin{bmatrix} \varepsilon^{1/2} & 0 & 1 \\ 0 & \varepsilon^{1/2} & 0 \\ 0 & 1 & -\varepsilon^{1/2} \end{bmatrix}.$$

Можно поставить задачу о поиске хорошо обусловленных корней из заданной матрицы A. Рассмотрим один из подходов к ее

численному решению [64]. Предлагается рассматривать поиск корня X как задачу минимизации функции

$$f(X)=\sum_{i,\,j=1}^{n}|x_{ij}|^2$$

на множестве $X^2-A=0$. Применяя стандартные методы оптимизации, можно для сокращения арифметической работы пользоваться специальной формой дифференциала $G(X)$ матрицы ограничений $C(X)\equiv \equiv X^2-A$. Именно действие дифференциала на приращении H выражается формулой $G(X)H=XH+HX$; если же нужно определить матрицу H из условия $G(X)H=Y$, где X и Y заданы, то дело сводится к решению матричного уравнения Сильвестра.

Чтобы снизить число переменных оптимизационной задачи, можно комбинировать этот подход с методом Парлетта. Как и в последнем, исходная матрица A приводится к форме Шура, так что остается вычислить корень из треугольной матрицы T. Представим ее в блочной форме:

$$T=\begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix},$$

где T_{11} — подматрица наибольшего порядка, *треугольный* корень U_{11} из которой хорошо обусловлен. Дополнительно потребуем, чтобы спектры матриц T_{11} и T_{22} не пересекались. Запишем искомую матрицу U в аналогичном блочном виде:

$$U=\begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix},$$

где U_{22} уже *не обязательно треугольный* корень из T_{22}. Он может быть найден оптимизационными средствами. Так как

$$\begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}^2=\begin{bmatrix} U_{11}^2 & U_{11}U_{12}+U_{12}U_{22} \\ 0 & U_{22}^2 \end{bmatrix}=\begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix},$$

то блок U_{12} определяется из уравнения Сильвестра $U_{11}U_{12}+ +U_{12}U_{22}=T_{12}$.

В заключение данного раздела рассмотрим процедуру вычисления кубического корня из матрицы A [64]. Выполняя приведение к форме Шура, можно считать, что кубический корень U нужно извлечь из верхней треугольной матрицы T. Вводя вспомогательную треугольную матрицу

$$R=U^2, \tag{13.32}$$

имеем

$$UR=T. \tag{13.33}$$

Приравнивание элементов в матричных соотношениях (13.32), (13.33) дает для $i<j$

$$t_{ij}=\sum_{k=i}^{j}u_{ik}r_{kj}=u_{ii}r_{ij}+u_{ij}r_{jj}+\sum_{k=i+1}^{j-1}u_{ik}r_{kj}, \tag{13.34}$$

$$r_{ij}=\sum_{k=i}^{j}u_{ik}u_{kj}=u_{ij}(u_{ii}+u_{jj})+z_{ij}, \tag{13.35}$$

где принято

$$z_{i,i+1}=0, \qquad i=1, \ldots, n-1,$$

$$z_{ij}=\sum_{k=i+1}^{j-1} u_{ik}u_{kj}, \quad i<j-1. \tag{13.36}$$

Из (13.34)—(13.36) получаем расчетные формулы:

$$u_{ii}=t_{ii}^{1/3}, \quad r_{ii}=u_{ii}^2, \qquad i=1, \ldots, n,$$

$$u_{ij}=\frac{t_{ij}-u_{ii}z_{ij}-\sum_{k=i+1}^{j-1} u_{ik}r_{kj}}{r_{ii}+u_{ii}u_{jj}+r_{jj}},$$

$$r_{ij}=u_{ij}(u_{ii}+u_{jj})+z_{ij}, \qquad i<j.$$

ДОПОЛНЕНИЯ К § 13

1. Задача вычисления вещественных квадратных корней из вещественной невырожденной матрицы A рассматривается в [121]. Необходимым и достаточным условием существования таких корней является присутствие в жордановой форме матрицы четного — при одном и том же порядке — числа клеток для каждого вещественного отрицательного собственного значения. Ограничиваясь же корнями, представимыми в виде вещественных многочленов от A, нужно потребовать, чтобы отрицательных собственных чисел не было вообще. Для вычисления корней можно, как и в комплексном случае, использовать приведение к форме Шура T, которая теперь представляет собой блочно треугольную матрицу с диагональными блоками порядка 1 или 2. Блоки порядка 2 отвечают парам сопряженных комплексных корней. Если

$$T_{ii}=\begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix}$$

есть такой блок, а $\lambda=\theta+i\mu$ и $\bar{\lambda}$ — его собственные значения, то существует ровно два вещественных квадратных корня из T_{ii} (два других корня чисто мнимы):

$$U_{ii}=\pm\left[\alpha I+\frac{1}{2\alpha}(T_{ii}-\theta I)\right];$$

число α определяется равенством $(\alpha+i\beta)^2=\theta+i\mu$. Выбрав определенные корни для диагональных блоков, можно затем последовательно определить внедиагональные блоки, решая относительно их элементов линейные системы 1-го, 2-го или 4-го порядка.

Пусть r и c обозначают соответственно число различных вещественных неотрицательных собственных значений и число различных пар сопряженных комплексных собственных значений матрицы A. Тогда имеется 2^{r+c} различных вещественных корней из A, являющихся многочленами от этой матрицы.

Обусловленность задачи вычисления корня X из матрицы A была охарактеризована в основном тексте посредством величины $\alpha(X)=\|X\|^2/\|A\|$ или — после перехода к форме Шура T — величины

$$\tilde{\alpha}(U)=\|U\|^2/\|T\|, \tag{13.37}$$

где $T=P^*AP$, $U=P^*XP$. Для унитарно-инвариантных норм обе характеристики совпадают.

Значения $\tilde{\alpha}$ для разных корней из матрицы T могут различаться весьма сильно. Так, среди 16 корней матрицы

$$T=\begin{bmatrix} 1 & -1 & -1 & -1 \\ & 1.1 & -1 & -1 \\ & & 1.5 & -1 \\ 0 & & & 2 \end{bmatrix}$$

есть два корня с $\tilde{\alpha}_1(U)=1.64$ (индекс 1 указывает, что в (13.37) взята норма $\|\cdot\|_1$) и два корня с $\tilde{\alpha}_1(U)=1990.35$.

Если для задачи, требующей отыскания квадратного корня из матрицы, все корни равноправны, то можно искать корень с минимальным или близким к минимальному значением $\tilde{\alpha}$. В [121] описан алгоритм такого поиска; основан он на той же идее, что и алгоритм оценивания обусловленности линейной системы, используемый в подпрограммах пакета LINPACK (см., например, [19, § 9]). Рассмотрим для простоты случай, когда форма Шура T является треугольной матрицей. Вместо того чтобы сразу зафиксировать значения диагональных элементов u_{ii} квадратного корня U, а затем, двигаясь по диагоналям, параллельным главной, вычислять внедиагональные элементы u_{ij}, будем определять элементы U *по столбцам*. Пусть уже найдены значения элементов u_{ik} для первых $j-1$ столбцов. Тогда

$$u_{jj}=\pm\sqrt{t_{jj}}, \tag{13.38}$$

$$u_{ij}=\frac{t_{ij}-\sum_{k=i+1}^{j-1} u_{ik}u_{kj}}{u_{ii}+u_{jj}}, \qquad i=j-1, \ldots, 1.$$

Не фиксируя пока определенный знак для u_{jj}, построим два набора значений элементов $u_{jj}, u_{j-1,j}, \ldots, u_{1j}$, отвечающих выбору того и другого знака в (13.38). Обозначим эти значения соответственно через u_{ij}^{+} и u_{ij}^{-}. Положим

$$c_j^{+}=\sum_{i=1}^{j}|u_{ij}^{+}|, \qquad c_j^{-}=\sum_{i=1}^{j}|u_{ij}^{-}|.$$

Если $c_j^{+}\leqslant c_j^{-}$, то в качестве окончательных значений для u_{ij} берем значения u_{ij}^{+}, в противном случае — значения u_{ij}^{-}.

Как видно из сказанного, алгоритм ориентирован на локальное уменьшение l_1-длин столбцов квадратного корня; расчет при этом тот, что в итоге и норма всей матрицы U будет близка к возможному минимуму. Результаты экспериментов, о которых сообщается в [121], подтверждают эффективность алгоритма. Арифметическая работа при таком поиске удваивается по сравнению с вычислением произвольного корня из T. Однако извлечение корня из формы Шура составляет лишь малую долю общих вычислительных затрат метода; гораздо большей работы требует преобразование U в X и в особенности первоначальное приведение к форме Шура. Поэтому введение процедуры поиска лишь незначительно увеличивает суммарное число арифметических операций при извлечении корня из A.

ГЛАВА 5. МЕТОДЫ ДЛЯ РАЗРЕЖЕННЫХ МАТРИЦ

QR-алгоритм, изучавшийся в гл. 4, является типичным представителем методов трансформационного типа. Общая идея методов этой группы состоит в том, чтобы посредством последовательности простых подобий привести матрицу к той или иной канонической форме, позволяющей уже без труда определить собственные значения и векторы. В QR-алгоритме это форма Шура, в более ранних методах использовались другие, например форма Фробениуса.

Если трансформационный метод применить к *разреженной* матрице, т. е. матрице, имеющей большое количество нулевых элементов, то одновременно с преобразованиями матрицы будет более или менее быстро происходить ее *заполнение*, т. е. в позициях, где первоначально стояли нули, появятся ненулевые числа. С другой стороны, есть позиции, где по самому смыслу алгоритма должны быть получены и сохранены нули, скажем элементы $(i, j)\,(i>j+1)$ при приведении к (верхней) форме Хессенберга. Однако число таких позиций не компенсирует, как правило, заполнения, и общим итогом будет то, что *заполненность* матрицы сильно возрастает. Для матрицы A большого порядка n это означает, что, даже если удается разместить A в оперативной памяти, пользуясь компактной формой записи (численные значения ненулевых элементов плюс индексная информация, определяющая их местоположение в матрице), после нескольких шагов алгоритма преобразованную матрицу хранить будет уже невозможно.

Та же проблема заполнения возникает при решении разреженных линейных систем, и за минувшие три десятилетия найдены довольно эффективные приемы, помогающие сохранить если не исходный, то все же приемлемый уровень разреженности на всех шагах прямого алгоритма. К сожалению, численная практика показывает (см., например, [95]), что для спектральных задач аналогичные приемы малоуспешны.

В этой ситуации на авансцену выходят методы другого класса, а именно методы, не требующие преобразования исходной матрицы. В таких методах матрица A выступает как оператор, превращающий задаваемый вектор x в вектор-произведение $y=Ax$. При обращении к программе метода пользователь составляет только подпрограмму умножения *своей* матрицы на вектор-параметр.

Можно сказать, что именно так обстоит дело в степенном методе. Далее, подпрограмму, решающую линейную систему $Ay=x$, тоже можно считать подпрограммой умножения матрицы на вектор,

а именно матрицы A^{-1} на вектор x. Поэтому и обратные итерации нужно причислить к данному классу. Отчего же эти методы были рассмотрены гораздо раньше, в гл. 3?

На то есть две причины. Во-первых, оба метода, в особенности обратные итерации, широко используются и для заполненных матриц. Так, в гл. 4 говорилось, что подпрограммы метода обратных итераций есть и в справочнике [43], и в пакете EISPACK, причем один из режимов работы QR-алгоритма предусматривает обращение к ним для вычисления собственных векторов. Во-вторых, для степенного метода и обратных итераций характерно отыскание собственных векторов «по одному», тогда как методы, изучаемые в настоящей главе, вычисляют (или могут вычислять) сразу группу векторов.

Методы для разреженных матриц можно разделить на две основные группы: методы *одновременных итераций* (или *итерирования подпространства*) и методы *типа Ланцоша*. В 60-е и 70-е годы при решении больших спектральных задач пользовались почти исключительно методами первой группы. В настоящее время в симметричных задачах эти методы сильно потеснены различными вариантами алгоритма Ланцоша, а его несимметричные обобщения привлекают все больший интерес в несимметричном случае.

§ 14. Методы одновременных итераций

Идея одновременных итераций есть непосредственное обобщение идеи степенного метода. Обобщение состоит в том, что итерирование с помощью матрицы A единственного вектора x заменяется итерированием целого подпространства (отсюда другое название — «методы итерирования подпространства»). При определенных условиях полученная последовательность подпространств сходится к инвариантному подпространству матрицы A, отвечающему группе старших ее собственных значений. Сами условия имеют тот же смысл, что и условия сходимости степенного метода: они исключают неудачный выбор начального подпространства и обеспечивают то обстоятельство, что остальные собственные значения действительно меньше старших (по модулю).

Первый метод этого типа под названием *ступенчатые итерации* был предложен в 1957 г. Бауэром [57]. Годом позже Бауэр же предложил [58] модификацию метода, называемую *двусторонними итерациями* (bi-iteration), в которой одновременно вычисляются правые и левые собственные векторы.

Эффективность одновременных итераций существенно возрастает, если время от времени «поворачивать» базис итерируемого подпространства, приближая его к искомым собственным векторам. Эту идею — в приложении к симметричным матрицам — первым высказал, по-видимому, Дженнингс [126]. До совершенства она доведена в работах Рутисхаузера [168, 169], а составленная им для симметричного случая процедура ritzit (см. [43, алгоритм II. 9]) и сейчас остается образцом продуманной до мелочей программной реализацией метода. Хорошее и полное изложение принципов одновременной итерации для этого случая дано в гл. 14 книги [35].

С идейной точки зрения отказ от симметрии изменяет в одновременных итерациях немногое. Однако, как всегда, появляются технические сложности, связанные с возможным отсутствием базиса из собственных векторов или — в случае существования — с его неортогональностью.

В первой части параграфа мы описываем и обсуждаем общую схему методов одновременных итераций; здесь мы близко следуем плану статьи [186]. После этого рассматриваются некоторые конкретные методы, прежде всего метод Стьюарта.

Предположим, что собственные значения $n \times n$-матрицы A пронумерованы в порядке убывания модулей:

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \ldots \geqslant |\lambda_n|,$$

причем для выбранного натурального числа r справедливо строгое неравенство $|\lambda_r| > |\lambda_{r+1}|$. Обсуждение метода сильно упростится, если потребовать, чтобы матрица A имела полную систему собственных векторов $u_1, \ldots, u_n$. Подчеркнем, что *для применимости*, скажем, метода Стьюарта это требование излишне.

Введем матрицы

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \quad \Lambda_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_r), \quad \Lambda_2 = \operatorname{diag}(\lambda_{r+1}, \ldots, \lambda_n), \tag{14.1}$$

$$U = [u_1 | \ldots | u_n], \quad U_1 = [u_1 | \ldots | u_r], \quad U_2 = [u_{r+1} | \ldots | u_n].$$

Из определения собственных значений и векторов вытекают равенства

$$AU_j = U_j \Lambda_j, \qquad j = 1, 2. \tag{14.2}$$

В простейшем варианте одновременные итерации можно задать предписанием: выбрать r-мерное начальное подпространство $\mathscr{L}_0$ и построить последовательность подпространств $\mathscr{L}_k$ по формуле

$$\mathscr{L}_{k+1} = A\mathscr{L}_k, \qquad k = 0, 1, \ldots \tag{14.3}$$

Чтобы показать, почему такая последовательность в общем случае сходится к инвариантному подпространству $\mathscr{U}_r$ матрицы A, натянутому на векторы $u_1, \ldots, u_r$, целесообразно перейти с геометрического языка подпространств на матричный язык. Пусть $n \times r$-матрица P_0 базисная для подпространства $\mathscr{L}_0$. Определяя последовательность матриц $\{P_k\}$ соотношением

$$P_{k+1} = AP_k, \qquad k = 0, 1, \ldots, \tag{14.4}$$

видим, что каждое из подпространств $\mathscr{L}_k$ есть линейная оболочка столбцов одноименной матрицы P_k.

Наряду с P_k будем рассматривать матрицы

$$Y_k = U^{-1} P_k, \tag{14.5}$$

которые удобно представить в блочном виде с квадратным $r \times r$-блоком вверху:

$$Y_k = \begin{bmatrix} Y_k^{(1)} \\ Y_k^{(2)} \end{bmatrix}.$$

Равенство (14.5) означает, что

$$P_k = UY_k = U_1 Y_k^{(1)} + U_2 Y_k^{(2)}, \tag{14.6}$$

и если $Y_k^{(1)}$ не вырождена (для чего, как увидим, достаточно невырожденности $Y_0^{(1)}$), то

$$P_k \left(Y_k^{(1)}\right)^{-1} = U_1 + U_2 Y_k^{(2)} \left(Y_k^{(1)}\right)^{-1}. \tag{14.7}$$

В качестве системы образующих для подпространства $\mathscr{L}_k$ можно вместо столбцов P_k взять столбцы матрицы $P_k\left(Y_k^{(1)}\right)^{-1}$. Из (14.7) видно, что последовательность $\mathscr{L}_k$ сходится к $\mathscr{U}_r$, если $Y_k^{(2)}\left(Y_k^{(1)}\right)^{-1} \to 0$ при $k \to \infty$. Покажем, что это условие выполнено, если матрица $Y_0^{(1)}$ не вырождена. Смысл последнего требования мы обсудим чуть позже.

Используя (14.6) при $k=0$ и (14.2), имеем

$$P_k = A^k P_0 = A^k U_1 Y_0^{(1)} + A^k U_2 Y_0^{(2)} = U_1 \Lambda_1^k Y_0^{(1)} + U_2 \Lambda_2^k Y_0^{(2)}.$$

Сопоставляя с (14.6), получаем $Y_k^{(i)} = \Lambda_i^k Y_0^{(i)}$ $(i=1, 2)$. По предположению $Y_0^{(1)}$ не вырождена и $|\lambda_r| > |\lambda_{r+1}|$, поэтому $Y_k^{(1)}$ не вырождена при всех k. Далее

$$Z_k \equiv Y_k^{(2)}\left(Y_k^{(1)}\right)^{-1} = \Lambda_2^k Z_0 \Lambda_1^{-k}, \tag{14.8}$$

или поэлементно

$$z_{ij}^{(k)} = z_{ij}^{(0)} \left(\lambda_{i+r}/\lambda_j\right)^k, \qquad i=1, \ldots, n-r;\ j=1, \ldots, r. \tag{14.9}$$

Отсюда и следует, что $Z_k \to 0$.

Столбцы матрицы Y_k составлены из координат одноименных столбцов матрицы P_k в базисе $u_1, \ldots, u_n$. При $r=1$ условие невырожденности блока $Y_0^{(1)}$ сводится к тому, чтобы начальный вектор p_0 имел ненулевую компоненту по старшему собственному вектору u_1. Это — знакомое нам условие сходимости степенного метода. По-другому то же самое можно сформулировать как требование, чтобы p_0 не был ортогонален к левому собственному вектору v_1. Этот же смысл имеет условие $\det Y_1^{(0)} \neq 0$ при любом r. Действительно, строки матрицы U^{-1} с точностью до операции сопряжения являются левыми собственными векторами матрицы A, и $\det Y_0^{(1)} \neq 0$ означает, что левое инвариантное подпространство старших собственных векторов и начальное подпространство P_0 не ортогональны в смысле § 8, т. е. наибольший канонический угол между ними не равен $\pi/2$.

Равенство (14.9) показывает, что сходимость подпространств $\mathscr{L}_k$ к предельному инвариантному подпространству $\mathscr{U}_r$ имеет скорость $O\left(|\lambda_{r+1}/\lambda_r|^k\right)$. В то же время столбцы матрицы Z_k в (14.8) сходятся к нулю с разной скоростью: для m-го столбца скорость убывания определяется отношением $|\lambda_{r+1}/\lambda_m|$ $(m=1, \ldots, r)$. Возвращаясь к (14.7), заключаем, что столбцы матрицы $P_k\left(Y_k^{(1)}\right)^{-1}$ сходятся к собственным векторам со скоростью $O\left(|\lambda_{r+1}/\lambda_m|^k\right)$. Если в последовательности $|\lambda_1|, \ldots, |\lambda_r|$ разброс величин велик, то получается, что в подпространствах $\mathscr{L}_k$ некоторые направления сходятся к предельному положению значительно быстрее, чем подпространство в целом. Насколько существенно могут различаться скорости, видно из следующего

примера. Если $\lambda_1 = 1.0$, $\lambda_2 = 0.99$, $\lambda_r = 0.5$ и $\lambda_{r+1} = 0.49$, то скорость сходимости подпространств $\mathscr{L}_k$ характеризуется отношением $\lambda_{r+1}/\lambda_r = 0.98$, тогда как существует последовательность векторов $\{s_k\}$ ($s_k \in \mathscr{L}_k$), сходящаяся к u_1 со скоростью 0.49^k, и еще одна последовательность $\{t_k\}$, почти так же быстро сходящаяся к u_2. Этот же пример объясняет, почему использование одновременных итераций может быть рентабельным даже тогда, когда требуется вычислить всего лишь один собственный вектор u_1, т. е. в ситуации, когда можно применить и степенной метод. Хотя каждый шаг одновременных итераций требует гораздо большей арифметической работы, резкое уменьшение — по отношению к степенному методу — числа шагов может дать в итоге большую экономию. В данном случае степенные итерации сходятся очень медленно — со скоростью 0.99^k — в сравнении со скоростью сходимости последовательности $\{s_k\}$.

Формула (14.4) мало пригодна для практических вычислений. Даже в случае степенного метода $p_{k+1} = Ap_k$ ($k = 0, 1, \ldots$) она приводит к быстрому росту или уменьшению компонент векторов p_k. Чтобы избежать переполнений или обращения компонент в машинные нули, в степенном методе прибегают к нормировке векторов. При проведении одновременных итераций возникает новый фактор. Расписывая (14.4) по столбцам, получим

$$p_i^{(k+1)} = Ap_i^{(k)}, \qquad i = 1, \ldots, r.$$

Тем самым одновременные итерации (в форме (14.4)) протекают так, как если бы мы проводили степенной метод для каждого из столбцов начальной матрицы P_0. В общем случае все r степенных последовательностей сходятся к доминирующему собственному вектору u_1, и с ростом k столбцы матрицы P_k приближаются к строгой пропорциональности, теоретически оставаясь линейно независимыми. Нормировка столбцов не изменит ситуации, и, работая с конечноразрядными числами, мы в конечном счете придем к вырождению итерируемого подпространства — потере им первоначальной размерности.

Воспрепятствовать указанному явлению можно, перестраивая время от времени базис подпространства $\mathscr{L}_k$, для чего базисная матрица P_k заменяется другой матрицей, линейная независимость столбцов которой не вызывает сомнения. Общим матричным описанием такой перестройки будет формула

$$AP_k = P_{k+1} R_{k+1}, \tag{14.10}$$

где R_{k+1} — (невырожденная) матрица перехода от нового базиса P_{k+1} к старому базису AP_k. В разных методах одновременных итераций вид разложения в правой части (14.10) выбирается по-разному. Для ступенчатых итераций Бауэра P_{k+1} — нижняя трапецеидальная матрица с единицами на дигонали, R_{k+1} — верхняя треугольная матрица. В метода Стьюарта P_{k+1} — матрица с ортонормированными столбцами, а R_{k+1} — снова верхняя треугольная. Есть и другие возможности.

Шаги типа (14.10) обеспечивают восстановление практической линейной независимости столбцов базисной матрицы, но не гарантируют их быстрой сходимости. Так, в методах Бауэра и Стьюарта

формула перестройки базиса не изменяет направления первого вектора $p_1^{(k)}$. Рассмотренный выше пример показывает, что его сходимость к u_1 может быть очень медленной, хотя в $\mathscr{L}_k$ можно выделить последовательность, сходящуюся к u_1 гораздо быстрее.

Для ускорения сходимости столбцов матриц P_k в метод помимо *степенных шагов* (14.4) и *шагов-нормировок* (14.10) вводят шаги третьего типа. Их смысл в том, чтобы лучше использовать спектральную информацию, заключенную в подпространстве $\mathscr{L}_k$. К общей конструкции такого рода шагов можно прийти, рассматривая соотношение (14.7): $P_k(Y_k^{(1)})^{-1} = U_1 + U_2 Z_k$. Столбцы матрицы в правой части сходятся к соответствующим собственным векторам со скоростью $O(|\lambda_{r+1}/\lambda_m|^k)$ $(m = 1, ..., r)$. Если бы удалось найти приближение к $(Y_k^{(1)})^{-1}$, то, еще раз перестроив базис, мы получили бы систему векторов, сходящихся к своим пределам с максимально возможной для процесса скоростью.

К способу вычисления приближенной матрицы $(Y_k^{(1)})^{-1}$ придем с помощью следующего рассуждения. Пусть Q_k есть $n \times r$-матрица, для которой матрица $Q_k^* P_k$ не вырождена; норма $\| Q_k \|$ должна быть ограничена по k, в остальном Q_k пока произвольна. Из (14.7) имеем

$$Q_k^* P_k (Y_k^{(1)})^{-1} = Q_k^* U_1 + Q_k^* U_2 Z_k.$$

С другой стороны, из (14.7) и (14.2) следует равенство

$$Q_k^* A P_k (Y_k^{(1)})^{-1} = Q_k^* U_1 \Lambda_1 + Q_k^* U_2 \Lambda_2 Z_k.$$

Отсюда

$$(Q_k^* A P_k)(Y_k^{(1)})^{-1} - (Q_k^* P_k)(Y_k^{(1)})^{-1} \Lambda_1 = Q_k^* U_2 (\Lambda_2 Z_k - Z_k \Lambda_1).$$

Учитывая, что правая часть стремится к нулю, заключаем, что приближение к $(Y_k^{(1)})^{-1}$ можно искать как решение W_k задачи

$$B_k W_k = C_k W_k M_k, \tag{14.11}$$

где

$$B_k = Q_k^* A P_k, \quad C_k = Q_k^* P_k, \quad M_k = \operatorname{diag}(\mu_1, ..., \mu_r). \tag{14.12}$$

Формула (14.11) задает так называемую *обобщенную проблему собственных значений*, характеризуемую двумя матрицами (а не одной, как обычно). Впрочем, пользуясь невырожденностью матрицы C_k, при желании легко свести (14.11) к обычной задаче для матрицы $C_k^{-1} B_k$. Подчеркнем, что *собственные значения* $\mu_1, ..., \mu_r$ задачи (14.11) должны располагаться на диагонали матрицы M_k в порядке убывания модулей, что определяет и порядок следования столбцов матрицы W_k.

Вычислив матрицу W_k, подправляем базисную матрицу подпространства:

$$P_{k+1} = P_k W_k. \tag{14.13}$$

Шаги этого типа будем называть *поворотами*.

По поводу трудоемкости выполнения поворотов заметим следующее. Поскольку матрица AP_k в любом случае строится степенным

шагом, то дополнительная работа при повороте состоит: 1) в формировании матриц B_k и C_k; 2) в решении задачи (14.11); 3) в перемножении по формуле (14.13). Вычисление каждой из матриц B_k и C_k требует r^2n операций умножения, т. е. столько же, как и при ортогонализации столбцов матрицы AP_k; это же верно в отношении произведения P_kW_k. Число r в методах одновременной итерации не бывает большим, поэтому работой, затрачиваемой на решение задачи (14.11), можно пренебречь. Надо сказать к тому же, что повороты *не нужно* выполнять для каждого значения k. Так, в них обычно нет смысла до тех пор, пока не обозначилась заметно сходимость первого столбца (или первых столбцов) матрицы P_k. И в дальнейшем шаги-повороты производят лишь время от времени, заполняя промежутки между ними чистыми степенными шагами или степенными шагами с нормировками. Поскольку линейная независимость утрачивается в степенных шагах постепенно, то и нормировки выполняются с паузами. В целом методы одновременных итераций можно описать такой схемой (см. [187]):

Пока не все столбцы матрицы P сошлись
  Пока не обозначилась сходимость какого-либо столбца
    Пока столбцы матрицы P в достаточной степени линейно независимы
      Степенной шаг: $AP \to P$.
  Шаг-нормировка: $(AP)R^{-1} \to P$.
Шаг-поворот: $PW \to P$.
Проверка условия выхода

Многочисленные методы одновременных итераций различаются способом нормировки, выбором матрицы Q_k для поворота и метода решения задачи (14.11), а также условиями выхода. Ниже мы рассмотрим некоторые варианты.

Двусторонние итерации Бауэра. В этом методе наряду с последовательностью подпространств $\mathscr{L}_k$, порождаемых формулой (14.3), генерируется еще одна последовательность:

$$\mathscr{M}_{k+1} = A^* \mathscr{M}_k, \qquad k = 0, 1, \ldots$$

В действительности, разумеется, итерации ведутся с базисной $n \times r$-матрицей Q_k:

$$Q_{k+1} = A^* Q_k, \qquad k = 0, 1, \ldots \tag{14.14}$$

Начальные матрицы P_0 и Q_0 выбирают так, чтобы выполнялось равенство

$$Q_0^* P_0 = I_r,$$

т. е. чтобы их столбцы образовывали биортогональные системы. Это же соотношение биортогональности поддерживается на всех шагах, вследствие чего простые степенные формулы (14.4) и (14.14) приходится заменить на формулы

$$AP_k = P_{k+1} R_{k+1}, \qquad A^* Q_k = Q_{k+1} S_{k+1}. \tag{14.15}$$

Верхние треугольные матрицы R_{k+1}, S_{k+1} строятся процессом, аналогичным ортогонализации Грама—Шмидта, так, чтобы обеспечить соотношения двойственности

$$Q_k^* P_k = I_r. \tag{14.16}$$

Подробней об этом процессе биортогонализации мы поговорим в § 17.

Если старшие собственные значения матрицы A различны по модулю, то столбцы матрицы Q_k сходятся с ростом k к соответствующим *левым* (а столбцы P_k, как мы знаем,—к *правым*) собственным векторам. Исключения из этого правила могут быть связаны лишь с неудачным выбором начальных матриц. Без поворотов сходимость столбцов в обеих матричных последовательностях была бы очень медленной. Для поворота (см. формулы (14.12)) берется матрица Q_k второй бауэровской последовательности. В силу (14.16) задача (14.11) принимает вид

$$\left(Q_k^* A P_k\right) W_k = W_k M_k, \tag{14.17}$$

т. е. вид обычной спектральной задачи для матрицы $B_k = Q_k^* A P_k$. Матрица W_k собственных векторов используется для пересчета по формуле (14.13) базисной матрицы P_k, а матрица W_k^{-*}—для пересчета матрицы Q_k: $Q_{k+1} = Q_k W_k^{-*}$. В результате условие двойственности соблюдается.

Метод двусторонних итераций реализован в программе Клинта и Дженнингса [83]. Как видно из описания, метод существенно опирается на посылку, что вычисляемые старшие собственные векторы у матрицы действительно имеются. Нарушение этого условия или плохая обусловленность векторов в случае их существования приводят к тому, что и задача (14.17) обусловлена плохо. Указанного недостатка лишен метод Стьюарта.

Метод Стьюарта. По сравнению с методом двусторонних итераций (и даже с общей схемой одновременных итераций, как она изложена выше) здесь изменяется целевая установка: вместо того чтобы строить векторы, в пределе сходящиеся к доминирующим собственным векторам, ставится задача отыскания инвариантных подпространств, определяемых этими последними векторами. Такие *доминирующие инвариантные подпространства* существуют даже в том случае, когда нужного числа доминирующих собственных векторов нет. Пусть, в самом деле, унитарная матрица U трансформирует A к форме Шура, причем в последней собственные значения на диагонали упорядочены по убыванию модулей:

$$U^* A U = T. \tag{14.18}$$

Если при произвольном натуральном r $(1 \leqslant r < n)$ представить матрицы T и U в виде

$$U = \left[U_1 \mid U_2\right], \quad T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}, \tag{14.19а}$$

где U_1—$n \times r$-, а T_{11}—$r \times r$-матрица, то из (14.18) следует равенство

$$A U_1 = U_1 T_{11}, \tag{14.19б}$$

т. е. столбцы матрицы U_1 образуют базис (причем ортонормированный) инвариантного подпространства матрицы A, отвечающего старшим r собственным значениям (вспомним аналогичную выкладку в (13.18)).

Последовательность вложенных одно в другое подпространств $\mathcal{U}_1 \subset \mathcal{U}_2 \subset \ldots \subset \mathcal{U}_r$ размерностей $n_1, \ldots, n_r$ принято называть *флагом типа* $(n_1, \ldots, n_r)$. *Базисом флага* назовем такую линейно независимую систему из n_r векторов, что первые ее n_k членов $(1 \leqslant k \leqslant r)$ составляют базис подпространства $\mathcal{U}_k$. Задачу метода Стьюарта можно теперь сформулировать так: найти ортонормированный базис доминирующего флага типа $(1, \ldots, r)$. Как только эта цель будет достигнута, вычисление старших собственных векторов не составит труда: если z — собственный вектор матрицы T_{11}, то $x = U_1 z$ — собственный вектор для A с тем же собственным значением.

Несмотря на изменение задачи, метод Стьюарта можно рассматривать как частную реализацию общей схемы одновременных итераций. Для нормировок (14.10) используется унитарно-треугольное разложение; при поворотах полагаем $Q_k = P_k$, и (14.11) сводится к задаче

$$(P_k^* A P_k) W_k = W_k \Delta_k. \tag{14.20}$$

Здесь, однако, вместо диагональной матрицы M_k разыскивается верхняя треугольная матрица Δ_k — форма Шура матрицы $B_k = P_k^* A P_k$; эта матрица вместе с унитарной трансформирующей матрицей W_k может быть вычислена посредством QR-алгоритма (см. в связи с этим п. 2 дополнений к § 11).

Отметим, что доминирующий флаг однозначно определен лишь в случае, если $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_r| > |\lambda_{r+1}|$. В такой ситуации, по существу, однозначно определен и (ортонормированный) базис доминирующего флага: свобода остается только в выборе для каждого базисного вектора скалярного множителя с модулем 1.

Если у матрицы A есть группа равных по модулю собственных значений

$$|\lambda_1| \geqslant \ldots \geqslant |\lambda_{l-1}| > |\lambda_l| = |\lambda_{l+1}| = \ldots = |\lambda_m| > |\lambda_{m+1}| \geqslant \ldots \geqslant |\lambda_r|,$$

то на участке от l до $m-1$ имеется произвол в выборе подпространств $\mathcal{U}_k$, связанный с возможностью разных упорядочений группы $\lambda_l, \ldots, \lambda_m$. При наличии кратных собственных значений произвол может еще увеличиться. Естественно, таков же произвол при выборе базиса флага. В то же время базисные векторы, отвечающие изолированным (по модулю) собственным значениям, по-прежнему однозначны (в указанном выше смысле).

В случае, когда $r = n$, указанная неоднозначность ответственна за множественность форм Шура данной матрицы. Но подобно тому, как эта множественность не мешает определению собственных значений и векторов посредством QR-алгоритма, неоднозначность доминирующего флага не существенна в методе Стьюарта.

Теоретическое исследование метода проведено в [187]. Главным его результатом является

Теорема 14.1. *Пусть* $|\lambda_r|>|\lambda_{r+1}|$, *и пусть для некоторого* l $(1\leqslant l\leqslant r)$ *выполняется неравенство* $|\lambda_l|>|\lambda_{l+1}|$. *Тогда* l*-мерное доминирующее подпространство* $\mathcal{U}_l$ *матрицы* A *определено однозначно, и при обычном ограничении на начальную матрицу* P_0 *линейная оболочка первых* l *столбцов матрицы* P_k *сходится к* $\mathcal{U}_l$ *со скоростью* $O_\varepsilon((|\lambda_{r+1}|/|\lambda_l|)^k)$.

Замечание 14.2. Символ $O_\varepsilon((|\lambda_{r+1}|/|\lambda_l|)^k)$ означает асимптотическое убывание не медленнее геометрической прогрессии со знаменателем $(|\lambda_{r+1}|+\varepsilon)(|\lambda_l|^{-1}+\varepsilon)$ при произвольном $\varepsilon>0$. Он учитывает возможное присутствие жордановых клеток с собственными значениями λ_l и λ_{r+1}. Если матрица A диагонализуема, то можно пользоваться обычным символом O.

Следствие 14.3. *Если в условиях теоремы* 14.1 *дополнительно потребовать, чтобы выполнялось неравенство* $|\lambda_{l-1}|>|\lambda_l|$, *то* l*-й столбец матрицы* P_k *сходится к* l*-му базисному вектору доминирующего флага со скоростью* $O_\varepsilon((|\lambda_{r+1}/\lambda_l|)^k)$.

Как и в QR-алгоритме, сходимость нужно понимать с точностью до скалярного множителя, по модулю равного 1. С такой же скоростью диагональный элемент (l, l) матрицы Δ_k сходится к λ_l. В случае нормальной матрицы A скорость сходимости диагонального элемента вдвое выше скорости сходимости вектора $p_l^{(k)}$ и составляет $O((|\lambda_{r+1}/\lambda_l|)^{2k})$.

Замечание 14.4. Условие $|\lambda_r|>|\lambda_{r+1}|$ является необходимым, как показывает рассмотрение следующего примера [187]. Пусть $l=1$, $r=2$,

$$A=\begin{bmatrix}2&0&0\\0&1&-6\\0&0&-1\end{bmatrix},\qquad P_0=\begin{bmatrix}1&0\\1&1\\1&-1\end{bmatrix}.$$

Тогда

$$A^{2k}P_0=\begin{bmatrix}2^{2k}&0\\1&1\\1&-1\end{bmatrix}.$$

Столбцы этой матрицы ортогональны. Нормируя их, можем затем построить матрицу B_{2k} (см. (14.12)):

$$B_{2k}=\begin{bmatrix}\dfrac{2-6\varepsilon_k^2}{1+2\varepsilon_k^2} & \dfrac{8\varepsilon_k}{\sqrt{2(1+2\varepsilon_k^2)}}\\ \dfrac{-4\varepsilon_k}{\sqrt{2(1+2\varepsilon_k^2)}} & 3\end{bmatrix},\qquad \varepsilon_k=2^{-2k}.$$

Большее собственное значение матрицы B_{2k} сходится к 3, а соответствующий собственный вектор имеет пределом $(0, 1)^T$. Последнее означает, что первые столбцы четных матриц P_k будут сходиться к вектору $t=(0, 1/\sqrt{2}, -1/\sqrt{2})^T$. Но у матрицы A нет собственного значения $\lambda=3$, а вектор t нисколько не похож на ее старший собственный вектор $(1, 0, 0)^T$.

Левыми собственными векторами матрицы A будут $v_1=(1, 0, 0)^T$, $v_2=(0, 1, -0)^T$ и $v_3=(0, 3, 1)^T$. Если $V_2=[v_1 \mid v_2]$ либо $V_2=[v_1 \mid v_3]$, то матрица $V_2^T P_0$ не вырождена. Следовательно, причиной «неправильной» сходимости является не ошибочный выбор начальной матрицы, а именно нарушение условия $|\lambda_2|>|\lambda_3|$.

Поговорим теперь о некоторых практических деталях метода Стьюарта. Относительно частоты, с какой необходимо производить шаги-нормировки, в [187] приведены такие соображения. Когда процесс одновременных итераций начинает сходиться, выполняется приближенное равенство (опускаем индекс k текущей итерации)

$$AP \approx P\Delta$$

(см. (14.19) и (14.20)). После m последовательных шагов получаем $A^m P \approx P\Delta^m$, и потому

$$\| A^m P \| \lesssim \| \Delta \|^m$$

(символ $\lesssim$ указывает приближенное неравенство). Так как, с другой стороны, $A^m P\Delta^{-m} \approx P$, то столбцы матрицы $A^m P\Delta^{-m}$ приблизительно ортонормированы. Если R — треугольная матрица, которая *точно* ортонормирует столбцы произведения $A^m PR$, то $\| R \| \lesssim \| \Delta^{-1} \|^m$. Анализ ошибок округлений при ортонормализации показывает, что число верных десятичных (для определенности) разрядов, теряемых в результатах, может доходить до $\lg(\operatorname{cond} R) \approx m \lg(\operatorname{cond} \Delta)$. Если s — максимальное число разрядов, потеря которых еще допускается пользователем, то число m, определяемое из неравенства

$$m \leqslant s/\lg(\operatorname{cond} \Delta), \tag{14.21}$$

задает частоту нормировок: шаг типа (14.10) нужно выполнять после каждых $m-1$ степенных шагов. Оценку нормы обратной матрицы Δ^{-1}, необходимую для пользования неравенством (14.21), можно получить приемом типа применяемого в подпрограммах пакета LINPACK (см. [19, § 9]). Однако значение $\| \Delta^{-1} \|$ нетрудно найти и прямым вычислением, поскольку порядок r треугольной матрицы Δ невелик.

Наиболее естественное место для проверки условия выхода — сразу вслед за шагом-поворотом, как это и показано в нашей общей схеме. В основу проверки могут быть заложены различные критерии. Можно следить за стабилизацией диагональных элементов в треугольных матрицах Δ_k: их предельными значениями являются старшие собственные числа матрицы A. Если собственные числа плохо обусловлены, они могут заметно изменяться от итерации к итерации, как бы долго ни шел процесс. В этом случае предпочтительней условие выхода, опирающееся на величину невязок. Последние нужно понимать в следующем обобщенном смысле. Сходимость первых l столбцов (они предполагаются ортонормированными) матрицы P_k можно характеризовать посредством нормы *матрицы-невязки*

$$R_1^{(k)} = AP_1^{(k)} - P_1^{(k)}\Delta_{11}^{(k)}, \tag{14.22}$$

где $P_1^{(k)}$ — $n \times l$-матрица, образованная первыми столбцами P, а $\Delta_{11}^{(k)}$ — ведущая главная подматрица порядка l в Δ_k. Хотя малость нормы $\| R_1^{(k)} \|$ не гарантирует хорошего качества приближений, так как оно определяется еще и обусловленностью спектральной задачи, все же (14.22) означает, что $P_1^{(k)}$ — базисная матрица *точного* инвариантного подпространства матрицы $\tilde{A}$, близкой к A:

$$\left[A - R_1^{(k)}\left(P_1^{(k)}\right)^*\right] P_1^{(k)} = P_1^{(k)} \Delta_{11}^{(k)}.$$

Спектральная (или евклидова) норма возмущения $F_k = R_1^{(k)}\left(P_1^{(k)}\right)^*$ равна норме самой матрицы-невязки $R_1^{(k)}$.

Отметим, что мы говорим о сходимости группы *первых* столбцов матрицы P по той причине, что в силу теоремы 14.1 раньше всего сходятся именно эти столбцы.

Как часто следует выполнять повороты? Можно рассуждать следующим образом. Невязки, порождаемые столбцами $p_l^{(k)}$, убывают приблизительно с линейными скоростями. Если сравнить величины норм невязок после двух последовательных поворотов, то можно оценить знаменатели прогрессий сходимости и число (степенных) шагов, необходимых для достижения заданного уровня малости невязок. После такого количества шагов можно произвести новый поворот и сравнить предсказание с реальной величиной невязок. Этот прием, возможно, придется применить неоднократно как из-за неточности прогноза, так и из-за того, что невязки разных столбцов сходятся с различными скоростями.

Различие скоростей сходимости можно использовать для сокращения вычислительной работы. Пусть, в самом деле, заданный критерий (стабилизации или малости невязок) для первых l столбцов матрицы P_k выполнен, а для остальных — нет. Последующие степенные шаги можно применять только к этим столбцам, так что матрицы $P_{k+\nu}$ имеют вид

$$P_k = [P_1 \mid P_2], \quad P_{k+\nu} = [P_1 \mid A^\nu P_2], \qquad \nu = 1, 2, \ldots$$

Когда производится ортогонализация, ее нужно начинать с $(l+1)$-го столбца, потому что столбцы подматрицы P_1 и так ортогональны. Наконец, при повороте, работая с уже ортогонализованной матрицей $P_\mu = [P_1 \mid \tilde{P}_2]$, мы можем в матрице

$$B_\mu = P_\mu^* A P_\mu = \begin{bmatrix} \Delta_{11} & P_1^* A \tilde{P}_2 \\ \tilde{P}_2^* A P_1 & \tilde{P}_2^* A \tilde{P}_2 \end{bmatrix}$$

пренебречь блоком $\tilde{P}_2^* A P_1$, поскольку (см. (14.22))

$$\| \tilde{P}_2^* A P_1 \|_2 = \| \tilde{P}_2^* \left(R_1^{(\mu)} + P_1 \Delta_{11}\right) \|_2 = \| R_1^{(\mu)} \|_2$$

(можно было бы взять и евклидову норму). Тогда приводить к форме Шура нужно лишь подматрицу $\tilde{P}_2^* A \tilde{P}_2$.

Если A — вещественная матрица, то все вычисления можно проводить в вещественной арифметике, начиная с вещественной же матрицы P_0. В этом случае на шагах-поворотах строится *вещественная* форма Шура матриц B_k.

Сделаем в заключение несколько замечаний, относящихся к произвольному методу одновременных итераций. Общую схему этой группы методов мы излагали в форме, отвечающей степенным итерациям; в результате определялись доминирующие собственные значения и векторы матрицы A. Если A не вырождена, то, заменяя в схеме A на A^{-1}, придем к *одновременным обратным итерациям*, позволяющим вычислить младшие собственные числа. Наконец, имея приближенное значение α для искомого собственного числа, можем найти группу собственных чисел, ближайших к α, и соответствующие собственные векторы посредством одновременных итераций с матрицей $(A-\alpha I)^{-1}$.

Удачный выбор начальной матрицы P_0 может ускорить сходимость одновременных итераций. В практических задачах нередко выбирают P_0, исходя из представлений о виде или характере поведения искомых собственных функций. Однако рецепта, пригодного на любой случай, не существует. В § 19 будет рассмотрено видоизменение метода одновременных итераций, равносильное неоднократному его проведению с последовательно улучшаемыми начальными приближениями.

Скорость сходимости одновременных итераций еще больше зависит от выбора числа r итерируемых векторов. Понятно, что оно должно быть не меньше требуемого числа s доминирующих собственных пар, а сверху ограничено необходимостью хранить $n \times r$-матрицу P. Однако в этих пределах еще остается некоторая свобода для r, и опять-таки трудно дать обоснованные рекомендации относительно наилучшего выбора: он неизбежно зависит от распределения собственных значений, обычно заранее неизвестного. Часто пользуются каким-нибудь эмпирическим правилом типа $r=\min\{s+8, 2s\}$ (см. [50]).

ДОПОЛНЕНИЯ К § 14

1. В [127] наряду с описанием двусторонних итераций рассматривается еще один алгоритм, который авторы называют *косыми итерациями* (lopsided iteration). Этот алгоритм по структуре схож с методом Стьюарта и отличается от него главным образом тем, что шаги-нормировки как таковые не производятся. Нормировки осуществляются неявно в ходе шагов-поворотов. Подробное обсуждение программной реализации алгоритма дано в [189]. Нужно сказать, что в подходе авторов сохранен основной дефект двусторонних итераций, а именно попытка непосредственно найти доминирующие собственные векторы (только правые в отличие от бауэровского процесса). Уже отмечалось, что полной системы таких векторов может не быть или же векторы могут быть очень плохо обусловлены, тогда как задача построения доминирующего флага, решаемая методом Стьюарта, всегда имеет смысл.

2. Изложение теории QR-алгоритма с позиций одновременной итерации предпринято в [205]. Согласно (12.12), при любом l $(1 \leqslant l \leqslant n)$ первые l столбцов матрицы $\tilde{Q}_k = Q_1 \ldots Q_k$ составляют ортонормированный базис подпространства $\mathcal{L}_k^{(l)} = A^k \mathcal{L}_0^{(l)}$, где в качестве $\mathcal{L}_0^{(l)}$ берется оболочка первых l координатных векторов $e_1, \ldots, e_l$. Если взять обычные достаточные условия сходимости QR-алгоритма (см. § 12) — все собственные значения матрицы A различны по модулю и в матрице левых собственных векторов ведущие главные

миноры отличны от нуля,— то, как следует из основного текста настоящего параграфа, эти условия обеспечивают сходимость каждой последовательности $\{\mathscr{L}_k^{(l)}\}$ $(l=1, ..., n)$. Тем самым столбцы матриц $\tilde{Q}_k$ сходятся (с точностью до множителей, по модулю равных 1), а потому сходятся к треугольной форме и матрицы, генерируемые QR-процессом.

§ 15. Метод Ланцоша и его обобщения

Алгоритм Ланцоша лишь сравнительно недавно стал рассматриваться как метод решения разреженных спектральных задач. В первой публикации 1950 г. [135] метод предназначался для трехдиагонализации симметричных (эрмитовых) матриц. В то время о разреженности никто не говорил; большими же считались задачи порядка нескольких десятков.

Хотя в данной книге не излагаются методы для симметричных матриц, для алгоритма Ланцоша нам придется сделать исключение: это облегчит понимание обобщений метода на несимметричный случай.

Итак, пусть A — эрмитова матрица порядка n; ее спектр $\lambda_1, ..., \lambda_n$ будем пока считать простым. Возьмем любой вектор x_0, в разложении которого по собственным векторам $u_1, ..., u_n$ матрицы A *все коэффициенты ненулевые*:

$$x_0 = \xi_1 u_1 + ... + \xi_n u_n. \tag{15.1}$$

Случайно выбранный вектор x_0 с вероятностью 1 удовлетворяет этому условию.

Построим последовательность векторов

$$x_0, \quad Ax_0, ..., A^{m-1}x_0. \tag{15.2}$$

При любом m $(1 \leqslant m \leqslant n-1)$ линейная оболочка $\mathscr{K}^m(x_0)$ системы (15.2) называется *крыловским подпространством*, а сама система — *степенной последовательностью* или *последовательностью Крылова*, порождаемой вектором x_0. При сделанном относительно x_0 предположении последовательность Крылова длины n $(m=n)$ образует базис в $\mathbf{C}^n$. Действительно, равенству (15.1) отвечает разложение

$$A^k x_0 = \xi_1 \lambda_1^k u_1 + \xi_2 \lambda_2^k u_2 + .. + \xi_n \lambda_n^k u_n, \qquad 1 \leqslant k \leqslant n-1,$$

а координаты векторов крыловской последовательности в базисе $u_1, ..., u_n$ составляют матрицу

$$\begin{bmatrix} \xi_1 & ... & \xi_n \\ \xi_1\lambda_1 & ... & \xi_n\lambda_n \\ . & . & . \\ \xi_1\lambda_1^{n-1} & ... & \xi_n\lambda_n^{n-1} \end{bmatrix},$$

которая с точностью до ненулевых столбцевых множителей $\xi_1, ..., \xi_n$ является матрицей Вандермонда для различных чисел $\lambda_1, ..., \lambda_n$, а потому не вырождена (см. хотя бы [27, с. 50]).

Разумеется, всякая крыловская последовательность длины, большей n, линейно зависима.

Метод Ланцоша основывается на следующих двух наблюдениях. Во-первых, если отождествить матрицу A с линейным оператором

$\mathscr{A}$, действующим в $\mathbf{C}^n$, то с помощью крыловской последовательности можно построить ортонормированный базис, в котором этот оператор имеет трехдиагональную матрицу. Иными словами, можно найти (унитарное) подобие, приводящее A к трехдиагональной форме. В качестве базиса нужно взять результат применения к последовательности Крылова $x_0, Ax_0, \ldots, A^{n-1}x_0$ процесса ортонормализации Грамма—Шмидта. В самом деле, пусть $q_1, \ldots, q_n$—найденный ортонормированный базис. Процесс Грама—Шмидта устроен таким образом, что при любом m $(1 \leqslant m \leqslant n-1)$ векторы $q_1, \ldots, q_m$ образуют базис линейной оболочки первых m векторов ортогонализуемой системы, т. е. базис крыловского подпространства $\mathscr{K}^m(x_0)$. Пусть T—матрица линейного оператора $\mathscr{A}$ в базисе $q_1, \ldots, q_n$. Ее элементом t_{ij} является скалярное произведение $(\mathscr{A}q_j, q_i)$, и при $i>j+1$ имеем

$$t_{ij}=(\mathscr{A}q_j, q_i)=0, \tag{15.3}$$

поскольку $\mathscr{A}q_j \in \mathscr{K}^{j+1}(x_0)$, а вектор q_i ортогонален к $\mathscr{K}^{j+1}(x_0)$. Матрица T, будучи матрицей самосопряженного оператора в ортонормированном базисе, сама самосопряжена, поэтому (15.3) означает, что T трехдиагональная.

Равенства (15.3) заключают в себе и другое важнейшее свойство алгоритма: векторы $q_1, \ldots, q_n$, называемые *векторами Ланцоша*, можно строить по простым рекуррентным формулам, в которых в отличие от общего случая ортогонализации число членов не зависит от номера шага. Действительно, пусть векторы $q_1, \ldots, q_m$ уже вычислены. На $(m+1)$-м шаге процесса Грама—Шмидта определяется вектор q_{m+1}, такой, что $q_{m+1} \in \mathscr{K}^{m+1}(x_0)$ и $q_{m+1} \perp \mathscr{K}^m(x_0)$. Будем искать этот вектор с помощью соотношения

$$r_{m+1,m}q_{m+1}=Aq_m-r_{1m}q_1-r_{2m}q_2-\ldots-r_{mm}q_m, \tag{15.4}$$

представляющего собой видоизмененную форму обычного правила Грама—Шмидта. Коэффициент $r_{m+1,m}$ обеспечивает нормировку вектора q_{m+1}, и, следовательно, его модуль равен длине вектора в правой части соотношения (15.4). Условимся выбирать для этого коэффициента неотрицательное значение. Из условий ортогональности $q_{m+1} \perp q_i$ $(i=1, \ldots, m)$ получаем

$$(q_{m+1}, q_i)=0=(Aq_m, q_i)-r_{im},$$

и при $m>i+1$

$$r_{im}=(Aq_m, q_i)=(q_m, Aq_i)=0,$$

так как $Aq_i \in \mathscr{K}^{i+1}(x_0)$. В правой части (15.4) остается лишь три члена:

$$r_{m+1,m}q_{m+1}=Aq_m-r_{mm}q_m-r_{m-1,m}q_{m-1}. \tag{15.5}$$

На самом деле коэффициенты $r_{m+1,m}$, r_{mm}, $r_{m-1,m}$ суть элементы трехдиагональной матрицы T, вычисляемой алгоритмом Ланцоша. Действительно, перепишем (15.5) как

$$Aq_m=r_{m-1,m}q_{m-1}+r_{mm}q_m+r_{m+1,m}q_{m+1}. \tag{15.6}$$

Заметим, что при $m=n$ формула (15.5) порождает нулевой вектор — иначе был бы найден $(n+1)$-й вектор ортонормированной системы в $\mathbf{C}^n$. Следовательно, $r_{n+1,n}=0$, и равенства (15.6), взятые при $m=1, \ldots, n$, в совокупности дают матричное соотношение

$$AQ=QT, \tag{15.7}$$

где $Q=[q_1 \mid q_2 \mid \ldots \mid q_n]$,

$$T=\begin{bmatrix} r_{11} & r_{12} & & & & \\ r_{21} & r_{22} & r_{23} & & 0 & \\ & r_{32} & r_{33} & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \ddots & \ddots & r_{n-1,n} \\ & & & & r_{n,n-1} & r_{nn} \end{bmatrix}.$$

Нет смысла сохранять двухиндексные обозначения для элементов трехдиагональной матрицы. Будем писать α_i и β_{i+1} для элементов в позициях (i, i) и $(i, i+1)$. Учитывая самосопряженность матрицы T, получаем

$$T=\begin{bmatrix} \alpha_1 & \beta_2 & & & & \\ \beta_2 & \alpha_2 & \beta_3 & & 0 & \\ & \beta_3 & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \ddots & \ddots & \beta_n \\ & & & & \beta_n & \alpha_n \end{bmatrix}. \tag{15.8}$$

Рекуррентной формуле (15.5) можно теперь придать окончательный вид:

$$\begin{aligned} &\beta_{m+1} q_{m+1}=Aq_m-\alpha_m q_m-\beta_m q_{m-1}, \\ &r_{m+1} \equiv Aq_m-\alpha_m q_m-\beta_m q_{m-1}, \\ &\alpha_m=(Aq_m, q_m),\ \beta_{m+1}=\| r_{m+1} \|_2. \end{aligned} \tag{15.9}$$

При $m=1$ нужно положить $q_0=0$. Вектор q_1 есть нормированный начальный вектор x_0.

Что произойдет, если в разложении (15.1) есть нулевые коэффициенты или же матрица A имеет кратные собственные значения? В обоих этих случаях число l членов разложения

$$x_0=\xi_1 u_1+\xi_2 u_2+\ldots+\xi_l u_l$$

строго меньше, чем n, а векторы $u_1, \ldots, u_l$ можно считать относящимися к *различным* собственным значениям. Будем называть l *индексом* вектора x_0 относительно матрицы A. Вектор $A^l x_0$ оказывается линейной комбинацией предыдущих векторов крыловской последовательности, до этого момента линейно независимой. Вычисления

по формулам (15.9) приведут к нулевому значению β_{l+1}, и вектор q_{l+1} определить не удастся. В этой ситуации, называемой *обрывом* или *вырождением* процесса Ланцоша, поступают так: берут нормированный вектор $\tilde{q}_{l+1}$, ортогональный к уже вычисленным векторам $q_1, \ldots, q_l$, и, рассматривая его как начальный, снова применяют формулы (15.9). Получаемые теперь векторы $\tilde{q}_{l+2}, \ldots, \tilde{q}_t$ будут автоматически ортогональны первой серии $q_1, \ldots, q_l$. Действительно, крыловское подпространство $\mathscr{K}^l(x_0)$ инвариантно относительно A, и это же в силу самосопряженности A верно для его ортогонального дополнения $\mathscr{L}^l$. Но $\tilde{q}_{l+1} \in \mathscr{L}^l$ по построению, поэтому и остальные векторы новой крыловской последовательности также лежат в $\mathscr{L}_l$. Если $t<n$, т. е. имеет место новый обрыв, то описанный прием применяют еще раз, ортогонализуя $\tilde{\tilde{q}}_{t+1}$ и к векторам q_i, и к векторам $\tilde{q}_j$. В конечном счете будет получен ортонормированный базис пространства $\mathbf{C}^n$. Построенную из векторов q_i, $\tilde{q}_j, \ldots$ унитарную матрицу будем, как и раньше, обозначать через Q. Нетрудно проверить, что по-прежнему действует равенство (15.7), где T— трехдиагональная матрица, составленная из коэффициентов α_i, β_i ланцошевых рекурсий (15.9). Отличие от предыдущего случая будет в том, что теперь T разложима. Это только облегчает последующее вычисление спектра.

Подведем итоги. Метод Ланцоша во всех случаях позволяет найти трехдиагональную форму T заданной эрмитовой матрицы A и унитарную матрицу Q, трансформирующую A к этой форме. В каждый момент процесса в оперативной памяти достаточно держать пять массивов длины n; два из них хранят коэффициенты α_i, β_i, три других нужны для рекуррентного счета по формулам (15.9). Впрочем, при необходимости можно обойтись и четырьмя массивами [151]. Если собственные векторы матрицы A вычислять не требуется, то векторы Ланцоша, кроме двух последних, не нужно хранить вообще; в противном случае их можно после того, как они использованы для рекурсии, вынести во внешнюю память. Вычисление собственных значений и векторов симметричной трехдиагональной матрицы T— сравнительно простое дело и может быть осуществлено несколькими эффективными методами, например QR-алгоритмом в сочетании с обратными итерациями. Если θ, s— собственная пара матрицы T, то θ, Qs— собственная пара для матрицы A. Внося поочередно векторы Ланцоша из внешней памяти, можно построить произведение Qs.

Математическая элегантность алгоритма Ланцоша обеспечила ему в момент появления горячий прием вычислителей. Вскоре появились обобщения метода на случай неэрмитовых матриц. Так как оба свойства процесса Ланцоша — трехдиагональность преобразованной матрицы и ортонормированность системы векторов $q_1, \ldots, q_n$— сохранить не удается, обобщения строились по-разному в зависимости от того, удержание какого свойства считалось более важным.

В методе Арнольди [54] выбор делается в пользу ортонормированности. Как и прежде, крыловская последовательность подвергается процессу ортогонализации, проводимому в следующей форме.

Пусть q_1 — нормированный начальный вектор. Тогда для $m=1, 2, \ldots$ вектор q_{m+1} разыскивается в виде

$$h_{m+1,m}q_{m+1}=Aq_m-h_{1m}q_1-\ldots-h_{mm}q_m, \tag{15.10}$$

где коэффициенты h_{im} $(1\leqslant i\leqslant m)$ определяются из условий ортогональности к ранее вычисленным векторам $q_1, \ldots, q_m$, а коэффициент $h_{m+1,m}$ — так, чтобы вектор q_{m+1} был нормированным. Отсюда получаем

$$\begin{aligned} &h_{im}=(Aq_m, q_i), \qquad i=1, \ldots, m, \\ &h_{m+1,m}=\| Aq_m-h_{1m}q_1-\ldots-h_{mm}q_m \|_2. \end{aligned} \tag{15.11}$$

Обращение в нуль коэффициента $h_{m+1,m}$ — *обрыв* процесса Арнольди — означает, что крыловское подпространство, натянутое на q_1, $Aq_1, \ldots, A^{m-1}q_1$ (т. е. на $q_1, \ldots, q_m$), является инвариантным для матрицы A. Так как (15.10) в этом случае не определяет вектора q_{m+1}, то для продолжения процесса берется произвольный нормированный вектор, ортогональный к $q_1, \ldots, q_m$. Однако до следующего обрыва вычисления по-прежнему ведутся по формулам (15.10), (15.11).

В совокупности соотношения (15.10) для $m=1, \ldots, n$ эквивалентны матричному равенству

$$AQ=QH, \tag{15.12}$$

где Q — снова унитарная матрица, образованная (по столбцам) векторами $q_1, \ldots, q_n$, а H — хессенбергова матрица

$$H=\begin{bmatrix} h_{11} & h_{12} & \ldots & h_{1,n-1} & h_{1n} \\ h_{21} & h_{22} & \ldots & h_{2,n-1} & h_{2n} \\ & h_{32} & \ldots & h_{3,n-1} & h_{3n} \\ & 0 & \ddots & \ddots & \ddots \\ & & & h_{n,n-1} & h_{nn} \end{bmatrix}. \tag{15.13}$$

Таким образом, алгоритм Арнольди был первым из методов унитарного приведения матрицы общего вида к форме Хессенберга.

Другое неэрмитово обобщение предложено самим Ланцошем. Ценой отказа от ортогональности вычисляемых векторов здесь сохраняется трехдиагональная форма матрицы T. Для этого приходится рассматривать наряду с последовательностью (15.2) еще одну крыловскую последовательность:

$$y_0, \quad A^*y_0, \ldots, (A^*)^m y_0, \ldots,$$

порождаемую сопряженной матрицей A^*. От начальных векторов x_0, y_0 потребуем выполнения равенства $(x_0, y_0)=1$. Принимая $v_1\equiv x_0$, $w_1\equiv y_0$, строим затем *биортогональные* последовательности *правых* и *левых векторов Ланцоша* $v_1, v_2, \ldots$ и $w_1, w_2, \ldots$ Именно векторы v_2 и w_2 ищем в виде

$$\beta_2 v_2=Av_1-\alpha_1 v_1, \qquad \bar{\gamma}_2 w_2=A^*w_1-\bar{\alpha}_1 w_1$$

так, чтобы выполнялись условия $v_2 \perp w_1$, $w_2 \perp v_1$ и $(v_2, w_2) = 1$. Первые два условия будут удовлетворены, если положить $\alpha_1 = (Av_1, w_1)$. Если далее обозначить $r_2 = Av_1 - \alpha_1 v_1$, $s_2 = A^* w_1 - \bar{\alpha}_1 w_1$, то последнее условие дает

$$\beta_2 \gamma_2 = (r_2, s_2) \equiv \omega_2.$$

Следовательно, значение одного из коэффициентов β_2, γ_2 может быть произвольным.

К шагу с номером m вычислены биортогональные последовательности $v_1, \ldots, v_m$ и $w_1, \ldots, w_m$. Векторы v_{m+1} и w_{m+1} ищем в виде

$$\begin{aligned} \beta_{m+1} v_{m+1} &= Av_m - \alpha_m v_m - \gamma_m v_{m-1} \equiv r_{m+1}, \\ \bar{\gamma}_{m+1} w_{m+1} &= A^* w_m - \bar{\alpha}_m w_m - \bar{\beta}_m w_{m-1} \equiv s_{m+1}. \end{aligned} \tag{15.14}$$

Из условий ортогональности $(v_{m+1}, w_m) = (v_m, w_{m+1}) = 0$ получаем

$$\alpha_m = (Av_m, w_m).$$

Ортогональность r_{m+1} и w_{m-1} обеспечена автоматически. Действительно,

$$\begin{aligned} (r_{m+1}, w_{m-1}) &= (Av_m, w_{m-1}) - \gamma_m = (v_m, A^* w_{m-1}) - \gamma_m = \\ &= (v_m, \bar{\gamma}_m w_m + \bar{\alpha}_{m-1} w_{m-1} + \bar{\beta}_{m-1} w_{m-2}) - \gamma_m = \gamma_m (v_m, w_m) - \gamma_m = 0. \end{aligned}$$

Аналогично доказывается, что $s_{m+1} \perp v_{m-1}$. Ортогональность к более ранним векторам устанавливается так:

$$\begin{aligned} (r_{m+1}, w_j) &= (Av_m, w_j) = (v_m, A^* w_j) = \\ &= \gamma_{j+1} (v_m, w_{j+1}) + \alpha_j (v_m, w_j) + \beta_j (v_m, w_{j-1}) = 0, \qquad j < m-1. \end{aligned}$$

Подобным же образом будем иметь $(s_{m+1}, v_j) = 0$, если $j < m-1$. Условие нормировки $(v_{m+1}, w_{m+1}) = 1$ приводит к равенству

$$\beta_{m+1} \gamma_{m+1} = (r_{m+1}, s_{m+1}) \equiv \omega_{m+1}.$$

Если $\omega_{m+1} \neq 0$, то, выбирая для одного из коэффициентов β_{m+1}, γ_{m+1} произвольное ненулевое значение, однозначно определяем другой коэффициент. После этого формулы (15.14) дают векторы v_{m+1} и w_{m+1}.

Если $\omega_m \neq 0$ для $m = 2, \ldots, n$, то будут построены биортогональные базисы $v_1, \ldots, v_n$ и $w_1, \ldots, w_n$, а равенства (15.14) вместе охватываются матричным соотношением

$$AV = VT. \tag{15.15}$$

Здесь $V = [v_1 | \ldots | v_n]$, а T — трехдиагональная матрица вида

$$T = \begin{bmatrix} \alpha_1 & \gamma_2 & & & \\ \beta_2 & \alpha_2 & \gamma_3 & & 0 \\ & \beta_3 & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot \\ & 0 & & \cdot & \cdot & \gamma_n \\ & & & & \beta_n & \alpha_n \end{bmatrix}. \tag{15.16}$$

По-другому (15.15) можно записать как

$$W^*AV=T, \tag{15.17}$$

где $W=V^{-*}$ — матрица, составленная по столбцам из векторов $w_1, \ldots, w_n$. Таким образом, матрица V приводит A к трехдиагональной форме, но подобие не является унитарным.

Рассмотренный метод называют *двусторонним* или *несимметричным алгоритмом Ланцоша*, а в советской литературе еще и *биортогональным алгоритмом* [45, § 64].

Метод Ланцоша и его несимметричные варианты были пока что описаны как методы подобного преобразования матрицы к специальным формам — трехдиагональной или форме Хессенберга. Собственно для того они и были предложены. Однако первоначальный энтузиазм в отношении этих методов скоро сменился разочарованием. Оказалось, что в условиях приближенных вычислений их поведение совсем не похоже на предсказываемое теорией. Векторы Ланцоша рано или поздно теряют даже приблизительную ортогональность к уже вычисленным векторам; более того, в системе $q_1, \ldots, q_m$ со временем появляются почти коллинеарные векторы. Коэффициент β_n, который при точной арифметике должен быть равен нулю, не будет, как правило, даже особенно мал, равно как и предыдущие β_m. Появляется возможность продолжать вычисления по формулам (15.9) неограниченно долго, не наталкиваясь на нулевые или очень малые значения коэффициентов β_m.

Появившиеся в 50-х годах новые ортогональные методы приведения к специальным формам — сначала метод вращений (Гивенс, 1954 г.), а затем метод отражений (Хаусхолдер, 1957 г.) — отличались куда большей численной устойчивостью. То, что они являются методами трансформационного типа, не было существенно для задач умеренного порядка, решавшихся в те годы. Поэтому новые методы быстро вытеснили из численной практики алгоритмы типа Ланцоша. Что касается последних, то им стала свойственна устойчивость другого рода, а именно устойчиво плохая репутация. Это объясняет, почему в 60-е годы при возникновении потребности в решении больших спектральных задач были воскрешены методы одновременных итераций, а не ланцошевы алгоритмы.

Положение стало меняться лишь в 70-е годы, что связано с рядом обстоятельств. Прежде всего, была осознана связь алгоритма Ланцоша с методом Рэлея — Ритца; вместе с обнаружением замечательных аппроксимационных свойств крыловских подпространств это привело к переоценке взгляда на область применимости ланцошевых процессов. Скажем об этом подробнее.

Метод Рэлея — Ритца для спектральной задачи $Ax=\lambda x$ заключается в следующем. Выбирается подпространство аппроксимации $\mathscr{L}$, т. е. подпространство, в котором ищутся приближения к собственным векторам. Если обозначить приближенный собственный вектор через y, а соответствующее приближенное собственное число через θ, то определяются y и θ условием, чтобы невязка

$$r(\theta, y)=Ay-\theta y \tag{15.18}$$

была ортогональна к подпространству $\mathscr{L}$.

Пусть $q_1, \ldots, q_m$ — произвольный ортонормированный базис подпространства $\mathscr{L}$, и пусть $Q_m = [q_1 | \ldots | q_m]$. Записывая y в виде $y = Q_m s$ $(s \in \mathbf{C}^m)$, из условия $r(\theta, y) \perp \mathscr{L}$ получаем

$$(Q_m^* A Q_m) s = \theta s. \tag{15.19}$$

Таким образом, искомые приближения θ_i оказываются собственными значениями редуцированной спектральной задачи (15.19), а ее собственные векторы s_i порождают по формуле

$$y_i = Q_m s_i \tag{15.20}$$

приближения к собственным векторам матрицы A.

Числа θ_i и векторы y_i называются соответственно *числами* и *векторами Ритца для подпространства* $\mathscr{L}$. Заметим, что в методе Стьюарта из § 14 и в более раннем симметричном алгоритме Рутисхаузера шаг-поворот есть, в сущности, применение к текущему подпространству $\mathscr{L}_k$ процедуры Рэлея — Ритца.

Возьмем в качестве $\mathscr{L}$ крыловское подпространство $\mathscr{K}^m(x_0)$ эрмитовой матрицы A, и пусть $q_1, \ldots, q_m$ — его ортонормированный базис, образованный векторами Ланцоша. Из вычисленных к m-му шагу коэффициентов α_i, β_i составим трехдиагональную $m \times m$-матрицу

$$T_m = \begin{bmatrix} \alpha_1 & \beta_2 & & & 0 \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_m \\ 0 & & & \beta_m & \alpha_m \end{bmatrix}. \tag{15.21}$$

Тогда справедливо матричное равенство (см. формулы (15.9))

$$A Q_m = Q_m T_m + R_m, \tag{15.22}$$

где $R_m = [0 | \ldots | 0 | r_{m+1}]$. Умножая (15.22) слева на Q_m^*, получаем

$$Q_m^* A Q_m = T_m. \tag{15.23}$$

Итак, собственные значения матрицы T_m суть числа Ритца для крыловского подпространства $\mathscr{K}^m$ и могут рассматриваться как приближения к собственным значениям матрицы A. Но какого качества будут эти приближения?

Ответ на поставленный вопрос дан в статье Каниэля [133]. Впоследствии ее результаты были уточнены Саадом [172]; еще более точные оценки приведены в книге Князева [25] (см. по этому поводу также [35, § 12.4; 18, § 5]).

Будем для простоты считать все собственные значения матрицы A различными и упорядоченными по убыванию: $\lambda_1 > \lambda_2 > \ldots > \lambda_n$. В этом же порядке располагаем числа Ритца $\theta_1 \geqslant \theta_2 \geqslant \ldots \geqslant \theta_m$. Предположим еще, что индекс начального вектора x_0 относительно A равен n. Введем величины

$$\gamma_j = (\lambda_j - \lambda_{j+1}) / (\lambda_{j+1} - \lambda_n), \tag{15.24}$$

имеющие смысл относительной *отделенности* собственных значений от расположенной слева части спектра.

В [133] даны оценки для относительных расстояний $(\lambda_j-\theta_j)/(\lambda_j-\lambda_n)$, а также для углов, составляемых векторами Ритца y_j с одноименными собственными векторами z_j матрицы A, и углов между z_j и крыловскими подпространствами $\mathscr{K}^m$. Не приводя здесь точного вида этих оценок (более общие оценки Саада будут рассмотрены в § 16), ограничимся обсуждением их содержания.

При фиксированном j и растущем m числа Ритца $\theta_j^{(m)}$ и векторы Ритца $y_j^{(m)}$ крыловского подпространства $\mathscr{K}^m$ сходятся соответственно к λ_j и z_j. Сходимость характеризуется наличием в знаменателях оценок величины $T_{m-j}(1+2\gamma_j)$, т. е. значения в точке $1+2\gamma_j$ чебышевского многочлена 1-го рода с номером $m-j$. Правее интервала $(-1, 1)$ многочлен $T_k(\xi)$ задается выражениями

$$T_k(\xi)=\operatorname{ch}(k\operatorname{arcch}\xi)=\frac{1}{2}\left[(\xi+\sqrt{\xi^2-1})^k+(\xi+\sqrt{\xi^2-1})^{-k}\right].$$

Для больших k и для значения ξ, близкого к 1, $T_k(\xi)$ может быть приближенно заменен на $\frac{1}{2}(1+\sqrt{2(\xi-1)})^k$. Поэтому сходимость последовательности $\{y_j^{(m)}\}$ к z_j можно описать качественно как имеющую скорость геометрической прогрессии со знаменателем $1-2\sqrt{\gamma_j}$; для последовательности $\{\theta_j^{(m)}\}$ знаменатель сходимости равен $(1-2\sqrt{\gamma_j})^2$.

Для определенности здесь говорилось о сходимости к собственным значениям из правой части спектра. В действительности аналогичное утверждение справедливо и по отношению к левой части, так что сходимость наблюдается с обоих концов спектра. Скорость сходимости к каждому конкретному собственному значению зависит от его отделенности, но обычно первыми сходятся крайние собственные значения. Нередко для задач даже очень большого порядка уже после 30 шагов Ланцоша $(m=30)$ эти значения получаются с полной машинной точностью [160].

Итак, согласно оценкам Каниэля, с ростом m в крыловском подпространстве $\mathscr{K}^m$ можно найти все лучшие приближения к собственным значениям и векторам матрицы A, в первую очередь к крайним собственным значениям. Отсюда вытекает, что алгоритм Ланцоша можно использовать как метод решения частичной проблемы собственных значений, а не как метод полной трехдиагонализации матрицы. При таком подходе выполняют некоторое число m шагов Ланцоша, существенно меньшее порядка n матрицы A, и собственные значения полученной трехдиагональной матрицы T_m принимают за приближения к собственным значениям A; приближенные собственные векторы для A вычисляют из собственных векторов матрицы T_m по формуле (15.20). Качество приближений к собственным значениям можно оценить посредством норм соответствующих невязок, и метод Ланцоша дает очень удобный способ вычисления этих норм. Пусть θ, s — собственная пара для матрицы T_m, так что $T_m s=\theta s$. Умножая обе части равенства (15.22) справа на s, получаем

$$AQ_m s=Q_m T_m s+R_m s=\theta Q_m s+\{s\}_m r_{m+1}.$$

Через $\{s\}_m$ обозначена последняя компонента m-мерного вектора s. Таким образом, невязка вектора Ритца $y=Q_m s$ равна $r(\theta, y)=\{s\}_m r_{m+1}$, а ее длина составляет

$$\| r(\theta, y) \|_2 = \beta_{m-1} |\{s\}_m|. \tag{15.25}$$

Следовательно, для вычисления *длины* невязки нет необходимости строить вектор Ритца. Это важно, иначе пришлось бы в любом случае хранить векторы Ланцоша $q_1, \ldots, q_m$, даже когда нужны лишь собственные значения. Но и в том случае, если векторы Ланцоша хранятся во внешней памяти, выгодно пользоваться все же формулой (15.25) — ведь невязка может оказаться большой, вектор Ритца не будет принят в качестве приближенного собственного вектора и немалая работа, затраченная на его вычисление, окажется напрасной!

Хотя с точки зрения теории применение метода Ланцоша к частичной проблеме собственных значений выглядит обоснованным, пока не ясно, как поведет себя метод в условиях приближенных вычислений. Объяснение этого поведения, предложенное в начале 70-х годов Пейджем [35, § 13.4; 18, § 6], стало вторым фактором, оживившим интерес к методу. Вкратце суть открытия, сделанного Пейджем, состоит в том, что момент, когда в системе *реально вычисляемых* векторов Ланцоша впервые происходит сильная потеря ортогональности, всегда совпадает с моментом появления в спектре матрицы T_m хорошего приближения к некоторому собственному значению матрицы A. До этого момента поведение реального процесса мало расходится с теорией.

Третьим обстоятельством, способствовавшим возрождению метода, было обнаружение так называемого «феномена Ланцоша» [87; 18, § 9, 10]. Если, несмотря на утерю ортогональности векторами Ланцоша, продолжать процесс (15.9), то при достаточно большом m в спектре матрицы T_m можно найти приближения ко *всем* собственным значениям матрицы A. Число шагов, необходимое для этого, может значительно превысить порядок n матрицы; практическая же возможность «шагать» так далеко обеспечивается тем, что β_n и последующие β в отличие от теории не только не равны нулю, но и, как правило, не очень малы.

Начиная с середины 70-х годов появляется ряд программных реализаций метода Ланцоша, основанных на различных принципах. В одних считается важным сохранение глобальной ортогональности векторов Ланцоша, и вычисляемые векторы подвергаются переортогонализации. Результаты Пейджа позволяют проводить переортогонализацию лишь время от времени — в «моменты сходимости» собственных значений, и это обстоятельство используют так называемые *методы выборочной ортогонализации*. Первый из методов этой группы описан в гл. 13 книги [35]. В программах другого направления не только не проводится переортогонализация, но и сами векторы Ланцоша не хранятся. Такие программы эксплуатируют феномен Ланцоша, и основная их проблема — выделить среди собственных значений матрицы T_m те, которые приближают собственные значения матрицы A. О программах этого типа дает представление двухтомник [87]. Существуют и иные программы, со своим обоснованием [39, 40].

В целом, если говорить об эрмитовых матрицах, алгебраисты отдают сейчас алгоритму Ланцоша явное предпочтение перед методами одновременных итераций. Хотя и в том и в другом случае приближения к собственным значениям и векторам черпаются из крыловского пространства, алгоритм Ланцоша использует заключенную в нем спектральную информацию полностью, тогда как одновременная итерация — лишь то, что можно найти в его r-мерном подпространстве, пусть даже улучшенном предыдущими итерациями. Нет проблемы с «угадыванием» значения для r, и можно находить любое нужное число собственных векторов. Сравнение обоих подходов проведено в [147]. Вывод авторов: современная, подобающим образом реализованная версия алгоритма Ланцоша для $n > 500$ должна работать на порядок быстрее, чем хорошая программа одновременных итераций. Чем больше требуется собственных векторов, тем выгодней сравнение для метода Ланцоша.

Преимущества алгоритма Ланцоша начинают теперь оценивать и на рынке спектральных программ. До недавнего времени программные реализации метода носили по преимуществу исследовательский характер и не имели необходимого коммерческого блеска. В последние годы ситуация меняется. Шведская программа STLM (авторы — Эриксон и Руэ [105]), вычисляющая все, в том числе внутренние точки спектра, приобретена самолетостроительной фирмой «Боинг». Льюис и Граймс [140] сообщают о программе метода Ланцоша, разработанной с целью ввода в прикладной пакет GTSTRUDL. Этот популярный пакет для расчетов по методу конечных элементов имел ограниченные возможности для решения больших спектральных задач: их порядок в случае обобщенной проблемы не превышал 400. Программа BCS Льюиса — Граймса не переортогонализует векторы Ланцоша и идейно близка к программам из книги [87]. Опыт эксплуатации BCS показал, что она работает на порядок быстрее обратной итерации из пакета NASTRAN или программы итерирования подпространства из пакета SAP IV (это также очень известные в области конечно-элементных вычислений пакеты), а ее преимущество над программой одновременных итераций из GTSTRUDL исчисляется несколькими порядками. В пакете программа BCS позволяет поднять размер решаемых спектральных задач до 1200—1500. Для задач порядка 10^3 и выше число шагов Ланцоша, необходимое для получения нужных собственных пар, в типичном случае колеблется от 100 до 150. Составленная Скоттом программа LAS 02 (в ней реализован *блочный* вариант выборочной ортогонализации; см. [35, § 13.10; 18, § 10]) интенсивно используется в Ок-Риджской национальной лаборатории США для того же круга задач, что и BCS. Программа решает задачи порядка более 2×10^3 и работает в 10 раз быстрее программы одновременных итераций из пакета SAP. Предварительная версия программы LAS 02 сравнивалась в рамках пакета SESAM 80 с программой одновременных итераций SSIT 25, очень схожей с аналогичной программой из книги [2]. Превосходство программы Скотта проявилось и здесь [50]. Программа LAS 02 доступна любому пользователю; ее распространитель — Аргоннский программный центр NESC в США. Еще одна версия метода Ланцоша включена в конечно-элементный пакет SAMCEF, эксплуатирующийся

в университете бельгийского города Льежа [76]. Наконец, упомянем отечественные программы метода Ланцоша. Помимо программы Собянина [40] это — программы, описываемые в [34, 36].

Таким образом, можно констатировать большие продвижения и в теории, и в программной реализации метода Ланцоша для симметричных матриц. В последние годы внимание алгебраистов переключилось на несимметричные задачи. Здесь прогресс пока не так велик и сама задача существенно сложнее. Однако основные направления исследований выявились — это различные варианты алгоритма Арнольди и биортогонального алгоритма. Они будут рассмотрены в нескольких следующих параграфах.

§ 16. Метод Арнольди

Суть метода Арнольди изложена в предыдущем параграфе (см. формулы (15.10) — (15.13). Напомним, что первоначально метод был предложен как способ унитарного преобразования матрицы к форме Хессенберга, однако в этом качестве разделил судьбу метода Ланцоша: появление более устойчивых методов — метода вращений и метода отражений — привело к тому, что метод был надолго забыт.

Метод снова привлек к себе внимание после публикации Саада [171], где сообщалось об экспериментальном его сравнении с методами одновременных итераций. В отношении времени работы преимущество метода Арнольди оказалось примерно таким же, как метода Ланцоша перед методом Рутисхаузера.

По сравнению с алгоритмом Ланцоша у метода Арнольди есть большой минус: матрица коэффициентов рекурсии хессенбергова, а не трехдиагональная. Необходимость хранить ее исключает проведение большого числа шагов и вынуждает пользоваться методом лишь для отыскания группы собственных значений и векторов. Однако есть ли и теперь — в неэрмитовом случае — основания надеяться на получение приближений хорошего качества в крыловском подпространстве относительно небольшой размерности m? Этот вопрос рассматривается Саадом [171, 174].

Обозначим через $\mathscr{P}_m$ ортогональный проектор на крыловское подпространство $\mathscr{K}^m$. Если Q_m — матрица с ортонормированными столбцами, составленная из *векторов Арнольди* $q_1, \ldots, q_m$, то $\mathscr{P}_m = Q_m Q_m^*$. Введем линейный оператор $\mathscr{A}_m = \mathscr{P}_m A \mathscr{P}_m$. Он равен нулю на ортогональном дополнении к подпространству $\mathscr{K}^m$, а на самом $\mathscr{K}^m$ его действие описывается хессенберговой матрицей

$$H_m = Q_m^* A Q_m = \begin{bmatrix} h_{11} & h_{12} & \ldots & h_{1,m-1} & h_{1m} \\ h_{21} & h_{22} & \ldots & h_{2,m-1} & h_{2m} \\ & h_{32} & \ldots & h_{3,m-1} & h_{3m} \\ & 0 & \ddots & \ddots & \ddots \\ & & & h_{m,m-1} & h_{mm} \end{bmatrix}. \tag{16.1}$$

Пусть λ и u — собственная пара матрицы A. Показать, что некоторая пара (θ, y), где θ — число, а y — вектор Ритца, аппроксимирует пару (λ, u), можно двумя способами: во-первых, проверить, что пара (θ, y) дает малую невязку относительно матрицы A; во-вторых, проверить, что пара (λ, u) дает малую невязку относительно $\mathscr{A}_m$. Мы воспользуемся вторым способом и оценим величину невязки через расстояние от вектора u до крыловского подпространства.

Теорема 16.1. *Пусть*

$$\gamma_m = \| \mathscr{P}_m A (I - \mathscr{P}_m) \|_2. \tag{16.2}$$

Тогда

$$\| \mathscr{A}_m u - \lambda u \|_2 \leqslant \sqrt{|\lambda|^2 + \gamma_m^2}\, \| (I - \mathscr{P}_m) u \|_2. \tag{16.3}$$

Доказательство. Запишем невязку в виде

$$(\mathscr{A}_m - \lambda I) u = \mathscr{P}_m (A - \lambda I) \mathscr{P}_m u - \lambda (I - \mathscr{P}_m) u =$$

$$= \mathscr{P}_m (A - \lambda I)(\mathscr{P}_m u - u) - \lambda (I - \mathscr{P}_m) u = -\mathscr{P}_m (A - \lambda I)(I - \mathscr{P}_m) u - \lambda (I - \mathscr{P}_m) u.$$

Так как $I - \mathscr{P}_m$ также проектор (на $(\mathscr{K}^m)^\perp$), то $(I - \mathscr{P}_m)^2 = I - \mathscr{P}_m$ и

$$(\mathscr{A}_m - \lambda I) u = -\mathscr{P}_m (A - \lambda I)(I - \mathscr{P}_m)(I - \mathscr{P}_m) u - \lambda (I - \mathscr{P}_m) u.$$

Слагаемые правой части ортогональны. Поэтому

$$\| (\mathscr{A}_m - \lambda I) u \|_2^2 = \| \mathscr{P}_m (A - \lambda I)(I - \mathscr{P}_m)(I - \mathscr{P}_m) u \|_2^2 + |\lambda|^2 \| (I - \mathscr{P}_m) u \|_2^2. \tag{16.4}$$

Норму первого члена можно оценить через $\| \mathscr{P}_m (A - \lambda I)(I - \mathscr{P}_m) \|_2 \times \| (I - \mathscr{P}_m) u \|_2$. Заметим, что

$$\mathscr{P}_m (A - \lambda I)(I - \mathscr{P}_m) = \mathscr{P}_m A (I - \mathscr{P}_m) - \lambda \mathscr{P}_m (I - \mathscr{P}_m) = \mathscr{P}_m A (I - \mathscr{P}_m).$$

Учитывая (16.2), выводим из (16.4) неравенство (16.3).

Замечание 16.2. Представляет интерес интерпретация величины γ_m. Предположим, что процесс Арнольди доведен до конца и получена хессенбергова матрица (15.13). Разобьем ее на блоки

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix},$$

где H_{11} совпадает с матрицей H_m (см. (16.1)). В базисе пространства $\mathbf{C}^n$, составленном из векторов Арнольди, оператор $\mathscr{P}_m A (I - \mathscr{P}_m)$ имеет матрицу

$$\begin{bmatrix} 0 & H_{12} \\ 0 & 0 \end{bmatrix}.$$

Итак, $\gamma_m = \| H_{12} \|_2$. В частности, для нормальной матрицы A справедливо равенство $\| H_{12} \|_E = \| H_{21} \|_E$, и поскольку в H_{21} отличен от нуля (и равен $h_{m+1,m}$) только элемент в правом верхнем углу, то окончательно $\gamma_m \leqslant |h_{m+1,m}|$.

Перейдем теперь к оценке расстояния между u и крыловским подпространством. Заметим, что каждый вектор x из $\mathscr{K}^m$, будучи

линейной комбинацией векторов $q_1, Aq_1, ..., A^{m-1}q_1$, может быть представлен в виде

$$x=(\alpha_0 I+\alpha_1 A+ ... +\alpha_{m-1}A^{m-1})q_1=\varphi(A)q_1,$$

где $\varphi(t)$ — это многочлен $\alpha_0+\alpha_1 t+ ... +\alpha_{m-1}t^{m-1}$. Таким образом, $\mathscr{K}^m$ можно описать как миожество векторов вида $\varphi(A)q_1$ при $\varphi(t)$, пробегающем пространство π_{m-1} миогочленов степени не больше $m-1$. Вместо q_1 можно взять любой коллинеарный ему вектор.

Для дальнейшего нам придется предположить, что все n собственных значений $\lambda_1, ..., \lambda_n$ матрицы A различны. Пусть $u_1, ..., u_n$ — базис из ее собственных векторов, причем все векторы имеют единичную евклидову длину. Предположим еще, что в разложении начального вектора q_1 по базису $\{u_i\}$ все коэффициенты отличны от нуля:

$$q_1=\sum_{i=1}^{n}\alpha_i u_i, \qquad \alpha_i\neq 0, \qquad i=1, ..., n.$$

Теорема 16.3. *Справедливы неравенства*

$$\|(I-\mathscr{P}_m)u_i\|_2\leqslant\xi_i\varepsilon_i^{(m)}, \qquad i=1, ..., n, \tag{16.5}$$

где

$$\xi_i=\sum_{\substack{j=1\\ j\neq i}}^{n}(|\alpha_j|/|\alpha_i|), \qquad \varepsilon_i^{(m)}=\min_{\substack{\varphi\in\pi_{m-1}\\ \varphi(\lambda_i)=1}}\max_{j\neq i}|\varphi(\lambda_j)|. \tag{16.6}$$

Доказательство. Положим

$$\tilde{q}_1=q_1/\alpha_i=u_i+\sum_{\substack{j=1\\ j\neq i}}^{n}(\alpha_j u_j/\alpha_i). \tag{16.7}$$

Так как $\mathscr{P}_m u_i\in\mathscr{K}^m$, то $\mathscr{P}_m u_i=\psi(A)\tilde{q}_1$ для некоторого многочлена $\psi\in\pi_{m-1}$. Поэтому

$$\|(I-\mathscr{P}_m)u_i\|_2=\min_{\varphi\in\pi_{m-1}}\|u_i-\varphi(A)\tilde{q}_1\|_2\leqslant\min_{\substack{\varphi\in\pi_{m-1}\\ \varphi(\lambda_i)=1}}\|u_i-\varphi(A)\tilde{q}_1\|_2.$$

Если $\varphi(\lambda_i)=1$, то (см. (16.7)) $u_i-\varphi(A)\tilde{q}_1=-\sum_{\substack{j=1\\ j\neq i}}^{n}(\alpha_j\varphi(\lambda_j)u_j/\alpha_i)$. Следовательно,

$$\|(I-\mathscr{P}_m)u_i\|_2\leqslant\xi_i\min_{\substack{\varphi\in\pi_{m-1}\\ \varphi(\lambda_i)=1}}\max_{j\neq i}|\varphi(\lambda_j)|=\xi_i\varepsilon_i^{(m)}.$$

При произвольном расположении спектра на комплексной плоскости получение оценки для величины $\varepsilon_i^{(m)}$ является трудной задачей (см. п. 2 дополнений к § 16). Поэтому потребуем, чтобы все собственные значения матрицы A были вещественными. Расположим их по убыванию:

$$\lambda_1>\lambda_2>...>\lambda_n$$

и введем, как в (15.24), величины $\gamma_j=(\lambda_j-\lambda_{j+1})/(\lambda_{j+1}-\lambda_n)$. Положим $\varkappa_1=1$,

$$\varkappa_i=\prod_{j=1}^{i-1}\frac{\lambda_j-\lambda_n}{\lambda_j-\lambda_i}, \qquad i>1. \tag{16.8}$$

Теорема 16.4. *Имеет место оценка*

$$\varepsilon_i^{(m)}\leqslant\frac{\varkappa_i}{T_{m-i}(1+2\gamma_i)}, \tag{16.9}$$

где T_k — многочлен Чебышева первого рода степени k.

Доказательство. Обозначим через ω_i множество всех многочленов вида $\varphi(t)=l_i(t)\tau(t)$, где $l_1(t)\equiv 1$,

$$l_i(t)=\frac{(t-\lambda_1)(t-\lambda_2)\ldots(t-\lambda_{i-1})}{(\lambda_i-\lambda_1)(\lambda_i-\lambda_2)\ldots(\lambda_i-\lambda_{i-1})}, \qquad i>1,$$

а $\tau(t)$ — многочлен степени не выше $m-i$, такой, что $\tau(\lambda_i)=1$. Понятно, что $\varphi\in\pi_{m-1}$ и $\varphi(\lambda_i)=1$, поэтому

$$\varepsilon_i^{(m)}\leqslant\min_{\varphi\in\omega_i}\max_{j\neq i}|\varphi(\lambda_j)|. \tag{16.10}$$

Так как $\varphi(\lambda_j)=0$ $(j=1, \ldots, i-1)$, то максимум в формуле (16.10) достигается для некоторого λ_j с индексом больше i. Следовательно,

$$\min_{\substack{\tau\in\pi_{m-i}\\ \tau(\lambda_i)=1}}\max_{i<j\leqslant n}\left|\frac{(\lambda_j-\lambda_1)\ldots(\lambda_j-\lambda_{i-1})}{(\lambda_i-\lambda_1)\ldots(\lambda_i-\lambda_{i-1})}\right||\tau(\lambda_j)|\leqslant\varkappa_i\min_{\substack{\tau\in\pi_{m-i}\\ \tau(\lambda_i)=1}}\max_{i<j\leqslant n}|\tau(\lambda_j)|\leqslant$$

$$\leqslant\varkappa_i\min_{\substack{\tau\in\pi_{m-i}\\ \tau(\lambda_i)=1}}\max_{\lambda_n\leqslant t\leqslant\lambda_{i+1}}|\tau(t)|.$$

Хорошо известно (см., например, [6, гл. 2, § 3; гл. 4, § 3]), что минимаксный член в правой части равен $\left[T_{m-i}(1+2\gamma_i)\right]^{-1}$. Это доказывает (16.9).

Замечание 16.5. Результат, аналогичный (16.9), можно получить для чисто мнимого спектра и, более общо, для случая, когда спектр сосредоточен на прямой в комплексной плоскости.

Итак, в соответствии с теоремами 16.1, 16.3, 16.4 в крыловском подпространстве достаточно большой размерности m существуют — по крайней мере для матриц с линейным спектром — хорошие приближения к собственным парам матрицы A. В качестве приближений берут собственные значения хессенберговой матрицы (16.1); их можно определить с помощью QR-алгоритма. Приближенными собственными векторами являются векторы Ритца, формируемые из собственных векторов матрицы H_m по формуле (15.20). При этом, как и в методе Ланцоша, длину невязки, отвечающей числу Ритца θ и вектору Ритца y, можно найти, *не вычисляя y в явном виде.* Действительно, формулы первых m шагов процесса Арнольди в совокупности равносильны матричному соотношению $AQ_m=Q_mH_m+R_m$, $R_m\equiv[0|\ldots|0|r_{m+1}]$, где вектором r_{m+1} служит разность

$Aq_m - h_{1m}q_1 - \ldots - h_{mm}q_m$. Учитывая (15.11), по аналогии с (15.25) получаем

$$\| r(\theta, y) \|_2 = h_{m+1,m} |\{s\}_m|, \tag{16.11}$$

где s — собственный вектор матрицы H_m, порождающий вектор y. Тем самым, вычислять или нет вектор Ритца, определяется в зависимости от величины произведения в правой части уравнения (16.11).

Необходимость хранить векторы Арнольди — *все* они нужны для рекуррентного счета по формуле (15.10) — и хессенбергову матрицу H_m приводит к тому, что ресурсы оперативной памяти обычно оказываются исчерпанными еще до того, как в полной мере сказывается сходимость, обещаемая теорией. Поэтому пользуются различными модификациями основного алгоритма Арнольди.

Одна из возможностей состоит в том, чтобы рассматривать алгоритм как итерационный. После заданного числа m шагов, определяемого, например, размером выделенной памяти, происходит *остановка* для вычисления собственных значений матрицы H_m и проверки условий сходимости (например, длин невязок по формуле (16.11)). Для чисел Ритца, удовлетворивших принятое условие, вычисляются соответствующие векторы Ритца. Если получены приближения ко всем желаемым собственным парам, процесс закончен. В противном случае выполняется новый *прогон* алгоритма. Информация, полученная в первом прогоне, может быть использована при выборе начального вектора $\tilde{q}_1$ для второго. Естественный способ — взять в качестве $\tilde{q}_1$ линейную комбинацию векторов Ритца, еще не сошедшихся с требуемой точностью. Нет нужды строить каждый из этих векторов. Если $y_i = Q_m s_i$ $(i = 1, \ldots, l)$ и $\tilde{q}_1 = \alpha_1 y_1 + \ldots + \alpha_l y_l$, то $\tilde{q}_1 = Q_m(\alpha_1 s_1 + \ldots + \alpha_l s_l)$, т. е. вычисляется только один n-мерный вектор. Если и второй прогон не даст всех нужных собственных пар, совершается третий и т. д.

Другая возможность, рассматриваемая в [171], — это так называемый *метод неполной ортогонализации*. В его основе наблюдение, сделанное во многих экспериментах: при фиксированном i элементы h_{ij} убывают с ростом j. Это обстоятельство можно использовать так: задавшись натуральным числом p, на каждом шаге ортогонализовать вектор Aq_m не ко всем предыдущим векторам, а только к q_{m-p}, q_{m-p+1}, $\ldots$, q_m. Матрица $\tilde{H}_m$, отвечающая такому процессу, имеет ленточную форму Хессенберга: элементы h_{ij} равны нулю при $j < i-1$ и $j > i+p$. Тем самым ее хранение требует меньшей памяти, чем в ортодоксальном алгоритме Арнольди, да и для проведения рекурсии нужны лишь последние $p+1$ векторов q_j. Однако система этих векторов теперь утрачивает глобальную ортогональность даже в точной арифметике; ортогональны лишь ее подсистемы из $p+1$ подряд идущих векторов. С потерей глобальной ортогональности связано и то, что матрица $\tilde{H}_m$ уже не соответствует проецированию на крыловское подпространство по методу Рэлея — Ритца. Пока не ясно, почему эта матрица вообще должна заключать в себе какую-то информацию об исходной спектральной задаче. Частичное объяснение дает обобщение теоремы 16.1, полученное Саадом [174].

Будем искать в крыловском подпространстве $\mathscr{K}^m$ приближение к собственному вектору матрицы A, но не ортогональным проектированием, как в методе Рэлея—Ритца, а в соответствии с более общим методом Петрова—Галёркина. Это значит, что наряду с $\mathscr{K}^m$ рассматривается еще одно m-мерное подпространство $\mathscr{L}^m$, а в качестве приближенной собственной пары берут вектор $y \in \mathscr{K}^m$ и число θ, такие, что

$$r(\theta, y) = Ay - \theta y \perp \mathscr{L}^m \tag{16.12}$$

(*условие Петрова—Галёркина*). Если $Q_m = [q_1 | \ldots | q_m]$ — базисная матрица крыловского подпространства, $W_m = [w_1 | \ldots | w_m]$ — базисная матрица подпространства $\mathscr{L}^m$, то, полагая $y = Q_m s$ $(s \in \mathbf{C}^m)$, получаем из (16.12) соотношение

$$(W_m^* A Q_m) s = \theta (W_m^* Q_m) s, \tag{16.13}$$

т. е. обобщенную проблему собственных значений порядка m (см. § 14). В частном случае, когда системы $\{q_i\}$ и $\{w_j\}$ двойственны, т. е. $W_m^* Q_m = I_m$, (16.13) превращается в обычную задачу для матрицы $B_m = W_m^* A Q_m$.

Наряду с ортогональным проектором $\mathscr{P}_m$ рассмотрим еще — в предположении, что подпространства $\mathscr{K}^m$ и $(\mathscr{L}^m)^\perp$ дополнительные (см. § 1), — косой проектор $\mathscr{P}_{\mathscr{K}, \mathscr{L}^\perp} \equiv \hat{\mathscr{P}}_m$.

Под оператором $\mathscr{A}_m$ будем теперь понимать произведение $\hat{\mathscr{P}}_m A \mathscr{P}_m$. Ядром этого оператора является $\mathscr{K}_m^\perp$, образом — само крыловское подпространство. Для сужения оператора $\mathscr{A}_m$ на подпространство $\mathscr{K}^m$ матрицей служит произведение $W_m^* A Q_m$. Пусть, как и в теореме 16.1, λ и u — собственная пара матрицы A. Положим

$$\gamma_m = \| \hat{\mathscr{P}}_m (A - \lambda I)(I - \mathscr{P}_m) \|_2. \tag{16.14}$$

Теорема 16.6. *Справедливы оценки*

$$\| (\mathscr{A}_m - \lambda I) \mathscr{P}_m u \|_2 \leqslant \gamma_m \| (I - \mathscr{P}_m) u \|_2, \tag{16.15}$$

$$\| (\mathscr{A}_m - \lambda I) u \|_2 \leqslant \sqrt{|\lambda|^2 + \gamma_m^2}\, \| (I - \mathscr{P}_m) u \|_2. \tag{16.16}$$

Доказательство. Имеют место равенства

$$(\mathscr{A}_m - \lambda I) \mathscr{P}_m u = \hat{\mathscr{P}}_m (A - \lambda I) \mathscr{P}_m u = \hat{\mathscr{P}}_m (A - \lambda I)(\mathscr{P}_m u - u) =$$
$$= -\hat{\mathscr{P}}_m (A - \lambda I)(I - \mathscr{P}_m)(I - \mathscr{P}_m) u. \tag{16.17}$$

В последнем переходе учтено, что $I - \mathscr{P}_m$ также проектор, и, следовательно, $(I - \mathscr{P}_m)^2 = I - \mathscr{P}_m$. Беря нормы левой и правой частей в (16.17) и используя (16.14), получаем (16.15).

Чтобы доказать (16.16), представим невязку $(\mathscr{A}_m - \lambda I) u$ в виде

$$(\mathscr{A}_m - \lambda I) u = (\mathscr{A}_m - \lambda I)[\mathscr{P}_m u + (I - \mathscr{P}_m) u] = (\mathscr{A}_m - \lambda I) \mathscr{P}_m u +$$
$$+ (\mathscr{A}_m - \lambda I)(I - \mathscr{P}_m) u = (\mathscr{A}_m - \lambda I) \mathscr{P}_m u - \lambda (I - \mathscr{P}_m) u.$$

В последнем переходе использовано равенство $\mathscr{A}_m (I - \mathscr{P}_m) = 0$. Первое слагаемое правой части принадлежит подпространству $\mathscr{K}^m$, а второе — его ортогональному дополнению. Поэтому

$$\| (\mathscr{A}_m - \lambda I) u \|_2^2 = \| (\mathscr{A}_m - \lambda I) \mathscr{P}_m u \|_2^2 + \| \lambda (I - \mathscr{P}_m) u \|_2^2.$$

Применяя (16.15), получаем нужный результат.

Замечание 16.7. Поскольку $\mathscr{P}_m(I-\mathscr{P}_m)=0$, то в частном случае $\hat{\mathscr{P}}_m=\mathscr{P}_m$ определение (16.14) сводится к (16.2). Однако если γ_m из формулы (16.2) можно было ограничить сверху величиной $\|A\|_2$, то в отношении (16.14) это не верно. Норма косого проектора $\hat{\mathscr{P}}_m$ больше 1 (подчас значительно больше); при этом значение $\|\hat{\mathscr{P}}_m\|$ или хотя бы оценка для него известны не всегда. Если привлечь базисные матрицы подпространств $\mathscr{K}^m$ и $\mathscr{L}^m$, считая, что их столбцы образуют биортогональные системы, то проектор $\hat{\mathscr{P}}_m$ можно представить выражением $\hat{\mathscr{P}}_m=Q_mW_m^*$. Но в то время как векторы $q_1, \ldots, q_m$ известны, векторы $p_1, \ldots, p_m$ могут в реальном процессе и не вычисляться. Именно так обстоит дело в методе неполной ортогонализации. Здесь $q_1, \ldots, q_m$ — вычисляемые векторы, составляющие неортогональную в целом систему, а $p_1, \ldots, p_m$ — двойственная система, которая в принципе может быть построена, но в действительности не строится.

Несмотря на сделанное замечание, теорема 16.6 вместе с теоремами 16.3 и 16.4 устанавливает принципиальную возможность пользоваться методами типа неполной ортогонализации. Если нормы проекторов $\hat{\mathscr{P}}_m$ ограничены в совокупности, то, согласно (16.16), качество приближения, вычисляемого методом, определяется тем, насколько хорошо искомый собственный вектор u аппроксимируется крыловским подпространством.

Отметим в заключение, что при прочих равных условиях эффективность различных вариантов метода Арнольди зависит от удачного выбора начального вектора q_1. Вопросу о выборе начального вектора посвящен § 19.

ДОПОЛНЕНИЯ К § 16

1. Пусть $\mathscr{K}^m(q_1)$ — крыловское подпространство с базисной матрицей Q_m, составленной из ортонормированных векторов, например векторов Арнольди. Положим $B_m=Q_m^*AQ_m$, и пусть $\mu(t)=|tI-B_m|$ — характеристический многочлен матрицы B_m.

Теорема (Саад [174]). *Многочлен $\mu(t)$ реализует минимум нормы $\|\varphi(A)q_1\|_2$ на множестве всех многочленов φ степени m со старшим коэффициентом* 1.

Доказательство. По теореме Гамильтона — Кэли $\mu(\mathscr{A}_m)=0$ (напомним, что на $(\mathscr{K}^m)^\perp$ оператор $\mathscr{A}_m=\mathscr{P}_mA\mathscr{P}_m$ совпадает с нулевым оператором). Поэтому

$$(\mu(\mathscr{A}_m)q_1, x)=0 \qquad \forall x\in\mathscr{K}^m. \tag{16.18}$$

Заметим, что при любом $k\,(1\leqslant k\leqslant m)$ справедливо равенство

$$(\mathscr{A}_m)^kq_1=\mathscr{P}_mA^kq_1. \tag{16.19}$$

Действительно, при $k=1$ имеем $\mathscr{A}_mq_1=\mathscr{P}_mA\mathscr{P}_mq_1=\mathscr{P}_mAq_1$. Пусть $k<m$ и выполнено (16.19), тогда

$$(\mathscr{A}_m)^{k+1}q_1=\mathscr{A}_m(\mathscr{A}_m)^kq_1=\mathscr{A}_m\mathscr{P}_mA^kq_1=\mathscr{P}_mA\mathscr{P}_mA^kq_1=\mathscr{P}_mA^{k+1}q_1.$$

Здесь мы пользовались тем обстоятельством, что $A^kq_1\in\mathscr{K}^m$ при $k\leqslant m-1$. По этой же причине можно опустить $\mathscr{P}_m$ в правой части (16.19), если $k\leqslant m-1$. Итак, равенство (16.19) установлено. Возвращаясь к (16.18), имеем

$$0=(\mu(\mathscr{A}_m)q_1, x)=(\mathscr{P}_m\mu(A)q_1, x)=(\mu(A)q_1, \mathscr{P}_mx)=(\mu(A)q_1, x) \tag{16.20}$$

для любого $x \in \mathscr{K}^m$. Положим $\mu(t) = t^m - \nu(t)$, где $\nu(t) \in \pi_{m-1}$. Равенство (16.20) означает, что вектор $A^m q_1 - \nu(A) q_1$ ортогонален к подпространству $\mathscr{K}^m$, т. е. длина его наименьшая возможная среди длин векторов $A^m q_1 - \varphi(A) q_1$ $(\varphi(t) \in \pi_{m-1})$.

В [35, § 12.3] теорема Саада доказана для случая эрмитовой матрицы A. Приведенный вывод показывает, что в действительности утверждение верно для произвольной матрицы.

2. По поводу оценки величины $\varepsilon_1^{(m)}$ приведем (без доказательства) следующий результат из [174]. Не ограничивая общности, положим $i = 1$.

Теорема (Саад [174]). *Пусть $m < n$. Тогда найдутся m собственных значений матрицы A (обозначим их через $\lambda_2, \ldots, \lambda_{m+1}$) такие, что*

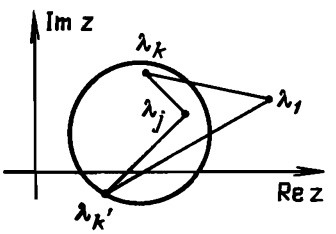

Рис. 6

$$\left[\varepsilon_1^{(m)}\right]^{-1} = \sum_{j=2}^{m+1} \prod_{\substack{k=2 \\ k \neq j}}^{m+1} \frac{|\lambda_k - \lambda_1|}{|\lambda_k - \lambda_j|}. \tag{16.21}$$

В некоторых случаях эта формула позволяет установить быстрое убывание величин $\varepsilon_1^{(m)}$. Пусть, например, удается заключить все собственные значения, кроме λ_1, в круг (см. рис. 6), и пусть λ_j — собственное значение, ближайшее к λ_1. Поскольку для прочих чисел λ_k, как правило, верно неравенство $|\lambda_1 - \lambda_k| > |\lambda_j - \lambda_k|$, то уже слагаемое $\prod_{k \neq 1, j} (|\lambda_k - \lambda_1| / |\lambda_k - \lambda_j|)$ в сумме (16.21) будет велико, обеспечивая малое значение для $\varepsilon_1^{(m)}$. Для этого же (и даже несколько более общего) случая можно дать точную оценку [174]. Пусть все, кроме λ_1, собственные значения матрицы A находятся внутри эллипса с центром d и фокусами в точках $d+e$ и $d-e$. Пусть $d+a$ указывает конец главной полуоси. Тогда a/e есть вещественное число (можно считать его положительным) и

$$\varepsilon_1^{(m)} \leqslant \frac{T_{m-1}(a/e)}{|T_{m-1}[(\lambda_1 - d)/e]|}.$$

Для комплексной переменной z многочлен Чебышева $T_m(z)$ задается формулой $T_m(z) = \operatorname{ch}(k \operatorname{arcch} z)$.

Однако величина $\varepsilon_1^{(m)}$ из (16.21) оказывается малой не при всяком распределении спектра на комплексной плоскости. Так, если $\lambda_k = \omega_0^{k-1}$ $(k = 1, \ldots, n)$, где ω_0 — первообразный корень n-й степени из 1, то даже при $m = n-1$ будем иметь $\varepsilon_1^{(m)} = 1/m$ [174].

3. В неэрмитовом случае длина невязки, отвечающей приближенной собственной паре (θ, y), не дает полноценного представления о величине погрешности даже в θ. Если приближаемое собственное значение λ простое, то более надежную оценку ошибки в θ можно получить посредством отношения $\|y\| \|z\| / |(y, z)|$, где z — приближенный *левый* собственный вектор для того же собственного значения (см. § 7). Исходя из этого, в [166] рассматривается модификация метода Арнольди, названная автором *двусторонним алгоритмом Арнольди*. Пусть после m шагов обычного процесса (15.10), (15.11) впервые для некоторой пары (θ, y) выполнен критерий сходимости, например длина невязки достигла нужного уровня малости. Если t — левый собственный вектор матрицы H_m для собственного значения θ, то нет никаких оснований считать вектор $w = Q_m t$ хорошим приближением к левому собственному вектору матрицы A (в отличие от вектора $y = Q_m s$, который *принят*

за приближение к правому собственному вектору!). Однако w можно взять начальным вектором для нового прогона, причем вести его уже не для матрицы A, а для сопряженной матрицы A^*. Результатом должно быть уже хорошее приближение z к искомому левому вектору.

Там же, в [166], отмечено, что положение с потерей ортогональности в реальио вычисляемой последовательности $q_1, \ldots, q_m$ существенно отличается от того, что наблюдается в методе Ланцоша. Матрицу Q_{m+1} можно рассматривать как результат применения к $n\times(m+1)$-матрице

$$\hat{A}=[q_1 \mid Aq_1 \mid \ldots \mid Aq_m]$$

так называемого *модифицированного* процесса ортогонализации Грама — Шмидта (см. [30, § 19]). Действительно, из формул (15.10) вытекает матричное равенство

$$\hat{A}=Q_{m+1}\begin{bmatrix} 1 & h_{11} & \ldots & h_{1m} \\ & h_{21} & \ldots & h_{2m} \\ & & & \\ & 0 & \ldots & \\ & & & h_{m+1,m} \end{bmatrix} = Q_{m+1}R_{m+1}$$

с верхней треугольной матрицей R_{m+1}. Анализ численного поведения модифицированного процесса Грама — Шмидта, выполненный Бьорком [62], приводит к оценке

$$\| I-\tilde{Q}_{m+1}^{*}\tilde{Q}_{m+1} \|_2 \leqslant f(n)\| A \|_E \| \tilde{R}_{m+1}^{-1} \|_2,$$

где $\tilde{Q}$ и $\tilde{R}$ — реально вычисленные матрицы, а $f(n)$, как и во всех прежних случаях, можно считать степенной функцией с небольшим показателем. Тем самым величина отклонения вычисляемой системы $\tilde{q}_1, \ldots, \tilde{q}_m$ от ортонормальной зависит от обусловленности (в смысле обращения) матрицы $\tilde{R}_m$.

Для норм $\| R_m^{-1} \|_2$ справедлива рекуррентная оценка

$$\| R_m^{-1} \|_2 \leqslant \rho_m, \tag{16.22}$$

где $\rho_1=1$ и при $m=1, 2, \ldots$

$$\rho_{m+1}=\rho_m\left(1+|h_{m+1,m}|\alpha_m\right)^{-1}+|h_{m+1,m}|^{-1}. \tag{16.23}$$

Через α_m обозначена евклидова длина вектора $(h_{1m}h_{2m}\ldots h_{mm})^{\mathrm{T}}$.

Из формул (16.22), (16.23) можно сделать такой вывод: в системе $\tilde{q}_1, \tilde{q}_2, \ldots$ не будет больших отклонений от ортогональности, пока коэффициенты $h_{m+1,m}$ не малы. Малость же коэффициента $h_{m+1,m}$ означает, что получено приближенное инвариантное подпространство матрицы A, и тогда H_m позволяет найти сразу m приближенных собственных пар!

4. Еще Ланцош [136] указал, что его метод может быть использован для решения системы линейных уравнений с симметричной матрицей A. В качестве начального вектора q_1 целесообразно брать иормированную правую часть b системы, т. е. $q_1=b/\beta_0$, $\beta_0=\| b \|_2$. При этом (что было замечено значительно позднее) не обязательно доводить дело до полной трехдиагонализации матрицы A. После m шагов процесса Ланцоша, где $m \ll n$, можно в качестве приближеиного решения системы $Ax=b$ взять вектор y крыловского подпространства $\mathscr{K}^m$, определяемый условием Галёркина:

$$r(y)\equiv Ay-b \perp \mathscr{K}^m. \tag{16.24}$$

Полагая $y=Q_m s$, где $Q_m=[q_1|\dots|q_m]$ и $q_1, \dots, q_m$ — векторы Ланцоша, выводим из (16.24) равенство $(Q_m^* A Q_m)s=Q_m^* b$, или (см. (15.21)—(15.23))

$$T_m s=\beta_0 e_1^{(m)}, \tag{16.25}$$

где $e_1^{(m)}$ — 1-й координатный вектор размерности m. Как и при вычислении собственных значений, оценить качество приближенного решения y можно, не формируя его в явном виде. Действительно, умножая обе части равенства (15.22) справа на вектор s и учитывая (16.25), получаем

$$r(y)=Ay-b=AQ_m s-Q_m\beta_0 e_1^{(m)}=AQ_m s-Q_m T_m s=r_{m+1}\{s\}_m,$$

откуда $\|r(y)\|_2=\beta_{m-1}|\{s\}_m|$.

Сходным образом можно использовать и другие методы типа Ланцоша. В частности, различные применения алгоритма Арнольди к решению линейных систем — как для вычисления итерационных параметров других методов, например чебышевской рекурсии, так и для уточнения приближенных решений по принципу Галёркина — рассмотрены в серии статей Саада [98, 173, 176].

§ 17. Двусторонний метод Ланцоша с «заглядыванием вперед»

В первом приближении двусторонний метод Ланцоша, называемый также биортогональным алгоритмом, описан в § 15 (см. формулы (15.14)—(15.17)). Однако в этом описании метод рассматривается как способ трехдиагонализации неэрмитовых матриц; кроме того, не был рассмотрен случай, когда скалярное произведение $\omega_{m+1}=(r_{m+1}, s_{m+1})$ равно нулю. Именно этому случаю посвящен в основном данный параграф. Но вначале мы обсудим, следуя [132], некоторые вопросы, важные для всех вариантов программной реализации метода.

Пусть проделано m шагов процесса, описываемого формулами (15.14). Положим

$$V_m=[v_1|\dots|v_m], \qquad W_m=[w_1|\dots|w_m], \tag{17.1}$$

$$T_m=\begin{bmatrix} \alpha_1 & \gamma_2 & & & \\ \beta_2 & \ddots & \ddots & & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \gamma_m \\ 0 & & & \beta_m & \alpha_m \end{bmatrix}.$$

Если говорить о точной арифметике, то выполняются соотношения биортогональности

$$W_m^* V_m=I_m. \tag{17.2}$$

Кроме того, сами формулы (15.14) за m шагов можно объединить матричными равенствами

$$AV_m-V_m T_m=R_m, \qquad W_m^* A-T_m W_m^*=S_m^*, \tag{17.3}$$

где $R_m=[0|\dots|0|\beta_{m+1}v_{m+1}]$, $S_m=[0|\dots|0|\gamma_{m+1}w_{m+1}]$.

Пусть θ, z, t — *собственная тройка* матрицы T_m:

$$T_m z = \theta z, \qquad T_m^* t = \bar{\theta} t.$$

Будем считать дополнительно, что $(z, t) = 1$. Полагая $p = V_m z$, $q = W_m t$, выводим из (17.3)

$$\begin{aligned} &A V_m z - V_m T_m z = Ap - \theta p = \beta_{m+1} \{z\}_m v_{m+1}, \\ &t^* W_m^* A - t_m^* T_m W_m^* = q^* A - \theta q^* = \bar{\gamma}_{m+1} \{\bar{t}\}_m w_{m+1}^*. \end{aligned} \tag{17.4}$$

Если рассматривать p и q соответственно как правый и левый приближенные собственные векторы матрицы A, то равенства (17.4) указывают отвечающие им (и числу θ) невязки. Привлекая теорему 7.5, можно утверждать, что тройка (θ, p, q) является точной собственной тройкой для возмущенной матрицы $B = A + F$, где

$$\| F \|_2 = \max \left\{ \frac{|\beta_{m+1} \{z\}_m| \, \| v_{m+1} \|_2}{\| p \|_2}, \frac{|\gamma_{m+1} \{t\}_m| \, \| w_{m+1} \|_2}{\| q \|_2} \right\}. \tag{17.5}$$

Пусть собственное значение λ матрицы A, приближаемое числом θ,— простое. Если норма $\| F \|_2$ мала, то с точностью до величины порядка $O(\| F \|^2)$

$$|\lambda - \theta| \approx k(\lambda) \| F \|_2.$$

Меняя ролями A и B, можно считать λ возмущением числа θ; тогда число обусловленности $k(\theta) = \| p \|_2 \| q \|_2$ (здесь учтено нормировочное условие $(p, q) = 1$) и

$$|\lambda - \theta| \approx \| F \|_2 \| p \|_2 \| q \|_2. \tag{17.6}$$

При пользовании формулами (17.5), (17.6) нужны значения для $\| p \|_2$ и $\| q \|_2$. Чтобы не формировать сами векторы p, q — эта работа может оказаться напрасной, если величина погрешности в θ больше заданного критерия,— целесообразно заменить их длины оценками

$$\| p \|_2 \leqslant \| V_m \|_E^{} \| z \|_2, \qquad \| q \|_2 \leqslant \| W_m \|_E \| t \|_2. \tag{17.7}$$

Значения для $\| V_m \|_E^2$ и $\| W_m \|_E^2$ легко пересчитываются от шага к шагу.

Заметим, что вывод выражений (17.4) для невязок не опирался на условия биортогональности (17.2), которые при практическом счете не будут выполнены даже приближенно. Напротив, равенства (17.3) с точностью до малых погрешностей остаются верными. Только они и нужны для формулы (17.5). Хотя векторы p и q теперь не будут бинормированными: $(p, q) \neq 1$, соотношением (17.6) (вместе с (17.7)) все же можно пользоваться при предварительной проверке условия выхода. Впоследствии, когда p и q будут вычислены в явном виде, можно будет учесть и скалярное произведение (p, q).

Теорема 7.5 полезна еще в одном отношении: она позволяет понять, почему после того, как в спектре матрицы T_m появляется хорошее приближение к некоторому собственному значению λ матрицы A, в спектре каждой последующей матрицы T_l $(l = m+1, m+2, \ldots)$ есть приближение к λ сравнимого качества. Этот факт для

симметричных матриц был подмечен и объяснен Пейджем. Объяснение в неэрмитовом случае дается сходным образом.

Пусть (θ, z, t) — по-прежнему собственная тройка матрицы T_m. Дополняя векторы z, t нулями до l-мерных векторов $\tilde{z}$ и $\tilde{t}$, будем рассматривать $(\theta, \tilde{z}, \tilde{t})$ как приближенную собственную тройку матрицы T_l. Легко проверить, что невязка $T_l\tilde{z}-\theta\tilde{z}$ имеет единственную ненулевую компоненту (а именно $(m+1)$-ю), равную $\beta_{m+1}\{z\}_m$. Точно так же в невязке $T_l^*\tilde{t}-\bar{\theta}\tilde{t}$ только одна компонента отлична от нуля и равна $\bar{\gamma}_{m+1}\{t\}_m$. По теореме 7.5 тройка $(\theta, \tilde{z}, \tilde{t})$ будет точной собственной тройкой для матрицы T_l+G_l, где

$$\|G_l\|_2 = \max\left\{\frac{|\beta_{m+1}\{z\}_m|}{\|z\|_2}, \frac{|\gamma_{m+1}\{t\}_m|}{\|t\|_2}\right\}.$$

Как видим, норма матрицы-возмущения G_l не зависит от l. Поскольку по предположению последние компоненты векторов z и t малы — они должны обеспечивать малость возмущения (17.5), — то малы и все матрицы G_l.

Мы переходим к исследованию *обрывов*, т. е. ситуаций, когда очередное произведение ω_{m+1} обращается в нуль. В этой части параграфа будет изложено содержание статьи [163].

Равенство $\omega_{m+1}=0$ может иметь три причины: 1) $r_{m+1}=0$; 2) $s_{m+1}=0$; 3) ненулевые векторы r_{m+1} и s_{m+1} ортогональны. В первом случае *правые векторы Ланцоша* $v_1, \ldots, v_m$ определяют инвариантное подпространство, и с помощью накопленной за m шагов матрицы T_m можно найти группу собственных значений матрицы A. Точно так же при $s_{m+1}=0$ получаем левое инвариантное подпространство (т. е. инвариантное подпространство сопряженной матрицы A^*) как линейную оболочку *левых векторов Ланцоша* $w_1, \ldots, w_m$ и опять-таки с помощью T_m вычисляем часть собственных значений для A. Однако в третьем случае ни та, ни другая линейная оболочка инвариантным подпространством не будет, а между тем счет по формулам (15.14) продолжить не удастся, так как хотя бы один из коэффициентов β_{m+1}, γ_{m+1} равен нулю и формулы не определяют v_{m+1} либо w_{m+1}. Эту ситуацию Уилкинсон назвал *серьезным обрывом* в отличие от безобидных обрывов первых двух типов. Возможность серьезных обрывов значительно изменяет к худшему положение дел в алгоритме Ланцоша для неэрмитовых матриц. Действительно, никто пока не предложил лучшего выхода из серьезного обрыва, как повторить процесс Ланцоша с другими начальными приближениями $\tilde{v}_1$, $\tilde{w}_1$. При этом из первого прогона никакой спектральной информации относительно A не получено, как нет и информации по поводу более удачного выбора начальных векторов. В результате и новый прогон алгоритма не застрахован от серьезного обрыва.

Смысл работы [163] в том, чтобы предложить прием «заглядывания вперед», существенно уменьшающий вероятность натолкнуться на серьезный обрыв. Платой за это является частичная потеря трехдиагональности матрицей T: она искажается «горбами» всякий раз, когда заглядывание вперед помогает предотвратить обрыв.

Суть предлагаемого приема будет легче понять, если взглянуть на двусторонний алгоритм Ланцоша как на своеобразный вариант процесса Грама—Шмидта, в котором речь идет не об обычной ортогонализации, а о построении двух биортогональных систем. В общем случае описание такой биортогонализации было бы следующим. Заданы две невырожденные $n \times n$-матрицы: $F=[f_1|\ldots|f_n]$ и $G=[g_1|\ldots|g_n]$. Нужно построить системы векторов $v_1, \ldots, v_n$ и $w_1, \ldots, w_n$, такие, что для матриц $V_n=[v_1|\ldots|v_n]$ и $W_n=[w_1|\ldots|w_n]$ справедливо равенство

$$W_n^* V_n=\Psi_n=\operatorname{diag}(\psi_1, \ldots, \psi_n), \tag{17.8}$$

причем для всякого $j=1, \ldots, n-1$

$$\operatorname{span}\{v_1, \ldots, v_j\}=\operatorname{span}\{f_1, \ldots, f_j\},$$
$$\operatorname{span}\{w_1, \ldots, w_j\}=\operatorname{span}\{g_1, \ldots, g_j\}.$$

Построение осуществляется в n шагов, причем 1-й шаг состоит в простом присваивании $v_1=f_1$, $w_1=g_1$, $\psi_1=(v_1, w_1)$, а на j-м шаге вычисления ведутся по формулам

$$v_j=f_j-\sum_{i=1}^{j-1}(f_j, w_i)v_i/\psi_i, \quad w_j=g_j-\sum_{i=1}^{j-1}(v_i, g_j)w_i/\psi_i, \quad \psi_j=(v_j, w_j). \tag{17.9}$$

Обрыв процесса биортогонализации наступает при появлении нулевого коэффициента ψ_j. То, что обрыв возможен для невырожденных матриц F и G, иллюстрируется следующим примером.

Пример 17.1 (см. [163]). Пусть

$$F=\begin{bmatrix}1 & 1 & 1\\ 1 & 2 & 1\\ 1 & 1 & 2\end{bmatrix}, \qquad G=\begin{bmatrix}1 & 1 & 1\\ 0 & 0 & 1\\ 0 & 1 & 0\end{bmatrix}.$$

Шаг 1: $v_1=f_1$, $w_1=g_1$, $\psi_1=1$.

Шаг 2: $v_2=f_2-v_1=(0, 1, 0)^{\mathrm{T}}$, $w_2=g_2-2w_1=(-1, 0, 1)^{\mathrm{T}}$, $\psi_2=0$.

Отметим, что если не требовать невырожденности матриц F и G, то обрыв может наступить в связи с обращением v_j либо w_j в нулевой вектор. Обрыв такого сорта выявляет линейные зависимости в системе столбцов матрицы F или G. Нас же интересует *серьезный обрыв*, когда $\psi_j=0$ при ненулевых векторах v_j и w_j. Как продолжать процесс биортогонализации в этом случае?

Предлагается перевычислить вектор v_j, заменив в первой формуле из (17.9) f_j на f_{j+1}. Если и для нового вектора v_j скалярное произведение ψ_j равно нулю, то можно «попробовать» вектор f_{j+2} и т. д. Если матрица F не вырождена, то для некоторого $i>0$ вектор f_{j+i} должен привести к ненулевому значению для ψ_j. В самом деле, $\psi_j=(v_j, w_j)=(f_j, w_j)$ в силу ортогональности вектора w_j к $v_1, \ldots, v_{j-1}$. Поэтому $\psi_j=0$ означает, что $w_j \perp f_j$. Если и последующие векторы f_{j+i} «неудачны», то w_j ортогонален и к векторам $f_{j+1}, \ldots, f_n$. Ортогональность к $v_1, \ldots, v_{j-1}$ равносильна ортогональности к $f_1, \ldots, f_{j-1}$. Таким образом, вектор w_j ортогонален ко всем векторам $f_1, \ldots, f_n$, и, поскольку $w_j \neq 0$, эти последние векторы должны быть линейно зависимы.

В двустороннем методе Ланцоша роль матриц F_n и G_n играют крыловские матрицы

$$K=[v_1\,|\,Av_1\,|\,\ldots\,|\,A^{n-1}v_1],\quad \tilde{K}=[w_1\,|\,A^*w_1\,|\,\ldots\,|\,(A^*)^{n-1}w_1]. \tag{17.10}$$

Этот специальный выбор обеспечивает трехчленность формул рекурсии (15.14). Используется также то обстоятельство, что $\operatorname{span}\{v_1, \ldots, v_j, Av_j\}=\operatorname{span}\{v_1, \ldots, v_j, A^jv_1\}$, и аналогично для векторов w_i.

Рассмотрим матрицу

$$M=M(v_1, w_1)=\tilde{K}^*K. \tag{17.11}$$

Ее элементами являются числа

$$m_{ij}=w_1^*A^{i+j-2}v_1, \qquad i, j=1, \ldots, n, \tag{17.12}$$

называемые *моментами* матрицы A (относительно v_1 и w_1). Матрица M называется *матрицей моментов*. Понятно, что матрица M симметрична; более того, она — хотя это и не важно для нас — имеет *ганкелеву структуру*: значение ее элемента зависит только от суммы индексов $i+j$.

Оказывается, что метод Ланцоша тесно связан с треугольным разложением матрицы моментов, а серьезный обрыв при биортогонализации крыловских матриц соответствует появлению нуля в последовательности ведущих главных миноров матрицы M. Другими словами, обрыв наступает тогда, когда нельзя продолжить треугольное разложение.

Чтобы обосновать сделанное утверждение, заметим, что векторы v_{m+1} и w_{m+1} можно представить в виде

$$v_{m+1}=\mu_m(A)v_1/\prod_{j=2}^{m+1}\beta_j, \qquad w_{m+1}=\bar{\mu}_m(A^*)w_1/\prod_{j=2}^{m+1}\bar{\gamma}_j. \tag{17.13}$$

Здесь $\mu_m(t)$ — характеристический многочлен трехдиагональной матрицы T_m: $\mu_m(t)=\det(tI-T_m)$, а $\bar{\mu}_m(t)$ — многочлен с сопряженными коэффициентами (т. е. характеристический многочлен сопряженной матрицы T_m^*). Формулы (17.13) нетрудно доказать, сличая (15.14) с трехчленными соотношениями, связывающими характеристические многочлены последовательных матриц T_m.

Если обрывов в процессе Ланцоша не происходит, то в совокупности для $m=0, 1, \ldots, n-1$ формулы (17.13) эквивалентны матричным равенствам

$$V=[v_1\,|\,\ldots\,|\,v_n]=KR\mathrm{B}, \qquad W=[w_1\,|\,\ldots\,|\,w_n]=\tilde{K}\bar{R}\bar{\Gamma}, \tag{17.14}$$

где R — верхняя унитреугольная матрица, составленная по столбцам из коэффициентов многочленов μ_m, а B и Γ — диагональные матрицы:

$$\mathrm{B}^{-1}=\operatorname{diag}(1, \beta_2, \beta_2\beta_3, \ldots, \beta_2\beta_3\ldots\beta_n),$$
$$\Gamma^{-1}=\operatorname{diag}(1, \gamma_2, \gamma_2\gamma_3, \ldots, \gamma_2\gamma_3\ldots\gamma_n).$$

Подставляя выражения (17.14) в соотношение биортогональности $W^*V=I$, получаем

$$\Gamma R^{\mathrm{T}}\tilde{K}^*KR\mathrm{B}=I,$$

или

$$M = R^{-\mathrm{T}}\Gamma^{-1}\mathbf{B}^{-1}R^{-1} = L\Omega L^{\mathrm{T}}. \tag{17.15}$$

Мы положили здесь $L = R^{-\mathrm{T}}$, $\Omega = \Gamma^{-1}\mathbf{B}^{-1} = \mathrm{diag}(1, \omega_2, \omega_2\omega_3, \ldots, \omega_2\omega_3\ldots\omega_n)$. Это и есть искомое треугольное разложение. При его вычислении произведение $\omega_2\ldots\omega_m$ играет роль главного элемента m-го шага.

Обращение ω_m в нуль означает вырождение главного элемента; в этом случае, чтобы продолжить разложение, нужно было бы переставить в M строки или столбцы. В силу равенства

$$\mathscr{P}_{ml} M \mathscr{P}_{ml} = (\tilde{K}\mathscr{P}_{ml})^{*}(K\mathscr{P}_{ml})$$

перестановка строк (столбцов) в M равносильна такой же перестановке столбцов в $\tilde{K}(K)$. Это и есть «заглядывание вперед» — использование информации о еще не вычисленных векторах Ланцоша.

Чем дальше глядеть вперед, тем дороже это обходится. В [163] предлагается заглядывать ровно на один шаг.

Пусть проведено $i-1$ шагов обычного двустороннего алгоритма и вычислены векторы r_i, s_i и число $\omega_i = (r_i, s_i)$. Выбрав значения для β_i и γ_i и поделив на них r_i и s_i, мы получили бы очередные векторы Ланцоша v_i, w_i. Однако мы не станем пока делать этого, а поищем векторы r_{i+1}, s_{i+1} из условий

$$\mathrm{span}\,\{r_i, r_{i+1}\} = \mathrm{span}\,\{v_i, v_{i+1}\}, \quad \mathrm{span}\,\{s_i, s_{i+1}\} = \mathrm{span}\,\{w_i, w_{i+1}\}. \tag{17.16}$$

Простейший выбор для r_{i+1} и s_{i+1} описывается формулами

$$r_{i+1} = Ar_i - \omega_i v_{i-1}, \quad s_{i+1} = A^{*}s_i - \bar{\omega}_i w_{i-1}. \tag{17.17}$$

В самом деле, $r_{i+1} \in \mathscr{K}^{i+1}$, причем $(r_{i+1}, w_j) = (Ar_i, w_j) = (r_i, A^{*}w_j) = 0$, если $j < i-1$, и $(r_{i+1}, w_{i-1}) = (Ar_i, w_{i-1}) - \omega_i = (r_i, A^{*}w_{i-1}) - \omega_i = (r_i, \bar{\gamma}_i w_i + \ldots) - \omega_i = (r_i, \bar{\gamma}_i w_i) - \omega_i = (r_i, s_i) - \omega_i = 0$. Таким образом, оба вектора r_i и r_{i+1} принадлежат двумерному подпространству крыловского пространства $\mathscr{K}^{i+1}$, выделяемому условиями ортогональности к $w_1, \ldots, w_{i-1}$. Это оправдывает первое равенство из (17.16). Аналогично проверяется второе.

На i-м шаге разложения главный элемент матрицы M, т. е. элемент в позиции (i, i), равен $\omega_2\omega_3\ldots\omega_i$. Мы хотим сравнить его с тремя соседними элементами в позициях $(i, i+1)$, $(i+1, i)$, $(i+1, i+1)$ и по результатам сравнения выбрать способ продолжения процесса.

Чтобы найти нужные элементы, воспользуемся рассуждением, аналогичным выводу формул (17.14). К равенствам (17.13), взятым для $m = 0, 1, \ldots, i-2$, добавим выражения

$$\begin{aligned} r_i &= \mu_{i-1}(A)v_1 \Big/ \prod_{j=2}^{i-1}\beta_j, \qquad s_i = \bar{\mu}_{i-1}(A^{*})w_1 \Big/ \prod_{j=2}^{i-1}\bar{\gamma}_j, \\ r_{i+1} &= \varphi_i(A)v_1 \Big/ \prod_{j=2}^{i-1}\beta_j, \qquad s_{i+1} = \bar{\psi}_i(A^{*})w_1 \Big/ \prod_{j=2}^{i-1}\bar{\gamma}_j. \end{aligned} \tag{17.18}$$

Здесь φ_i, ψ_i — некоторые многочлены степени i со старшим коэффициентом 1. Последние два соотношения вытекают из (17.16), (17.17). Все это вместе эквивалентно матричным равенствам

$$V_i=[v_1|\ldots|v_{i-1}|r_i|r_{i+1}|A^{i+1}v_1|\ldots|A^{n-1}v_1]=KT_i\mathrm{B}_i,$$
$$W_i=[w_1|\ldots|w_{i-1}|s_i|s_{i+1}|(A^*)^{i+1}w_1|\ldots|(A^*)^{n-1}w_1]=\tilde{K}\bar{U}_i\bar{\Gamma}_i, \qquad (17.19)$$

где T_i, U_i — верхние унитреугольные матрицы вида

$$T_i=\begin{bmatrix}\Delta_{i+1} & 0\\ 0 & I_{n-i-1}\end{bmatrix}, \qquad U_i=\begin{bmatrix}\theta_{i+1} & 0\\ 0 & I_{n-i-1}\end{bmatrix},$$

$$\mathrm{B}_i^{-1}=\mathrm{diag}\,(1,\ \beta_2,\ \beta_2\beta_3,\ \ldots,\ \overbrace{\beta_2\ldots\beta_{i-1},\ \beta_2\ldots\beta_{i-1},\ \beta_2\ldots\beta_{i-1}}^{\text{3 раза}},\ 1,\ \ldots,\ 1),$$

$$\Gamma_i^{-1}=\mathrm{diag}\,(1,\ \gamma_2,\ \gamma_2\gamma_3,\ \ldots,\ \gamma_2\ldots\gamma_{i-1},\ \gamma_2\ldots\gamma_{i-1},\ \gamma_2\ldots\gamma_{i-1},\ 1,\ \ldots,\ 1).$$

Перемножая равенства (17.19), получаем

$$W_i^*V_i=\Gamma_iU_i^*\tilde{K}^*KT_i\mathrm{B}_i=\Gamma_iU_i^*MT_i\mathrm{B}_i. \qquad (17.20)$$

Системы v_1, ..., v_{i-1} и w_1, ..., w_{i-1} биортогональны; векторы r_i и r_{i+1} по построению ортогональны к w_1, ..., w_{i-1}, а векторы s_i и s_{i+1} — к v_1, ..., v_{i-1}. Поэтому ведущая главная подматрица порядка $i+1$ в $W_i^*V_i$ выглядит так:

$$\begin{bmatrix}1 & & & & & \\ & 1 & & & 0 & \\ & & \ddots & & & \\ & & & 1 & & \\ 0 & & & (r_i,s_i) & (r_{i+1},s_i) \\ & & & (r_i,s_{i+1}) & (r_{i+1},s_{i+1})\end{bmatrix}.$$

Нас интересует только нижний блок, а именно

$$Z_i=\begin{bmatrix}(r_i,\ s_i) & (r_{i+1},\ s_i)\\ (r_i,\ s_{i+1}) & (r_{i+1},\ s_{i+1})\end{bmatrix}. \qquad (17.21)$$

Из равенства (17.20) следует, что таков же с точностью до множителя $\omega_2\ldots\omega_{i-1}$ одноименный блок в матрице M, преобразованной $i-1$ шагами разложения.

Если элемент $\omega_i=(r_i,\ s_i)$ достаточно велик, то безопасно пользоваться обычными формулами (15.14). Возвращаясь к r_i и s_i, нормируем их делением на β_i и γ_i, после чего i-й шаг процесса можно считать законченным (если не включать в него, как это сделано в п. 1 дополнений к § 17, некоторые подготовительные вычисления для следующего шага). Лишние матрично-векторные умножения, потребовавшиеся для формул (17.17), в действительности не пропали даром — они будут использованы на шаге $i+1$.

Положим

$$\omega_l=(r_i,\ s_l),\quad \theta_l\equiv(r_{i+1},s_l)=(r_i,\ s_{i+1}),\quad \omega_{i+1}=(r_{i+1},\ s_{i+1}). \qquad (17.22)$$

Обычное продолжение процесса Ланцоша соответствует обычному же треугольному разложению матрицы (17.21):

$$Z_i = X_i Y_i = \begin{bmatrix} 1 & 0 \\ \theta_i/\omega_i & 1 \end{bmatrix} \begin{bmatrix} \omega_i & \theta_i \\ 0 & \omega_{i+1} - \theta_i^2/\omega_i \end{bmatrix}. \tag{17.23}$$

Альтернативный вариант — перестановка строк в Z перед исключением:

$$Z = \tilde{X}_i \tilde{Y}_i = \begin{bmatrix} \omega_i/\theta_i & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \theta_i & \omega_{i+1} \\ 0 & \theta_i - \omega_i \omega_{i+1}/\theta_i \end{bmatrix}. \tag{17.24}$$

Если выбран этот вариант, то в качестве следующих *двух пар* векторов Ланцоша принимаются

$$\begin{aligned} \tilde{V}_i &= [v_i \mid v_{i+1}] = [r_i \mid r_{i+1}]\, \tilde{Y}_i^{-1} \equiv R_i \tilde{Y}_i^{-1}, \\ \tilde{W}_i &= [w_i \mid w_{i+1}] = [s_i \mid s_{i+1}]\, \tilde{X}_i^{-*} \equiv S_i \tilde{X}_i^{-*}. \end{aligned} \tag{17.25}$$

Действительно, эти пары находятся в нужных двумерных подпространствах и биортогональны: $\tilde{W}_i^* \tilde{V}_i = \tilde{X}_i^{-1} S_i^* R_i \tilde{Y}_i^{-1} = \tilde{X}_i^{-1} Z_i \tilde{Y}_i^{-1} = I_2$.

Определим (наименьшие) углы между одноименными векторами v_i, w_i и v_{i+1}, w_{i+1}. Равенства (17.25) означают, что $v_i \| r_i$, $v_{i+1} \| \tilde{r}_{i+1} \equiv r_{i+1} - \omega_{i+1} r_i/\theta_i$, $w_i \| s_{i+1}$, $w_{i+1} \| \tilde{s}_i \equiv s_i - \bar{\omega}_i s_{i+1}/\bar{\theta}_i$. Следовательно,

$$\psi_i \equiv |\cos(v_i, w_i)| = |\cos(r_i, s_{i+1})| = \frac{|\theta_i|}{\|r_i\|_2 \|s_{i+1}\|_2}, \tag{17.26}$$

$$\psi_{i+1} \equiv |\cos(v_{i+1}, w_{i+1})| = \frac{|\theta_i - \omega_i \omega_{i+1}/\theta_i|}{\left\| r_{i+1} - \dfrac{\omega_{i+1}}{\theta_i} r_i \right\|_2 \left\| s_i - \dfrac{\bar{\omega}_i}{\bar{\theta}_i} s_{i+1} \right\|_2}.$$

Отметим — это важно с точки зрения требований к памяти, — что для пользования второй формулой из (17.26) нет необходимости в формировании векторов $\tilde{r}_{i+1}$ и $\tilde{s}_i$. В самом деле,

$$\|\tilde{s}_i\|_2^2 = \|s_i\|_2^2 - 2\,\mathrm{Re}\left[\frac{\omega_i}{\theta_i}(s_i, s_{i+1})\right] + \left|\frac{\omega_i}{\theta_i}\right|^2 \|s_{i+1}\|_2^2,$$

$$\|\tilde{r}_{i+1}\|_2^2 = \|r_{i+1}\|^2 - 2\,\mathrm{Re}\left[\frac{\omega_{i+1}}{\theta_i}(r_i, r_{i+1})\right] + \left|\frac{\omega_{i+1}}{\theta_i}\right|^2 \|r_i\|_2^2.$$

Таким образом, нужны длины векторов r_i, r_{i+1}, s_i, s_{i+1} и скалярные произведения (s_i, s_{i+1}), (r_i, r_{i+1}) помимо уже вычисленных значений ω_i, θ_i, ω_{i+1}.

В качестве меры качества (меры линейной независимости) биортогональных систем $v_1, \ldots, v_k$ и $w_1, \ldots, w_k$ примем величину

$$\varphi = \min_{1 \le j \le k} |\cos(v_j, w_j)|. \tag{17.27}$$

Наибольшее значение 1 она принимает, если системы $\{v\}$ и $\{w\}$ совпадают (с точностью до числовых множителей). Чем ближе φ к нулю, тем хуже: в системах имеется почти ортогональная пара одноименных векторов (точная ортогональность невозможна, поскольку $(v_j, w_j) = 1$), и, когда эти системы будут достроены до базисов пространства, базисы обязательно будут плохо обусловленными.

Теперь можно сформулировать критерий выбора продолжения — обычный шаг Ланцоша или двойной шаг (17.25). Прежде всего вычисляются

$$\varphi_1 = |\cos(r_i, s_i)|, \qquad \varphi_2 = \min\{|\psi_i|, |\psi_{i+1}|\}.$$

Если обе эти величины меньше заданного малого числа $\varepsilon > 0$, то алгоритм заканчивает работу, поскольку оба варианта продолжения приведут к плохим результатам. В противном случае производится сравнение

$$\varphi_1 \geqslant \varkappa \varphi_2,$$

где $\varkappa$ — также заданное положительное число. Если это условие выполнено, то делается одинарный шаг, иначе — двойной.

Для последующего удобно представить информацию относительно i-го шага в блочной форме:

$$\begin{aligned} [r_i | r_{i+1}] &= A[v_{i-1} | r_i] - [v_{i-1} | r_i] \begin{bmatrix} \alpha_{i-1} & \omega_i \\ 0 & 0 \end{bmatrix} - v_{i-2}(\gamma_{i-1}\ 0), \\ [s_i | s_{i+1}] &= A^*[w_{i-1} | s_i] - [w_{i-1} | s_i] \begin{bmatrix} \bar{\alpha}_{i-1} & \bar{\omega}_i \\ 0 & 0 \end{bmatrix} - w_{i-2}(\bar{\beta}_{i-1}\ 0). \end{aligned} \tag{17.28}$$

Напомним, что в случае двойного шага действуют формулы (17.24)—(17.25). Вместе с (17.28) они означают, что, хотя вектор Av_{i-1} по-прежнему выражается линейной комбинацией векторов v_{i-2}, v_{i-1} и v_i, для вектора Av_i в линейной комбинации будет уже четыре члена — с v_{i-2}, v_{i-1}, v_i и v_{i+1}. Следовательно, в i-м столбце матрицы T трехдиагональная форма будет нарушена. То же самое верно в отношении строк $i-1$ и i.

Поэтому матрицу T_m более естественно представлять себе как блочно-трехдиагональную матрицу

$$T_m = \begin{bmatrix} A_1 & \Gamma_2 & & & \\ B_2 & A_2 & \Gamma_3 & & 0 \\ & B_3 & A_3 & \ddots & \\ & & \ddots & \ddots & \Gamma_j \\ 0 & & & B_j & A_j \end{bmatrix} \tag{17.29}$$

с диагональными блоками первого и второго порядка. Блоки B_i могут иметь одну из следующих форм:

$$[*], \quad [0\ *], \quad \begin{bmatrix} * \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & * \\ 0 & 0 \end{bmatrix}.$$

Таковы же возможные размеры блоков Γ_i, однако все их элементы могут быть ненулевыми и, в случае матрицы 2×2, ранг должен быть равен единице.

Разбиению матрицы T на блоки соответствует группировка правых и левых векторов Ланцоша в виде $V_1, \ldots, V_j$ и $W_1, \ldots, W_j$, где блоки V_i, W_i суть обычные векторы, если i-й шаг одинарный, и $n \times 2$-матрицы,

если этот шаг двойной. (Отметим, что в (17.19) те же обозначения использованы для матриц другого вида.)

Подробная схема шага в методе Ланцоша с заглядыванием вперед приведена в п. 1 дополнений к § 17. Из нее следует, в частности, что для вычисления двух очередных пар векторов Ланцоша v_l, w_l и v_{l+1}, w_{l+1} в любом случае требуется четыре матрично-векторных умножения. Если назвать векторной операцией произвольное вычисление, сводящееся к n скалярным умножениям или делениям, то для продвижения на два одинарных шага нужны 30 векторных операций, а для одного двойного шага — 25 векторных операций. При двух шагах обычного алгоритма Ланцоша (без заглядывания вперед) выполняются те же четыре матрично-векторных умножения и 20 векторных операций. Таким образом, стоимость заглядывания вперед равна лишним пяти или десяти векторным операциям за два шага.

В [163] приведен пример (правда, только один), когда заглядывание вперед помогает преодолеть численную неустойчивость обычного алгоритма Ланцоша. Для матрицы

$$A=\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

собственными значениями являются все корни шестой степени из единицы. Если начальными векторами взять $v_1=w_1=(1, 2, 3, 4, 5, 6)^T$, то на 4-м шаге косинус угла между векторами v_4 и w_4 равен 3.025×10^{-8}, т. е. левый и правый вектор почти ортогональны. Это приводит к появлению в матрице T_4 элементов порядка 10^6; что еще хуже, тот же порядок приобретает ее спектральный радиус. Алгоритм с заглядыванием вперед выполняет в этом месте двойной шаг и не встречает проблем с ростом элементов. Собственные значения полученной в конце процесса блочно-трехдиагональной матрицы T_6 совпадают с собственными значениями матрицы A практически во всех разрядах.

В остальном обсуждаемом в [163] экспериментальном материале обращает на себя внимание следующее явление: в тех случаях, когда выполнялось двойное (по отношению к порядку матрицы) число шагов — будь то обычный алгоритм или алгоритм с «заглядыванием вперед»,— в спектре матрицы T_m присутствовали приближения ко *всем* собственным значениям матрицы A. При этом для обычного алгоритма характерно, что «лишние» числа Ритца θ также аппроксимировали — в качестве повторных копий — собственные значения матрицы A. Тем самым эти результаты снова подтверждают отмеченный в § 15 феномен Ланцоша. Более подробно он будет рассмотрен в следующем параграфе.

Метод Ланцоша в любом варианте предполагает неоднократное вычисление собственных значений матриц T_m, т. е. матриц трехдиагональной формы, обычной или блочной. В отличие от эрмитова случая QR-алгоритм не сохраняет специфику этой формы, превращая

ее после нескольких шагов в хессенбергову. Поэтому в [163] расчет спектра матриц T_m проводился посредством метода Лагерра для вычисления корней многочлена. Необходимые значения характеристического многочлена матрицы T_m определялись по схеме Хаймана. Комбинация методов Лагерра и Хаймана использовалась Парлеттом для спектральных вычислений еще в начале 60-х годов [153]. Краткое описание обоих методов дано в пп. 3, 4 дополнений к § 17; подробности можно найти в [42, гл. 7, § 14, 28].

ДОПОЛНЕНИЯ К § 17

1. Приведем (см. [163]) формализованное описание одного шага алгоритма Ланцоша с «заглядыванием вперед», считая для простоты все величины вещественными. Схема описания такова: предполагается, что на входе i-го шага имеются $n \times 2$-матрицы R_i, S_i, составленные из двух очередных векторов r_i, r_{i+1} и s_i, s_{i+1}; эти векторы и матрицы называются в [163] невязками. Алгоритм определяет, каким способом модифицировать невязки, превращает их в новые векторы Крылова, а затем частично вычисляет матрицы-невязки следующего шага. Таким образом, результатом шага будут новые матрицы V_i и W_i, т. е. попросту векторы v_i, w_i, если i-й шаг окажется обычным, и $n \times 2$-матрицы, если шаг двойной.

Итак, на входе i-го шага задаются: V_{i-1}, W_{i-1} (обе матрицы нулевые при $i=1$); число или двумерный вектор z_i, кратные 1-му столбцу матрицы $\Gamma_i (z_1 = 1)$; невязки r_i, s_i и числа $\omega \equiv \omega_i = (r_i, s_i)$, $\| r_i \|$, $\| s_i \|$. Норма векторов всюду евклидова. (В конце описания каждого этапа в скобках указывается его вычислительная стоимость.)

1. «Заглядывание вперед».

а) Вычислить $R_i = [r_i | r_{i+1}]$, $S_i = [s_i | s_{i+1}]$ по формулам

$$r_{i+1} = A r_i - \omega V_{i-1} z_i, \qquad s_{i+1} = A^T s_i - \omega w_{i-1}$$

(два матрично-векторных произведения и две-три векторные операции).

Комментарий. Сколько именно потребуется векторных операций, зависит от того, каким был предыдущий шаг: если обычным, то z_i — число, иначе z_i — двумерный вектор.

б) Вычислить необходимые скалярные произведения и длины:

$$\theta = (r_{i+1}, s_i) = (r_i, s_{i+1}), \qquad \hat{\omega} = (r_{i+1}, s_{i+1}),$$

$$(r_i, r_{i+1}), \qquad (s_i, s_{i+1}), \qquad \| r_{i+1} \|, \quad \| s_{i+1} \|$$

(шесть векторных операций).

в) Вычислить необходимые косинусы: $\varphi_1 = \cos (r_i, s_i) = \omega / (\| r_i \| \| s_i \|)$, $\varphi_2 = 0$; если $\theta = 0$, перейти к этапу 2; в противном случае $\tau_1 = \omega / \theta$, $\tau_2 = \hat{\omega} / \theta$,

$$\| \tilde{r}_{i+1} \| = \sqrt{\| r_{i+1} \|^2 - 2\tau_2 (r_i, r_{i+1}) + \tau_2^2 \| r_i \|^2},$$

$$\| \tilde{s}_i \| = \sqrt{\| s_i \|^2 - 2\tau_1 (s_i, s_{i+1}) + \tau_1^2 \| s_{i+1} \|^2},$$

$$\psi_1 = \cos (r_i, s_{i+1}) = \theta / (\| r_i \| \| s_{i+1} \|),$$

$$\psi_2 = \cos (\tilde{r}_{i+1}, \tilde{s}_i) = (\theta - \omega\tau_2) / (\| \tilde{r}_{i+1} \| \| \tilde{s}_i \|),$$

$$\varphi_2 = \min \{ |\psi_1|, |\psi_2| \}$$

(векторных операций нет).

2. Проверка условия выхода. Если $|\varphi_1| < \varepsilon$ и $\varphi_2 < \varepsilon$, где $\varepsilon > 0$ задано, то выход из алгоритма с сообщением об обрыве.

3. Выбор продолжения. Если $|\varphi_1| \geqslant \varkappa \varphi_2$, где $\varkappa > 0$ задано, то будет выполняться обычный шаг; в противном случае — двойной.

4. Шаг алгоритма Ланцоша. Операции, выполняемые в обычном и двойном шаге, описываются параллельно.

Обычный шаг	*Двойной шаг*
	а) Разложение $Z=\begin{bmatrix}\omega & \theta\\ \theta & \hat{\omega}\end{bmatrix}=XY$:
$\beta_l=\|r_l\|\sqrt{\lvert\varphi_1\rvert}$,	$\Delta=\operatorname{diag}(\beta_l,\ \beta_l\beta_{l+1})=$
$\gamma_l=\omega/\beta_l$	$=\operatorname{diag}\left(\|r_l\|\sqrt{\lvert\psi_1\rvert},\ \|r_{l+1}\|\sqrt{\lvert\psi_2\rvert}\right)$
(векторных операций нет).	

Комментарий. Выбор значений для β_l и γ_l в обычном шаге равносилен нормированию векторов v_l и w_l к одинаковой длине $1/\sqrt{|\varphi_1|}$. По аналогии с этим выбирается Δ — диагональная матрица нормирующих множителей для двойного шага. Введению матрицы Δ соответствует следующее разложение матрицы Z:

$$X=\begin{bmatrix}\tau_1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\theta & 0\\ 0 & \theta-\omega\tau_2\end{bmatrix}\Delta^{-1},\qquad Y=\Delta\begin{bmatrix}1 & \tau_2\\ 0 & 1\end{bmatrix}.$$

б) Вычисление новых матриц V_i, W_i:

$v_l=r_l/\beta_l$,	$V_i=R_iY^{-1}$,
$w_l=s_l/\gamma_l$	$W_i=S_iX^{-\mathrm{T}}$
(две векторные операции)	(шесть векторных операций)

Комментарий. Вычисление каждой из матриц V_i, W_i сводится к прибавлению к одному из столбцов матрицы R_i (или S_i) кратного другого столбца и нормировке обоих столбцов, поэтому векторных операций нужно шесть, а не восемь.

в) Достройка блоков Γ_i, B_i:

$\Gamma_i=\gamma_l z_i$,	$\Gamma_i=z_i\left[\omega\ \psi_2\right]\Delta^{-1}$,
$B_i=\beta_l$ либо $[0\ \beta_l]$	$B_i=\begin{bmatrix}\beta_l\\ 0\end{bmatrix}$ либо $\begin{bmatrix}0 & \beta_l\\ 0 & 0\end{bmatrix}$
(векторных операций нет)	

Комментарий. Если предыдущий шаг был обычным (двойным), то для B_i выбирается первое (второе) значение; соответственно Γ_i будет числом, вектор-столбцом, вектор-строкой или 2×2-матрицей в зависимости от конкретной комбинации шагов.

г) Формирование новых невязок:

$r_{l+1}\leftarrow r_{l+1}/\beta_l$,	$r_{l+2}=(AV_i-V_{i-1}\Gamma_i)\begin{bmatrix}0\\ 1\end{bmatrix}$,
$s_{l+1}\leftarrow s_{l+1}/\gamma_l$	$s_{l+2}=A^{\mathrm{T}}w_{l+1}-\beta_l w_{l-1}=$
	$=(A^{\mathrm{T}}W_l-W_{i-1}B_i^{\mathrm{T}}\pi_2)\begin{bmatrix}0\\ 1\end{bmatrix}$, $\pi_2=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$
(две векторные операции)	(два матрично-векторных произведения и две-три векторные операции)

Комментарий. Формула для r_{l+2} выражает вектор из крыловского подпространства $\mathcal{K}^{l+2}$, ортогональный к векторам $w_1, \ldots, w_{l-1}$. Количество

векторных операций в двойном шаге снова зависит от предыдущего шага, т. е. от строчного размера блока Γ_i. Так как в выражение для w_{l+1} входит s_{l+1}, то вектор $s_{l+2} \in \mathscr{L}^{l+2}$ и ортогонален к $v_1, \ldots, v_{l-1}$.

д) Вычисление блока A_i:

$$\alpha_l = 1/\tau_1 \qquad A_i = \begin{bmatrix} \tau_2 & (Av_{l+1}, w_l) \\ \beta_{l+1} & \dfrac{1}{\theta - \omega\tau_2}\left[\hat{\omega} - \tau_2\theta - \beta_{l+1}\tau_1\theta(Av_{l+1}, w_l)\right] \end{bmatrix}$$

(векторных операций нет) (одна векторная операция)

Комментарий. Вектор Av_{l+1} вычислен на этапе г), поэтому единственная векторная операция при построении блока A_i—это вычисление скалярного произведения (Av_{l+1}, w_l). Вывод выражений для α_l и элементов матрицы A_i приведен вслед за описанием алгоритма.

е) Биортогонализация:

$$r_{l+1} \leftarrow r_{l+1} - \alpha_l v_l, \qquad r_{l+2} \leftarrow r_{l+2} - V_i A_i \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

$$s_{l+1} \leftarrow s_{l+1} - \alpha_l w_l \qquad s_{l+2} \leftarrow s_{l+2} - W_i A_i^{\mathrm{T}} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

(две векторные операции) (четыре векторные операции)

ж) Вычисление скалярных произведений для следующего шага:

$$\|r_{l+1}\| \leftarrow \left(\|r_{l+1}\|^2 - 2\alpha_l(r_l, r_{l+1}) + \alpha_l^2\|r_l\|^2\right)^{1/2}/\beta_l,$$

$$\|s_{l+1}\| \leftarrow \left(\|s_{l+1}\|^2 - 2\alpha_l(s_l, s_{l+1}) + \alpha_l^2\|s_l\|^2\right)^{1/2}/\gamma_l,$$

$$\omega_{l+1} \leftarrow \hat{\omega}/\omega + \alpha_l^2$$

Вычислить $\|r_{l+2}\|$ и $\|s_{l+2}\|$, $\omega_{l+2} = (r_{l+2}, s_{l+2})$

(векторных операций нет) (три векторные операции)

з) Задание значения для z_{i+1}:

$$z_{i+1} = [1] \qquad z_{i+1} = \begin{bmatrix} 1 \\ -\dfrac{\beta_{l+1}\tau_1\theta}{\theta - \omega\tau_2} \end{bmatrix}$$

Комментарий. Объяснение последней формулы дано ниже.

Конец i-го шага

Отметим, что в режиме обычного шага два векторно-матричных умножения выполняются только на этапе «заглядывания вперед».

Проверим теперь формулы для α_l и элементов блока A_i. Имеем

$$\alpha_l = (Av_l, w_l) = \frac{1}{\beta_l\gamma_l}(Ar_l, s_l) = (r_{l+1} + \omega V_{i-1}z_i, s_l)/\omega = (r_{l+1}, s_l)/\omega = \theta/\omega = 1/\tau_1.$$

Согласно (17.29), для блока A_i справедливо представление

$$A_i = W_i^* A V_i = \begin{bmatrix} (Av_l, w_l) & (Av_{l+1}, w_l) \\ (Av_l, w_{l+1}) & (Av_{l+1}, w_{l+1}) \end{bmatrix}.$$

Так что формулу для элемента (1, 2) проверять не нужно. Что касается остальных трех формул, то для их проверки придется расписать подробно соотношения шага 4б):

$$v_l = r_l/\beta_l, \qquad v_{l+1} = (r_{l+1} - \tau_2 r_l)/(\beta_l \beta_{l+1}),$$
$$w_l = \beta_l s_{l+1}/\theta, \qquad w_{l+1} = \delta(s_l - \tau_1 s_{l+1}), \tag{17.30}$$

где $\delta = \beta_l \beta_{l+1}/(\theta - \omega\tau_2)$. Заметим еще, что

$$\theta - \omega\tau_2 = \theta - \omega\hat{\omega}/\theta = \theta - \hat{\omega}\tau_1.$$

Теперь имеем

$$(Av_l, w_l) = (Ar_l, s_{l+1})/\theta = (r_{l+1} + \omega V_{i-1} z_l, s_{l+1})/\theta = (r_{l+1}, s_{r+1})/\theta = \hat{\omega}/\theta = \tau_2,$$
$$(Av_l, w_{l+1}) = \delta(Ar_l, s_l - \tau_1 s_{l+1})/\beta_l =$$
$$= \delta(r_{l+1}, s_l - \tau_1 s_{l+1})/\beta_l = \frac{\beta_{l+1}}{\theta - \omega\tau_2}(\theta - \hat{\omega}\tau_1) = \beta_{l+1}.$$

Перепишем две последние формулы из (17.30) в виде

$$w_{l+1} = \delta s_l - (\delta\tau_1\theta) w_l/\beta_l. \tag{17.31}$$

Тогда

$$(Av_{l+1}, w_{l+1}) = \delta(Av_{l+1}, s_l) - \delta\tau_1\theta(Av_{l+1}, w_l)/\beta_l.$$

Далее

$$(Av_{l+1}, s_l) = (v_{l+1}, A^T s_l) = (v_{l+1}, s_{l+1} + \omega w_{l-1}) =$$
$$= (v_{l+1}, s_{l+1}) = (r_{l+1} - \tau_2 r_l, s_{l+1})/(\beta_l \beta_{l+1}) = (\hat{\omega} - \tau_2\theta)/(\beta_l \beta_{l+1}).$$

Остается заметнть, что $\delta/(\beta_l \beta_{l+1}) = 1/(\theta - \omega\tau_2)$, $\delta\tau_1\theta/\beta_l = \beta_{l+1}\tau_1\theta/(\theta - \omega\tau_2)$.

Наконец, объясним выражение для вектора z_{i+1} в случае двойного шага. Этот вектор есть кратное первого столбца будущей матрицы Γ_{i+1}. Точными значениями элементов первого столбца были бы (Av_{l+2}, w_l) и (Av_{l+2}, w_{l+1}). Но (см. (17.31))

$$(Av_{l+2}, w_{l+1}) = \delta(Av_{l+2}, s_l) - \delta\tau_1\theta(Av_{l+2}, w_l)/\beta_l.$$

При этом $(Av_{l+2}, s_l) = (v_{l+2}, A^T s_l) = (v_{l+2}, s_{l+1}) + \omega(v_{l+2}, w_{l-1}) = 0$. Итак,

$$(Av_{l+2}, w_{l+1}) = -\delta\tau_1\theta(Av_{l+2}, w_l)/\beta_l.$$

Следовательно, если первой компоненте вектора z_{l+1} присвоить значение 1, то вторая непременно будет равна $(-\delta\tau_1\theta)/\beta_l = -\beta_{l+1}\tau_1\theta/(\theta - \omega\tau_2)$.

2. По ряду причин удобней иметь дело с нормированными векторами Ланцоша: $\|v_m\|_2 = \|w_m\|_2 = 1$ для всех m. Переход к таким векторам требует некоторого (впрочем, несущественного) изменения формул метода. В частности, произведение $W_m^* V_m$ будет диагональной матрицей $\Phi_m = \mathrm{diag}(\varphi_1, ..., \varphi_m)$ с положительной диагональю, а не единичной матрицей. Как следствие, числа Ритца определяются уже не из трехдиагональной или блочно-трехдиагональной матрицы T_m, а из обобщенной проблемы вида $T_m s = \theta\Phi_m s$; однако комбинация методов Лагерра и Хаймана применима здесь в той же мере, что и к самой матрице T_m. К достоинствам этого варианта метода Ланцоша относится возможность простого контроля качества базисных матриц V_m, W_m крыловских подпространств. Именно если за меру качества принять спектральное число обусловленности $\mathrm{cond}_2 V_m = \|V_m\|_2 \|V_m^+\|_2$, то выполняется оценка $\mathrm{cond}_2 V_m \leqslant$ $\leqslant m/\min_{1 \leqslant l \leqslant m} \psi_i$. Та же граница справедлива для $\mathrm{cond}_2 W_m$. Если ортогональность векторов v_i и w_j с разными номерами нарушается в сильной степени — так

будет, когда какое-то число Ритца сойдется к собственному значению матрицы A,— оценки могут потерять силу. Чтобы можно было пользоваться ими и после момента сходимости, необходима перебиортогонализация векторов Ланцоша.

3. Метод Лагерра для вычисления корня комплексного многочлена $f(z)$ степени n имеет итерационную формулу

$$z_{k+1}=z_k-nf(z_k)/(f'(z_k)\pm H_k^{1/2}), \tag{17.32}$$

где

$$H_k=(n-1)^2[f'(z_k)]^2-n(n-1)f(z_k)f''(z_k).$$

В окрестности простого корня λ, выбирая знак в (17.32) так, чтобы увеличить модуль знаменателя, получим кубическую сходимость к λ. Кубическая сходимость сохранится и при замене n в формулах процесса любым другим фиксированным числом. Если же корень λ кратный, то сходимость будет лишь линейной.

Поскольку вычисление значения f'' сравнимо по трудоемкости с вычислением f и f', то метод Лагерра оказывается экономичнее метода Ньютона, в котором f'' не используется.

4. Метод Хаймана предназначен для вычисления значения в точке z характеристического многочлена хессенберговой матрицы A. Для определенности будем считать A верхней хессенберговой матрицей, причем неразложимой, иначе порядок можно было бы понизить. Положим $P\equiv A-zI=[p_1|p_2|\ldots|p_n]$, $f(z)=\det(A-zI)$. Пусть числа $x_1, \ldots, x_{n-1}$ таковы, что

$$x_1p_1+\ldots+x_{n-1}p_{n-1}+p_n=k(z)e_1, \tag{17.33}$$

где e_1 — первый координатный вектор. Тогда, сохраняя значение определителя, можно заменить в P последний столбец вектором $k(z)e_1$. Новый определитель тривиально вычисляется разложением по последнему столбцу, что приводит к равенству

$$f(z)=(-1)^{n+1}a_{21}a_{32}\ldots a_{n,n-1}k(z).$$

Если значение $f(z)$ используется в процедуре вычисления корня, то, игнорируя постоянный множитель, можем находить только значения функции $k(z)$ и ее производных.

Соотношение (17.33), расписанное покомпонентно, позволяет легко определить нужные числа $x_1, \ldots, x_{n-1}$, а затем и $k(z)$. Действительно, из равенства $a_{n,n-1}x_{n-1}+(a_{nn}-z)=0$ находим x_{n-1}, из равенства $a_{n-1,n-2}x_{n-2}+(a_{n-1,n-1}-z)x_{n-1}+a_{n-1,n}=0$ вычисляем x_{n-2} и т. д. Когда мы дойдем таким образом до равенства первых компонент, все числа $x_1, \ldots, x_{n-1}$ известны и $(x_n=1)$

$$k(z)=(a_{11}-z)x_1+a_{12}x_2+\ldots+a_{1n}x_n. \tag{17.34}$$

Если (как в методах Ньютона и Лагерра) нужны значения производных многочлена $k(z)$, то их можно получить аналогичным способом. Дифференцируя (17.34) и предыдущие равенства (в которые введено $x_n=1$)

$$a_{r,r-1}x_{r-1}+(a_{rr}-z)x_r+a_{r,r+1}x_{r+1}+\ldots+a_{rn}x_n=0, \qquad r=n, \ldots, 2,$$

приходим к соотношениям

$$a_{r,r-1}x'_{r-1}+(a_{rr}-z)x'_r+a_{r,r+1}x'_{r+1}+\ldots+a_{rn}x'_n=x_r, \qquad r=n, \ldots, 2, \tag{17.35}$$

$$(a_{11}-z)x'_1+a_{12}x'_2+\ldots+a_{1n}x'_n-x_1=k'(z). \tag{17.36}$$

Так как $x_n \equiv 1$, то $x'_n \equiv 0$, и из уравнения (17.35) последовательно находим x'_{n-1}, x'_{n-2}, ..., x'_1. После этого (17.36) определяет искомое значение $k'(z)$. Точно так же можно вычислить $k''(z)$.

В [42, гл. 7, § 12—14] показано, что метод Хаймана очень устойчив в смысле § 2 как вычислительный алгоритм. Именно получаемое им значение определителя является точным для матрицы, близкой к $A - zI$ и также хессенберговой. Априорная оценка матрицы эквивалентного возмущения настолько хороша, что ей уступает аналогичная оценка при вычислении определителя ортогональным методом вращений. Это тем более удивительно, что метод Хаймана эквивалентен последовательности элементарных (следовательно, неортогональных) преобразований, выполняемых без выбора главного элемента. Действительно, x_{n-1} можно интерпретировать как коэффициент при $(n-1)$-м столбце, такой, что вычитание этого кратного из последнего столбца аннулирует элемент (n, n); x_{n-2} есть аналогичный коэффициент $(n-2)$-го столбца, обеспечивающий при последующем вычитании из n-го столбца уничтожение элемента $(n-1, n)$, и т. д.

§ 18. Как использовать феномен Ланцоша

Теоретически процесс Ланцоша должен закончиться не более чем в n шагов получением нулевого значения для коэффициента β_{m+1} $(m \leqslant n)$. Если же ланцошева рекурсия ведется в машинной арифметике и вычисляемые векторы Ланцоша не подвергаются переортогонализации, то, как уже говорилось, значения β_m даже для $m > n$ не только не равны нулю, но и, как правило, не очень малы. Несмотря на такое очевидное расхождение с теорией, в спектре трехдиагональных матриц T_m со временем появляются приближения к каждому собственному значению исходной матрицы A. Чтобы найти приближения ко всем, подчас приходится выполнять число шагов, значительно превышающее n.

Возможность отказаться от переортогонализации векторов Ланцоша и тем не менее получать приближения к собственным значениям матрицы A — это и есть феномен Ланцоша, обнаруженный экспериментально американскими математиками Джейн Каллэм и Ральфом Уиллафби в конце 70-х годов. Он был положен ими в основу ряда программ, решающих спектральные задачи для симметричных и эрмитовых матриц, а также некоторые смежные задачи (например, вычисление сингулярных чисел и векторов больших разреженных матриц). Сводное описание этих программ и использованной в них методики дает книга [87].

В последующих экспериментах Каллэм и Уиллафби установили, что программная реализация биортогонального алгоритма, основанная на сходных принципах, также дает хорошие результаты [86, 88]. В настоящем параграфе будет кратко резюмировано основное содержание указанных публикаций.

Суть рассматриваемого подхода в том, чтобы разделить в спектре матриц T_m собственные значения, которые, по-видимому, могут интерпретироваться как приближения к собственным значениям матрицы A, и собственные значения, скорей всего приближениями не являющиеся. Условно назовем те и другие соответственно «хорошими» и «подозрительными». Насколько хороши «хорошие» собственные значения, судят по оценкам норм соответствующих невязок; для «подозрительных» собственных значений никакого оценивания не

проводится. Если нужное число собственных значений вычислено с требуемой точностью, процесс закончен. В противном случае значение m увеличивается, и описанные процедуры повторяются.

Придадим сказанному форму макропрограммы. Считаются заданными пользователем (или самой программой по умолчанию) значения: M — ограничитель числа шагов ланцошевой рекурсии; $m_1, ..., m_k$ — порядки трехдиагональных матриц T_m, для которых выполняется сортировка собственных значений; ε_1, ε_2, ε_3 — положительные числа, играющие роль допусков в программе. Через $\hat{T}_m$ обозначается матрица, полученная из T_m вычеркиванием первой строки и первого столбца.

Практическая процедура Ланцоша для неэрмитовых матриц (см. [88]).

Шаг 1. Провести M шагов ланцошевой рекурсии.

Для значений $m = m_1, m_2, ..., m_k$) выполнить следующее:

Шаг 2. Вычислить собственные значения $\theta_1, ..., \theta_m$ матрицы T_m. Выявить численно кратные собственные значения (собственные значения θ_s и θ_t считаются версиями одного и того же кратного собственного значения, если $|\theta_s - \theta_t| \leqslant \varepsilon_1$). Численно кратные собственные значения диагностируются как «хорошие».

Шаг 3. Вычислить собственные значения $\upsilon_1, ..., \upsilon_{m-1}$ матрицы $\hat{T}_m$. «Подозрительными» считаются те собственные значения матрицы T_m, которые находятся на расстоянии, не превышающем ε_2, от некоторого собственного значения υ_i.

Шаг 4. Вычислить оценку ошибки для каждого простого «хорошего» собственного значения. Если нужное число собственных значений вычислены с требуемой точностью (т. е. для них удовлетворен тест, заданный числом ε_3), то закончить выполнение программы. То же, если $m = M$. В противном случае увеличить значение m и перейти к шагу 2.

Дадим пояснения этому описанию. Прежде всего ланцошева рекурсия проводится в симметризованной форме:

$$\beta_{i+1} v_{i+1} = A v_i - \alpha_i v_i - \beta_i v_{i-1} \equiv r_{i+1},$$

$$\bar{\beta}_{i+1} w_{i+1} = A^* w_i - \bar{\alpha}_i w_i - \bar{\beta}_i w_{i-1} \equiv s_{i+1}, \tag{18.1}$$

где

$$\alpha_i \equiv (\alpha_i^v + \alpha_i^w)/2, \qquad \alpha_i^v \equiv (A v_i - \beta_i v_{i-1}, w_i),$$

$$\alpha_i^w \equiv (v_i, A^* w_i - \bar{\beta}_i w_{i-1}), \qquad \beta_{i+1}^2 \equiv (r_{i+1}, s_{i+1}).$$

Таким образом, неопределенность в выборе коэффициентов β_i, γ_i биортогонального алгоритма разрешается наложением условия симметрии $\beta_i = \gamma_i \; \forall_i$. Для обеих ланцошевых последовательностей выбирается один и тот же начальный вектор $v_1 = w_1$. Все эти решения направлены на то, чтобы, во-первых, симметризовать распространение ошибок и, во-вторых, получить *симметричные* матрицы T_m. Правда, эти матрицы будут комплексными, вообще говоря, даже для вещественной матрицы A.

Симметрия (в отличие от самосопряженности) и трехдиагональность комплексных матриц не сохраняются обычным, унитарным

QR-алгоритмом. Поэтому вычисление собственных значений матрицы T_m проводится посредством специальной комплексной версии QR-алгоритма. В ней основное QR-разложение использует комплексные ортогональные (а не унитарные) матрицы Q. Такое разложение не всегда существует, а когда существует, не всегда может быть найдено численно устойчивым образом (ситуация здесь схожа со стандартным треугольным разложением). Но зато комплексный QR-алгоритм поддерживает и симметрию, и трехдиагональность. В результате вычисление всех собственных значений матрицы T_m достигается за $O(m^2)$ операций.

Оценки погрешностей хороших собственных значений находят примерно так, как об этом говорилось в начале § 17 (см. формулы (17.5)—(17.7)); нужно только учесть, что вследствие симметрии матрицы T_m векторы z и t комплексно сопряжены.

Если z — собственный вектор неразложимой трехдиагональной матрицы T_m, то для его компоненты $\{z\}_m$ справедливо представление

$$(\{z\}_m)^2 = \varphi_{m-1}(\theta)/\varphi'_m(\theta). \tag{18.2}$$

Здесь θ — соответствующее собственное значение, φ_{m-1} и φ_m — характеристические многочлены матриц T_{m-1} и T_m. С другой стороны, имеет место равенство

$$\hat{\varphi}(\theta)\varphi_{m-1}(\theta) = \prod_{i=2}^{m} \beta_i^2, \tag{18.3}$$

где $\hat{\varphi}$ — характеристический многочлен матрицы $\hat{T}_m$. Оба эти соотношения доказаны в [87, гл. 3, § 2]. Согласно (17.4), (17.5), (18.2) и (18.3), выполняются приближенные равенства

$$\frac{\|Ap-\theta p\|}{\|p\|} \approx \frac{|\beta_{m+1}|\,\|v_{m+1}\|\,|\{z\}_m|}{\|p\|} = \frac{|\beta_{m+1}|\,\|v_{m+1}\|}{\|p\|}\left|\frac{\varphi_{m-1}(\theta)\hat{\varphi}(\theta)}{\varphi'_m(\theta)\hat{\varphi}(\theta)}\right|^{1/2} =$$

$$= \frac{\left|\prod_{i=2}^{m+1}\beta_i\right|\|v_{m+1}\|}{\|p\|\,|\varphi'_m(\theta)\hat{\varphi}(\theta|^{1/2}}.$$

Тест, отсеивающий подозрительные собственные значения от хороших, основан именно на полученном представлении нормы невязки. Если θ близко к корню многочлена $\hat{\varphi}$, можно ожидать больших значений этой нормы, и такое θ отвергается.

Обращает на себя внимание то, что в программах Каллэм — Уиллафби не принимается никаких предупредительных мер против серьезных обрывов. Исходя из своей численной практики, эти авторы считают возможность натолкнуться на такой обрыв крайне маловероятной.

ДОПОЛНЕНИЯ К § 18

1. То обстоятельство, что собственные значения трехдиагональных матриц Каллэм и Уиллафби вычисляют посредством комплексного QR-алгоритма, а не, скажем, методом Лагерра, как в [163], объясняется тем, что все необходимое программное обеспечение у них уже имелось. Один из разделов

книги [87] посвящен применению метода Ланцоша к решению спектральных задач для комплексных симметричных матриц. Ланцошева рекурсия здесь естественным образом приводит к симметричным трехдиагональным матрицам.

2. Объяснение феномена Ланцоша для симметричного случая можно найти в [87, гл. 4] и с другой точки зрения в [17].

§ 19. О выборе начальных приближений

Для всех методов этой главы выбор начальных приближений имеет немалое значение. От него зависит, насколько быстро будут получены и будут ли получены вообще желаемые собственные пары. Между тем кроме общих рекомендаций типа $\det Y_0^{(1)} \neq 0$ в методах одновременных итераций сказать по этому поводу нечего. Проверить же условия такого рода невозможно, поскольку, например, матрица $Y_0^{(1)}$ пользователю не известна.

Основу настоящего параграфа составляет статья Саада [175]. В ней для некоторого класса частичных задач на собственные значения, а именно для задач, где требуется найти группу собственных значений с наибольшими или, наоборот, наименьшими вещественными частями, предлагается комбинировать методы для разреженных матриц с техникой так называемого *чебышевского ускорения*. В результате получаются самокорректирующиеся вычислительные процессы. Если начальные значения выбраны неудачно, это будет обнаружено, и процесс повторится с новыми начальными данными, отражающими информацию, накопленную в проделанных вычислениях.

Так как большую роль в комбинированных методах играют чебышевские многочлены, нам придется вначале остановиться на их свойствах.

Многочлены Чебышева встречаются во многих задачах теории приближений и численного анализа. Одно из наиболее полезных и известных их применений связано с решением линейных систем с вещественными симметричными, положительно определенными матрицами или, более общо, с матрицами, имеющими вещественный спектр, расположенный по одну сторону (для определенности — справа) от нуля. Если для решения такой системы $Ax=b$ используется *трехслойная итерационная схема*

$$x^{k+1}=x^k-\alpha_{k+1}(Ax^k-b)-\beta_{k+1}(x^k-x^{k-1}), \quad k=0, 1, \ldots, \beta_1=0, \quad (19.1)$$

то вектор ошибки k-го шага $\varepsilon^k=x-x^k$ (x — точное решение системы) связан с начальной ошибкой ε^0 равенством

$$\varepsilon^k=\varphi_k(A)\,\varepsilon^0,$$

где $\varphi_k(t)$ — многочлен k-й степени со свободным членом 1, определяемый выбранными значениями итерационных параметров α_i, β_i $(1\leqslant i\leqslant k)$.

Считая для простоты матрицу A диагонализуемой, разложим ε^0 по ее собственным векторам:

$$\varepsilon^0=\xi_1 u_1+\xi_2 u_2+\ \ldots\ +\xi_n u_n. \quad (19.2)$$

Если $\lambda_1, \ldots, \lambda_n$ — соответствующие собственные значения, то

$$\varepsilon^k=\xi_1\varphi_k(\lambda_1)\,u_1+\ \ldots\ +\xi_n\varphi_k(\lambda_n)\,u_n. \quad (19.3)$$

Формула (19.3) показывает, что выгодно выбирать многочлены $\varphi_k(t)$ с возможно меньшими значениями в точках $\lambda_1, \ldots, \lambda_n$. Эти точки не известны, зато нередко известен отрезок $[m, M]$ $(m>0)$, которому все они принадлежат. В таком случае можно искать многочлен $\psi_k(t)$ из условия

$$\max_{m\leqslant t\leqslant M}|\psi_k(t)| = \min_{\substack{\deg\varphi(t)\leqslant k\\ \varphi(0)=1}} \max_{m\leqslant t\leqslant M}|\varphi(t)|. \tag{19.4}$$

Решение этой задачи дают нормированные многочлены Чебышева 1-го рода [189, 190]:

$$\psi_k(t)=\frac{T_k\left(\dfrac{M+m-2t}{M-m}\right)}{T_k(\theta)}, \qquad \theta=\frac{M+m}{M-m}. \tag{19.5}$$

Многочлены $T_k(t)$ обычно определяются соотношениями

$$\begin{aligned} &T_0(t)\equiv 1, \qquad T_1(t)=t,\\ &T_{k+1}(t)=2tT_k(t)-T_{k-1}(t), \qquad k=1, 2, \ldots \end{aligned} \tag{19.6}$$

Выбору (19.5) для многочлена $\varphi_k(t)$ соответствуют итерационные параметры

$$\alpha_{k+1}=4\delta_{k+1}/(M-m), \qquad \beta_{k+1}=-\delta_k\delta_{k+1}, \tag{19.7}$$

где

$$\delta_0=0, \quad \delta_1=1/\theta, \quad \delta_{k+1}=1/(2\theta-\delta_k), \qquad k=1, 2, \ldots \tag{19.8}$$

Наряду с (19.6) многочлены Чебышева на отрезке $[-1, 1]$ часто вводятся посредством формулы $T_k(t)=\cos(k\arccos t)$. Сейчас эти многочлены будут интересовать нас для комплексных значений, внешних по отношению к отрезку $[-1, 1]$. В этой области многочлены Чебышева определяются посредством формулы

$$T_k(z)=\operatorname{ch}(k\operatorname{arcch} z), \tag{19.9}$$

где обозначение переменной изменено с t на z.

Напомним, что функция $z=\operatorname{ch}\zeta$ периодическая с периодом $2\pi i$. Каждая полуполоса $\pi_m=\{\zeta\,|\,\zeta=u+iv, \ 0<u<\infty, \ 2m\pi\leqslant v<2(m+1)\pi\}$ отображается этой функцией взаимно однозначно на плоскость $z(=x+iy)$ с исключенным отрезком $[-1, 1]$. При этом образом полуинтервала $u=\text{const}$ является эллипс

$$x^2/\operatorname{ch}^2u+y^2/\operatorname{sh}^2u=1$$

с фокусами в точках $(-1, 0)$ и $(1, 0)$. Отрезку $u=0$, $2m\pi\leqslant v\leqslant 2(m+1)\pi$ отвечает дважды обходимый отрезок $-1\leqslant x\leqslant 1$, $y=0$.

Обратная функция $\zeta=\operatorname{Arcch} z$ многозначна; главной ее ветвью $\zeta=\operatorname{arcch} z$ будем считать функцию, определенную во внешности отрезка $[-1, 1]$, со значениями в π_0, так что образом эллипса

$$\frac{x^2}{a^2}+\frac{y^2}{a^2-1}=1, \qquad a>1, \tag{19.10}$$

является полуотрезок $u=\ln(a+\sqrt{a^2-1})$, $0\leqslant v<2\pi$.

Сложная функция $T_k(z)$ действует следующим образом: семейство эллипсов (19.10) отображается в семейство вертикальных полуинтервалов, лежащих в π_0; эти полуинтервалы растягиваются в k раз, а затем функцией $\operatorname{ch}\zeta$ снова переводятся в эллипсы того же семейства (19.10), но с полуосями $\operatorname{ch}(k \operatorname{arcch} a)$, $\operatorname{sh}(k \operatorname{arcch} a)$, где через $\operatorname{arcch} a$ обозначена функция $\ln(a+\sqrt{a^2-1})$ вещественной переменной $a>1$. При этом одному обходу эллипса (19.10) соответствует k-кратный обход эллипса-образа.

Установим еще, с какой скоростью растут значения многочленов $T_k(z)$ в фиксированной точке z. Положим $w=e^{\zeta}$, где $\zeta=\operatorname{arcch} z$. Так как $\operatorname{Re}\zeta>0$, то $|w|>1$. Из представления

$$T_k(z)=\operatorname{ch}(k\zeta)=(e^{k\zeta}+e^{-k\zeta})/2=(w^k+w^{-k})/2$$

выводим асимптотическое равенство

$$|T_k(z)|\cong\frac{1}{2}|w|^k.$$

Следовательно, модуль значения $T_k(z)$ возрастает приблизительно со скоростью геометрической прогрессии со знаменателем $|w|$. Для всех точек любого эллипса (19.10) скорость роста (асимптотическая) одинакова, поскольку одинаково значение $|w|=e^{\operatorname{Re}\zeta}=e^{\ln(a+\sqrt{a^2-1})}=$ $=a+\sqrt{a^2-1}$. Сравнение скоростей роста в разных точках z сводится, таким образом, к сравнению соответствующих величин $a+\sqrt{a^2-1}$.

Мы видим, что для многочленов Чебышева естественными областями рассмотрения являются эллипсы с фокусами в точках $(-1, 0)$ и $(1, 0)$. По отношению к эллипсам семейства $\mathscr{F}(d, c)$, т. е. эллипсам с центром в d и фокусами в точках $d+c$ и $d-c$, такими же естественными партнерами будут многочлены $T_k\left(\frac{d-z}{c}\right)$ от смещенного аргумента.

Пусть $F(d, c, a)$ — конкретный эллипс семейства $\mathscr{F}(d, c)$, характеризуемый значением $a>0$ большей полуоси. Поставим по отношению к $F(d, c, a)$ задачу о наилучшем равномерном приближении нуля, аналогичную (19.4): найти многочлен $\psi_k(z)$, такой, что

$$\max_{z\in F(d, c, a)}|\psi_k(z)|=\min_{\substack{\deg\varphi(z)=k\\ \varphi(0)=1}}\max_{z\in F(d, c, a)}|\varphi(z)|. \tag{19.11}$$

Сама постановка задачи предполагает, что точка $z=0$ будет внешней для эллипса $F(d, c, a)$.

Из теории приближений [122] известно, что задача (19.11) всегда имеет решение, причем единственное. Это верно не только для эллипсов, но вообще для любого замкнутого и ограниченного бесконечного подмножества $\mathscr{E}$ комплексной плоскости. Заметим, кстати, что решение задачи (19.11), поставленной на множестве $\mathscr{E}$, в теории приближений принято называть многочленом Чебышева для $\mathscr{E}$ (отличается, правда, условие нормировки — единице должен быть равен не младший, а старший коэффициент многочлена; см., например,

[15]). Подчеркнем, что это общее понятие лишь в частных случаях приводит к обычным многочленам Чебышева (19.9).

Вернемся к эллипсам $F(d, c, a)$ и перечислим известные для них результаты.

Теорема 19.1 (см. [201]). *Пусть* $c=0$, *т. е. эллипс* $F(d, c, a)$ *вырождается в окружность радиуса* a *с центром в* d. *Если точка* $z=0$ *лежит во внешности круга* $|z-d|\leqslant a$, *то решение задачи* (19.11) *выражается формулой*

$$\omega_k(z)=\left(\frac{d-z}{d}\right)^k.$$

Для случая, когда фокусы эллипса находятся на вещественной оси, справедлива

Теорема 19.2 (теорема Клейтона; см. [213]). *Пусть* $0<c\leqslant a\leqslant d$. *Тогда решением задачи* (19.11) *для эллипса* $F(d, c, a)$ *будет многочлен*

$$\omega_k(z)=\frac{T_k((d-z)/c)}{T_k(d/c)}. \tag{19.12}$$

Этот результат не может быть распространен на комплексные d и c [142]. Например, при $d>0$, $0<a<d$, $c=i\dfrac{d}{10}$ (т. е. большая ось эллипса параллельна мнимой оси)

$$\max_{z\in F(d, c, a)}|\omega_2(z)|<\max_{z\in F(d, c, a)}|\omega_3(z)|.$$

Однако асимптотически многочлен (19.12) и для комплексных d, c не уступает многочлену, наименее отклоняющемуся от нуля. Точный смысл этого высказывания выражает

Теорема 19.3 (см. [142]). *Пусть точка* $z=0$ *внешняя для эллипса* $F(d, c, a)$. *Положим*

$$M(\varphi_k)=\max_{z\in F(d, c, a)}|\varphi_k(z)|.$$

Тогда для многочлена ψ_k, *решающего задачу* (19.11), *и многочлена* (19.12) *имеет место равенство*

$$\lim_{k\to\infty}\sqrt[k]{M(\psi_k)}=\lim_{k\to\infty}\sqrt[k]{M(\omega_k)}.$$

Вернемся к вопросу о решении линейной системы $Ax=b$ методом (19.1). Пусть матрица A, вообще говоря, неэрмитова и ее спектр $\lambda_1, ..., \lambda_n$ расположен по одну сторону от мнимой оси. Если $F(d, c, a)$ — эллипс, заключающий в себе $\lambda_1, ..., \lambda_n$ и не пересекающийся с мнимой осью, то итерационные параметры трехслойной схемы можно выбрать так, чтобы они соответствовали семейству многочленов (19.12). В этом случае, согласно теореме 19.3, будет обеспечено оптимальное или почти оптимальное — на данном эллипсе — *подавление* вектора ошибки ε^k.

Множество из n точек $\lambda_1, ..., \lambda_n$, расположенных в одной комплексной полуплоскости, можно бесконечно многими способами по-

грузить в эллипс требуемого вида. Каждому эллипсу отвечает своя скорость подавления вектора ошибки. Как найти эллипс наискорейшего подавления?

Если при поиске оптимума ограничиться эллипсами с центром на вещественной прямой и осями, параллельными координатным осям, то решение этой задачи получено Мантейфелем [142], правда, в предположении, что точки $\lambda_1, ..., \lambda_n$ известны. Это решение изложено в п. 1 дополнений к § 19. В действительности нужны не все числа λ_i, а лишь те, что лежат на границе выпуклой оболочки множества $\{\lambda_1, ..., \lambda_n\}$. Однако в реальных условиях не известны и эти граничные собственные числа. Вместо них при решении линейных систем используют приближенные значения, которые можно получить посредством некоторой модификации степенного метода, применяемого к невязкам $r^k = b - Ax^k$; эти последние все равно вычисляются в схеме (19.1). Подробнее об этом можно прочесть в [143]. Мы же возвращаемся к спектральным задачам.

Упорядочим спектр матрицы A так, чтобы первые номера $1, ..., p$ получили искомые собственные значения. Положим $\mathscr{S} = \{\lambda_1, ..., \lambda_p\}$, $\mathscr{R} = \{\lambda_{p+1}, ..., \lambda_n\}$. Снова будем считать матрицу A диагонализуемой и рассмотрим задачу о вычислении вектора, лежащего в инвариантном подпространстве $\mathscr{L}^p$ только первых p собственных векторов. Если такой вектор взять в качестве начального вектора x^0 для любого из вариантов метода Ланцоша, то p шагов процесса дадут $\mathscr{L}^p$, и из соответствующей матрицы T_p (или H_p) определятся желаемые собственные значения. В реальных условиях мы, разумеется, не добьемся точного включения $x^0 \in \mathscr{L}^p$ и точного выделения инвариантного подпространства, но нечто близкое получить можно.

Предположим, что множество $\mathscr{R}$ можно заключить в эллипс $F(d, c, a)$, центр которого вещественен, оси параллельны координатным и внутри нет чисел из множества $\mathscr{S}$. Исходя из произвольно выбранного вектора x^0, построим последовательность $\{x^k\}$ по формулам

$$\sigma_1 = c/(\lambda_p - d), \quad x^1 = \frac{\sigma_1}{c}(A - dI)x^0, \quad \sigma_{k+1} = 1/(2/\sigma_1 - \sigma_k), \tag{19.13}$$

$$x^{k+1} = 2\frac{\sigma_{k+1}}{c}(A - dI)x^k - \sigma_k \sigma_{k+1} x^{k-1}, \quad k = 1, 2, ...$$

Через λ_p обозначено собственное значение из $\mathscr{S}$, ближайшее к $\mathscr{R}$; поскольку нумерация *внутри* $\mathscr{S}$ (как и внутри $\mathscr{R}$) произвольна, можно считать, что это собственное значение имеет номер p.

Процесс (19.13) есть спектральный эквивалент трехслойной схемы (19.1) с чебышевскими итерационными параметрами. Для векторов x^k справедливы представления

$$x^k = \omega_k(A)x^0,$$

где

$$\omega_k(z) = \frac{T_k((z-d)/c)}{T_k((\lambda_p - d)/c)}.$$

Таким образом, чебышевские многочлены теперь нормируются условием $\omega_k(\lambda_p) = 1$.

Разлагая x^0 по собственным векторам матрицы A:

$$x^0 = \xi_1 u_1 + \ldots + \xi_n u_n, \tag{19.14}$$

находим

$$x^k = \omega_k(\lambda_1)\xi_1 u_1 + \ldots + \omega_k(\lambda_{p-1})\xi_{p-1} u_{p-1} + \xi_p u_p + \\ + \omega_k(\lambda_{p+1})\xi_{p+1} u_{p+1} + \ldots + \omega(\lambda_n)\xi_n u_n. \tag{19.15}$$

Следовательно, в разложении вектора x^k компоненты по последним $n-p$ собственным направлениям будут убывать; а по первым $p-1$ направлениям, напротив, возрастать. После достаточного числа шагов чебышевского процесса (19.13) получим вектор, который можно считать расположенным в $\mathscr{L}^p$.

Для построения эллипса оптимального подавления в соответствии с процедурой Мантейфеля [142] нужны собственные значения матрицы A, т. е. как раз то, чего мы не имеем. Поэтому Саад предлагает объединить чебышевские итерации и процедуру Мантейфеля с некоторым спектральным алгоритмом для разреженных матриц, например с методом Арнольди. Если говорить для определенности о задаче вычисления p собственных значений с наибольшими вещественными частями, то комбинированный алгоритм задается следующим предписанием.

Пусть заданы нормированный вектор v_1 и натуральные числа m, l.

Алгоритм Саада.

1. Выполнить m шагов метода Арнольди, исходя из вектора v_1. Вычислить собственные значения полученной хессенберговой матрицы H_m. Выбрать p собственных значений $\theta_1, \ldots, \theta_p$ с наибольшими вещественными частями; положить $\mathscr{S} = \{\theta_1, \ldots, \theta_p\}$, $\mathscr{R} = \{\theta_{p+1}, \ldots, \theta_m\}$. Проверить условия сходимости для $\theta_1, \ldots, \theta_p$; если все они удовлетворены, закончить процесс; в противном случае перейти к шагу 2.

2. Обратиться к процедуре вычисления параметров d, c эллипса оптимального подавления для множества $\mathscr{R}$. Присвоить номер p собственному значению из $\mathscr{S}$, ближайшему к $\mathscr{R}$. Тем самым определены (с точностью до замены λ_p на θ_p) параметры чебышевских итераций (19.13). Вычислить начальный вектор x^0 чебышевского процесса как линейную комбинацию векторов Ритца $y_1, \ldots, y_p$. Перейти к шагу 3.

3. Выполнить l шагов чебышевского процесса, исходя из вектора x^0. Положить $v_1 = x^l / \|x^l\|_2$ и перейти к шагу 1.

Дадим небольшой комментарий к этому описанию. Прежде всего эллипс F, который строится на шаге 2, учитывает только «отвергаемые» числа Ритца $\theta_{p+1}, \ldots, \theta_m$. Так как число шагов m в одном прогоне метода Арнольди существенно меньше порядка n матрицы A, то многие собственные значения последней не будут отражены в спектре матрицы H_m. Ничто не мешает им находиться во внешности эллипса F, и в этом случае, хотя они и не представляют интереса для нас, соответствующие им компоненты в разложении (19.15) будут расти вместе с теми компонентами, которые мы усиливаем сознательно. Как избавиться от нежелательных собственных значений?

Ответ на этот вопрос такой: процесс избавляется от них сам. При определенном уровне усиления некоторые из нежелательных

собственных направлений будут настолько хорошо представлены в начальном векторе нового прогона метода Арнольди, что приближения к соответствующим собственным значениям попадут в спектр очередной хессенберговой матрицы. При новом применении процедуры Мантейфеля они окажутся внутри эллипса подавления, и на следующем этапе чебышевского процесса одноименные компоненты станут затухать.

Может случиться, что эллипс F будет вбирать в себя и некоторые из первых p чисел Ритца, так как построен он был, исходя из информации только о множестве $\mathscr{R}$. В этой ситуации может потребоваться изменение условия нормировки. Подробности можно найти в [175].

В отношении формирования вектора x^0 сохраняет силу замечание, высказанное в § 15: если $y_i = Q_m s_i$ $(i=1, ..., p)$ — векторы Ритца и $x^0 = \alpha_1 y_1 + \ldots + \alpha_p y_p$, то вычисляется x^0 по формуле

$$x^0 = Q_m(\alpha_1 s_1 + \ldots + \alpha_p s_p), \tag{19.16}$$

чем избегается построение векторов Ритца y_i, не сошедшихся с требуемой точностью.

Напомним, что сходимость собственных значений определяется в методе Арнольди по величинам невязок $\|(A-\theta_i I)y_i\|_2 = h_{m+1,m} \times \times |\{s_i\}_m|$, для вычисления которых нужны только собственные векторы s_i матрицы H_m (и даже только последние компоненты этих векторов). Саад рекомендует брать в качестве коэффициентов линейной комбинации (19.16) числа $\alpha_1, ..., \alpha_p$, пропорциональные длинам невязок. За счет этого в конечном векторе x^l чебышевского этапа будут хорошо представлены все желаемые собственные направления.

К сожалению, трудно сказать что-либо определенное о выборе параметров m, l комбинированной процедуры. Ограничимся поэтому изложением нескольких общих принципов.

Типичное значение для m должно в три-четыре раза превосходить число p вычисляемых собственных чисел. Относительно выбора l следует помнить, что слишком малое l может замедлить комбинированный процесс, не давая сказаться преимуществам чебышевского ускорения. Так, при $l=0$ получаем обычный итеративный метод Арнольди. С другой стороны, при чересчур большом l вектор x^l слишком сместится в направлении доминирующего собственного вектора и будет плохим начальным приближением для очередного прогона по методу Арнольди.

Метод Арнольди как часть составной процедуры был взят лишь в качестве конкретного примера. Вместо него можно использовать любой другой метод ланцошева типа или же метод одновременных итераций. В последнем случае степенные шаги, которые в совокупности — между двумя нормировками — описываются формулой

$$Q \leftarrow A^l Q,$$

заменяются чебышевским процессом

$$Q \leftarrow \omega_l(A) Q.$$

Матрица $\omega_l(A)Q$ строится по столбцам: к каждому столбцу матрицы Q по очереди применяют процесс (19.13). Так как число r итерируемых векторов превосходит число p определяемых собственных значений, то для построения эллипса подавления, как и в методе Арнольди, можно привлечь приближенные собственные значения $\theta_{p+1}, \ldots, \theta_r$; они вычисляются на шагах-поворотах.

Чебышевские итерации как средство ускорения степенного процесса использовались еще в программе Рутисхаузера [169]. Однако там определение области подавления, которой в симметричном случае является отрезок, практически не требовало дополнительных вычислений. Новым обстоятельством для неэрмитовых матриц оказывается необходимость специальной процедуры построения эллипса оптимального подавления.

В [175] обсуждаются результаты численных экспериментов с комбинированным алгоритмом. Приведем данные, относящиеся к одному из них. Собственные значения краевой задачи Дирихле для линейного дифференциального оператора

$$Lu=\frac{\partial}{\partial x}\left(a\frac{\partial u}{\partial x}\right)+\frac{\partial}{\partial y}\left(b\frac{\partial u}{\partial y}\right)+\frac{\partial(gu)}{\partial x}+f\frac{\partial u}{\partial x}, \quad 0\leqslant x, y\leqslant 1, \tag{19.17}$$

вычислялись путем дискретизации задачи $Lu=\lambda u$ по обычной пятиточечной разностной схеме с постоянным шагом h. Коэффициентами оператора (19.17) в описываемом эксперименте были функции

$$a(x, y)=e^{-xy}, \quad b(x, y)=e^{xy},$$

$$g(x, y)=\gamma(x+y), \quad f(x, y)=1/(1+x+y).$$

В результате дискретизации получается спектральная задача для большой разреженной матрицы. Посредством параметра γ можно управлять степенью ее несимметричности. В эксперименте были взяты значения $\gamma=20$, $h=1/31$, что приводит к сильно несимметричной матрице порядка 900. Вычислялись четыре самых правых собственных значения.

Для комбинированного алгоритма Арнольди — Чебышева были заданы значения параметров $m=15$, $l=80$. Процесс развивался следующим образом: 15 шагов метода Арнольди; вычисление параметров эллипса; 80 шагов чебышевского метода; наконец, еще 15 шагов метода Арнольди. После этого критерий сходимости был выполнен. Самыми правыми оказались две пары сопряженных комплексных чисел; их вычисленным приближениям соответствовали невязки с длинами 5.4×10^{-13} и 8.4×10^{-8}. Общее число матрично-векторных умножений в процессе составило 110.

Для сравнения та же спектральная задача была решена (при том же условии выхода) методом Стьюарта с $r=8$ и $l=50$. Потребовалось 220 итераций метода, в которых 1708 раз вычислялось произведение матрицы на вектор. Время счета возросло примерно в 8.5 раза. Комбинированный алгоритм Стьюарта — Чебышева для этой же задачи при $l=15$ выполнил 104 итерации с общим числом матрично-векторных умножений, равным 928.

Эксперименты, проведенные Саадом, показали, что чебышевские итерации могут заметно ускорять алгоритмы для разреженных матриц. Однако ускорение наблюдалось не во всех экспериментах. В некоторых случаях комбинированные процессы испытывали трудности при поиске хорошего эллипса, что можно было заключить из хаотического поведения параметров d, c эллипсов, вычисляемых в разных прогонах. Форма спектра задачи может быть такой, что он сопротивляется локализации его «лишней части» в эллипсе с приемлемой скоростью подавления.

ДОПОЛНЕНИЯ К § 19

1. Опишем алгоритм Мантейфеля для построения эллипса оптимального подавления. Предполагается, что заданы числа $\lambda_1, \ldots, \lambda_n$, принадлежащие полуплоскости $\operatorname{Re}\lambda > 0$. В каждом семействе эллипсов с фиксированными фокусами мы выбираем наименьший эллипс, содержащий — в себе или в своей внутренности — заданные точки. Следовательно, хотя бы одна из них удовлетворяет уравнению наименьшего эллипса, что при известных d и c однозначно определяет a.

Чтобы сформулировать оптимизационную задачу, нужно определить функцию качества. Сопоставим наименьшему эллипсу семейства $\mathscr{F}(d, c)$ число

$$\varphi(d, c^2) = \max_{1 \leqslant i \leqslant n} r(\lambda_i), \tag{19.18}$$

где

$$r(\lambda) \equiv r(\lambda, d, c^2) = \left| \frac{(d-\lambda) + \sqrt{(d-\lambda)^2 - c^2}}{d + \sqrt{d^2 - c^2}} \right|. \tag{19.19}$$

Пусть $T_k\left(\dfrac{d-z}{c}\right)$ — «смещенный» чебышевский многочлен. Числитель дроби (19.19) характеризует (см. обсуждение свойств чебышевских многочленов в основном тексте параграфа) скорость роста значения T_k в точке λ, а знаменатель — значения T_k в точке $z=0$. После нормировки в нуле, т. е. перехода от $T_k\left(\dfrac{d-z}{c}\right)$ к многочлену $\omega_k(z)$, дробь (19.19) выражает скорость убывания компоненты вектора ошибки, отвечающей собственному значению λ, в чебышевском процессе с параметрами d, c. Функция (19.18) таким же образом выражает скорость убывания всего вектора ошибки.

Напомним, что к рассмотрению допускаются только эллипсы с центром на вещественной прямой и осями, параллельными координатным. В итоге приходим к вещественной экстремальной задаче: найти

$$\min_{d, c} \varphi(d, c^2) = \min_{d, c} \max_{i} r(\lambda_i). \tag{19.20}$$

В то время как d вещественно, число c может быть и чисто мнимым. Однако функция φ зависит лишь от всегда вещественного значения c^2.

В работах Мантейфеля [142, 143] исследуется задача решения вещественной несимметричной линейной системы $Ax = b$ и в качестве $\lambda_1, \ldots, \lambda_n$ берется множество собственных значений матрицы A. Оно симметрично относительно вещественной оси, и это свойство используется в процедуре построения оптимального эллипса.

Легко показать, что при решении задачи (19.20) можно ограничиться значениями d и c^2, для которых $d > 0$, $c^2 < d^2$. При этом внутренний максимум достаточно искать среди чисел λ_i, образующих вершины выпуклой оболочки

множества $\{\lambda_1, ..., \lambda_n\}$, а ввиду симметрии — только среди вершин, лежащих выше вещественной оси. Множество таких вершин обозначим через H^+.

Для искомой точки минимума (d_0, c_0^2) должен осуществляться один из следующих трех вариантов (т. е. имеет место *трилемма* [56]):

1. Точка (d_0, c_0^2) реализует локальный минимум функции $r(\lambda_i)$ для некоторого значения i.

2. Для некоторых значений i, j точка (d_0, c_0^2) реализует локальный минимум вдоль линии

$$r(\lambda_i)=r(\lambda_j).$$

3. В точке (d_0, c_0^2) для некоторых значений i, j, k выполняется равенство

$$r(\lambda_i)=r(\lambda_j)=r(\lambda_k).$$

Для двух простых частных случаев задача (19.20) решается до конца:

А. Множество H^+ состоит из единственной точки $\lambda_1=x_1+iy_1$, $y_1>0$. Другими словами, все множество $\{\lambda_1, ..., \lambda_n\}$ расположено на вертикальном отрезке $[\bar{\lambda}_1, \lambda_1]$. Единственный локальный минимум функции $r(\lambda_1)$ достигается при $d_0=x_1$, $c_0^2=-y_1^2$. Последнее равенство означает, что c_0 — чисто мнимое число, и оптимальный эллипс имеет фокусы в точках λ_1 и $\bar{\lambda}_1$. В то же время λ_1 и $\bar{\lambda}_1$ должны лежать на границе оптимального эллипса, откуда следует, что $a=0$, т. е. эллипс вырождается в отрезок $[\bar{\lambda}_1, \lambda_1]$. Минимальное значение функции $\varphi(d, c^2)$, которая в данном случае совпадает с $r(\lambda_1)$, равно

$$\varphi(x_1, -y_1^2)=r(\lambda_1, x_1, -y_1^2)=\frac{y_1}{x_1+\sqrt{x_1^2+y_1^2}}.$$

Б. Множество H^+ содержит две точки: $\lambda_1=x_1+iy_1$, $\lambda_2=x_2+iy_2$. Можно показать, что

$$r(\lambda_1, x_1, -y_1^2)<r(\lambda_2, x_1, -y_1^2), \qquad r(\lambda_1, x_2, -y_2^2)>r(\lambda_2, x_2, -y_2^2);$$

это значит, что точки локального минимума функций $r(\lambda_1)$, $r(\lambda_2)$ не являются точками минимума функции

$$\varphi(d, c^2)=\max\{r(\lambda_1), r(\lambda_2)\}.$$

Поэтому должен выполняться второй вариант трилеммы:

$$r(\lambda_1, d_0, c_0^2)=r(\lambda_2, d_0, c_0^2).$$

Положим

$$A=\frac{x_2-x_1}{2}, \quad B=\frac{x_1+x_2}{2}, \quad S=\frac{y_2-y_1}{2}, \quad T=\frac{y_1+y_2}{2}.$$

Для определенности считаем, что $x_2>x_1$. Тогда $A>0$, $B>0$, $T\geqslant 0$.

Если $S=0$, то параметры d_0 и c_0^2 оптимального эллипса выражаются формулами

$$d_0=B, \qquad c_0^2=\frac{a_0^2(a_0^2-(A^2+T^2))}{a_0^2-A^2}.$$

Через a_0 обозначена длина большей полуоси. Она определяется из кубического уравнения

$$q_1y^3+q_2y^2+q_3y+q_4=0, \tag{19.21}$$

в котором $q_1=B^2+T^2$, $q_2=-3A^2B^2$, $q_3=3A^4B^2$, $q_4=-A^4B^2(A^2+T^2)$. Число $y_0=a_0^2$ является единственным корнем уравнения (19.21), расположенным в интервале $[A^2, B^2)$.

Пусть теперь $S \neq 0$. Тогда параметры c_0^2 и a_0^2 оптимального эллипса можно выразить как функции от d_0:

$$c_0^2 = \frac{(d_0-(B+ST/A))(d_0-(B-AT/S))(d_0-(B-AS/T))}{d_0-B},$$

$$a_0^2 = (d_0-(B-AT/S))(d_0-(B-AS/T)).$$

Полагая $z=d-B$, находим, что z_0, отвечающее искомому параметру d_0, есть корень уравнения

$$p_1 z^5 + p_2 z^4 + p_3 z^3 + p_4 z^2 + p_5 z + p_6 = 0, \tag{19.22}$$

лежащий в интервале $(0, A)$, если $S>0$, или в интервале $(-A, 0)$, если $S<0$. Коэффициенты уравнения (19.22) равны:

$$p_1 = \left(2B - A\left(\frac{T}{S}+\frac{S}{T}\right)\right)\left(2B+\frac{ST}{A}-A\left(\frac{T}{S}+\frac{S}{T}\right)\right),$$

$$p_2 = \left(2B+\frac{ST}{A}-A\left(\frac{T}{S}+\frac{S}{T}\right)\right)\left((2AB+ST)\left(\frac{T}{S}+\frac{S}{T}\right)+4A^2\right)+$$
$$+B^2\left(2B-A\left(\frac{T}{S}+\frac{S}{T}\right)\right)+B(B^2-A^2),$$

$$p_3 = 4A^4 - 4A^3B\left(\frac{T}{S}+\frac{S}{T}\right)+A^2ST\left(\left(\frac{T^3}{S^3}+\frac{S^3}{T^3}\right)-3\left(\frac{T}{S}+\frac{S}{T}\right)\right)+A^2B^2\left(\frac{T^2}{S^2}+\frac{S^2}{T^2}+3\right),$$

$$p_4 = AST\left(\left(B-A\frac{T}{S}\right)\left(B-3A\frac{T}{S}\right)+\left(B-A\frac{S}{T}\right)\left(B-3A\frac{S}{T}\right)\right),$$

$$p_5 = -3A^3ST\left(2B-A\left(\frac{T}{S}+\frac{S}{T}\right)\right), \quad p_6 = -3A^3ST(B^2-A^2).$$

Эллипс с параметрами d_0, c_0^2, a_0 будем называть (λ_1, λ_2)-наилучшим эллипсом.

В общем случае миожество H^+ содержит три и более вершин. Поэтому в принципе могут осуществляться все три варианта трилеммы. Второй вариант имеет место, если для некоторой пары вершин λ_i, λ_j из H^+ все остальные числа λ_l содержатся в (λ_i, λ_j)-наилучшем эллипсе. В третьем варианте на границе оптимального эллипса должны находиться (по меньшей мере) три различные точки λ_l. Обозначим их через λ_1, λ_2, λ_3: $\lambda_1 = x_1 + iy_1$, $\lambda_2 = x_2 + iy_2$, $\lambda_3 = x_3 + iy_3$; можно считать, что $x_1 < x_2 < x_3$. Оптимальный эллипс, проходящий через точки λ_1, λ_2, λ_3, однозначно определяется формулами

$$d_0 = \frac{1}{2}\frac{(y_1^2(x_2^2-x_3^2)+y_2^2(x_3^2-x_1^2)+y_3^2(x_1^2-x_2^2))}{(y_1^2(x_2-x_3)+y_2^2(x_3-x_1)+y_3^2(x_1-x_2))},$$

$$a_0^2 = d_0^2 - \frac{(y_1^2x_2x_3(x_2-x_3)+y_2^2x_1x_3(x_3-x_1)+y_3^2x_1x_2(x_1-x_2))}{(y_1^2(x_2-x_3)+y_2^2(x_3-x_1)+y_3^2(x_1-x_2))},$$

$$c_0^2 = a_0^2\left(1-\frac{(y_1^2(x_2-x_3)+y_2^2(x_3-x_1)+y_3^2(x_1-x_2))}{(x_1-x_2)(x_2-x_3)(x_3-x_1)}\right).$$

Этот эллипс будем называть $(\lambda_1, \lambda_2, \lambda_3)$-наилучшим эллипсом.

Проверка того, какие именно три точки из H^+ лежат на границе оптимального эллипса, облегчается необходимым условием:

$$(x_2-x_1)(y_3^2-y_1^2) < (x_3-x_1)(y_2^2-y_1^2).$$

В целом предлагаемый Мантейфелем алгоритм решения задачи (19.20) состоит в следующем:

1) Если H^+ содержит лишь одну вершину $\lambda_1 = x_1 + iy_1$, то оптимальный эллипс — это отрезок $[\bar{\lambda}_1, \lambda_1]$.

2) Если в H^+ две и более вершин, то для каждой пары λ_i, λ_j строится (λ_i, λ_j)-наилучший эллипс. Если какой-то из этих эллипсов содержит в себе все остальные вершины из H^+, то он и будет оптимальным.

3) Если никакой из парных эллипсов не охватывает все H^+, то строятся $(\lambda_i, \lambda_j, \lambda_k)$-наилучшие эллипсы. Среди них отбираются те, что содержат в себе H^+. Оптимальным будет тот из отобранных эллипсов, которому отвечает наименьшее значение функции φ.

2. Алгоритм Мантейфеля использует процедуру, выделяющую в множестве $\mathscr{R} = \{\lambda_1, ..., \lambda_n\}$ крайние точки его выпуклой оболочки $\operatorname{conv}\mathscr{R}$. Для решения этой последней задачи может быть применен ряд методов [37]. Опишем один из простейших — алгоритм Грэхема, ориентированный на работу с плоскими множествами. Алгоритм состоит из пяти этапов.

Этап 1. Его цель — найти точку P, принадлежащую внутренности множества $\operatorname{conv}\mathscr{R}$. Для этого берут три произвольные точки из $\mathscr{R}$ и проверяют на коллинеарность (т. е. на принадлежность одной прямой). Если они коллинеарны, средняя отбрасывается и заменяется еще не проверявшейся точкой из $\mathscr{R}$, снова проводится тест на коллинеарность и т. д. В конечном счете либо окажется, что все множество $\mathscr{R}$ лежит на одной прямой — в этом случае $\operatorname{conv}\mathscr{R}$ есть отрезок с концами в двух последних оставшихся точках,— либо будут найдены три неколлинеарные точки x, y, z из $\mathscr{R}$. Тогда в качестве P можно взять центр тяжести треугольника с вершинами x, y, z.

Трудоемкость этапа 1 равна $O(n)$ операций.

Этап 2. Вводим полярную систему координат с полюсом в точке P, найденной на первом этапе. Полярная ось может быть выбрана произвольно. Вычисляем координаты каждой точки из $\mathscr{R}$ в новой системе. Заметим, что по сравнению с этапом 1 число $\tilde{n}$ точек множества $\mathscr{R}$ могло уменьшиться за счет отбрасывания средних точек при проверках коллинеарности. Понятно, что трудоемкость этапа 2 не превосходит $O(n)$ операций.

Этап 3. Элементы множества $\mathscr{R}$ можно трактовать сейчас как пары (ρ_k, φ_k). Упорядочим их так, чтобы $0 \leqslant \varphi_1 \leqslant \varphi_2 \leqslant ... \leqslant \varphi_{\tilde{n}} < 2\pi$. Напомним, что наихудший алгоритм сортировки требует $O(\tilde{n}^2)$ операций.

Этап 4. Здесь проводится новое «прореживание» множества $\mathscr{R}$. Если $\varphi_i = \varphi_{i+1}$ и $\rho_i < \rho_{i+1}$, то пару (ρ_i, φ_i) можно отбросить, поскольку она не может соответствовать крайней точке выпуклой оболочки. При $\rho_i > \rho_{i+1}$ отбрасывается пара $(\rho_{i+1}, \varphi_{i+1})$. Если $\rho_i = 0$ (т. е. точка P принадлежит $\mathscr{R}$), то пара (ρ_i, φ_i) также отбрасывается. Новое число элементов в множестве $\mathscr{R}$ обозначим через n'. Трудоемкость этапа 4 — $O(\tilde{n})$ сравнений.

Этап 5. На каждом шаге этого этапа рассматриваются три последовательные точки: (ρ_k, φ_k), $(\rho_{k+1}, \varphi_{k+1})$, $(\rho_{k+2}, \varphi_{k+2})$, $\varphi_k < \varphi_{k+1} < \varphi_{k+2}$. Сохраняем обозначения координат, несмотря на изменения нумерации вследствие уже произведенных и предстоящих прореживаний. Имеются две возможности:

1. Точки P и $(\rho_{k+1}, \varphi_{k+1})$ находятся по одну сторону от отрезка с концами в k-й и $(k+2)$-й точках. В этом случае точка $(\rho_{k+1}, \varphi_{k+1})$ заведомо не может быть крайней для выпуклой оболочки, а потому удаляется из $\mathscr{R}$. Сдвинув нумерацию: $k+1 \to k+2$, $k+2 \to k+3$, ..., переходим к новому шагу.

2. Точки P и $(\rho_{k+1}, \varphi_{k+1})$ находятся по разные стороны от отрезка с концами в k-й и $(k+2)$-й точках. Переходим к аналогичной проверке для следующей тройки: $(\rho_{k+1}, \varphi_{k+1})$, $(\rho_{k+2}, \varphi_{k+2})$, $(\rho_{k+3}, \varphi_{k+3})$.

Трудоемкость этапа 5 составляет $O(n')$ операций, а общая трудоемкость алгоритма — не более чем $O(n^2)$ операций.

СПИСОК ЛИТЕРАТУРЫ

1. Апокорина В. С., Лебедев В. И. О применении метода обратных итераций в предельных случаях // Вычислительные методы линейной алгебры.— Новосибирск, 1973.— С. 23—33.
2. Бате К., Вилсон Е. Л. Численные методы анализа и метод конечных элементов.— М.: Стройиздат, 1982.
3. Бахвалов Н. С. Численные методы.— М.: Наука, 1973.
4. Беклемишев Д. В. Дополнительные главы линейной алгебры.— М.: Наука, 1983.
5. Белицкий Г. Р., Любич Ю. И. Нормы матриц и их приложения.— Киев: Наук. думка, 1984.
6. Березин И. С., Жидков Н. П. Методы вычислений. Т. 1.— М.: Наука, 1966.
7. Воеводин В. В. Вычислительные основы линейной алгебры.— М.: Наука, 1977.
8. Воеводин В. В. Линейная алгебра.— М.: Наука, 1980.
9. Воеводин В. В., Ким Г. Д., Агафонова З. И. О сходимости QR-алгоритма со сдвигом // Численный анализ на Фортране. Вып. 14.— М.: Изд-во МГУ, 1976.— С. 5—13.
10. Воеводин В. В., Кузнецов Ю. А. Матрицы и вычисления.— М.: Наука, 1984.
11. Воеводин В. В., Тыртышников Е. Е. Вычислительные процессы с теплицевыми матрицами.— М.: Наука, 1987.
12. Гантмахер Ф. Р. Теория матриц.— М.: Наука, 1967.
13. Гершгорин С. А. Über die Abgrenzung der Eigenwerte einer Matrix // ИАН. СССР. Сер. физ.-мат.— 1931.— С. 749—754.
14. Глазман И. М., Любич Ю. И. Конечномерный линейный анализ.— М.: Наука, 1969.
15. Голузин Г. М. Геометрическая теория функций комплексного переменного.— М.: Наука, 1966.
16. Гохберг И. Ц., Крейн М. Г. Основные положения о дефектных числах, корневых числах и индексах линейных операторов // УМН.— 1957.— Т. 12, № 2.— С. 43—118.
17. Друскин В. Л., Книжнерман Л. А. Использование операторных рядов по ортогональным многочленам при вычислении функций от самосопряженных операторов и обоснование феномена Ланцоша.— М., 1987.— 47 с.— Деп. ВИНИТИ 02.03.87, № 1535.
18. Икрамов Х. Д. Разреженные матрицы // Математический анализ. Итоги науки и техники.— М.: ВИНИТИ, 1982.
19. Икрамов Х. Д. Численное решение матричных уравнений.— М.: Наука, 1984.
20. Икрамов Х. Д. Численное решение матричных уравнений и симплектическая матричная алгебра // Вычислительные процессы и системы. Вып. 5.— М.: Наука, 1987.— С. 154—163.
21. Икрамов Х. Д. О неразложимых гамильтоновых матрицах с чисто мнимым спектром // ЖВМиМФ.— 1988.— Т. 28, № 12.— С. 1897—1902.

22. Икрамов Х. Д. О двух классах матриц, допускающих быстрое вычисление собственных значений // Вестн. МГУ. Сер. 15; Вычисл. математика и кибернетика — 1989. — С. 18—24.
23. Икрамов Х. Д., Сагитов М. С. О факторизации симплектических матриц в произведение элементарных // Методы и алгоритмы численного анализа. — М.: Изд-во МГУ, 1987. — С. 111—119.
24. Като Т. Теория возмущений линейных операторов. — М.: Мир, 1972.
25. Князев А. В. Вычисление собственных значений и векторов в сеточных задачах: алгоритмы и оценки погрешности. — М.: ОВМ АН СССР, 1986.
26. Кублановская В. Н. О некоторых алгоритмах для решения полной проблемы собственных значений // ЖВМиМФ. — 1961. — Т. I, № 4.
27. Курош А. Г. Курс высшей алгебры. — М.: Физматгиз, 1963.
28. Ланг В. Эффективные алгоритмы решения задач на собственные значения // ЖВМ и МФ. — 1985. — Т. 25, № 9. — С. 1409—1413.
29. Ланкастер П. Теория матриц. — М.: Наука, 1978.
30. Лоусон Ч., Хенсон Р. Численное решение задач метода наименьших квадратов. — М.: Наука, 1986.
31. Мальцев А. И. Основы линейной алгебры. — М.: Наука, 1970.
32. Маркус М., Минк Х. Обзор по теории матриц и матричных неравенств. — М.: Наука, 1972.
33. Марчук Г. И., Лебедев В. И. Численные методы в теории переноса нейтронов. — М.: Наука, 1981.
34. Нефедов Ю. В., Баранов В. И., Грибов Л. А. Программы диагонализации больших матриц специальной структуры методом Ланцоша. — М., 1985. — 45 с. — Деп. ВИНИТИ 22.10.85, № 7362.
35. Парлетт Б. Симметричная проблема собственных значений. — М.: Мир, 1983.
36. Пономарева И. А. Метод Ланцоша с выборочной ортогонализацией // Численный анализ: методы и алгоритмы. — М.: Изд-во МГУ, 1986.
37. Препарата Ф., Шамос М. Вычислительная геометрия. Введение. — М.: Мир, 1988.
38. Пупков В. А. Об изолированном собственном значении матрицы и структуре его собственного вектора // ЖВМиМФ. — 1983. — Т. 23, № 6.
39. Собянин А. В. Анализ ошибок округления и устойчивость в методах типа Ланцоша. — М.: ОВМ АН СССР, 1985.
40. Собянин А. В. Решение проблемы собственных значений для симметричных матриц большого порядка // Вычислительные процессы и системы. Вып. 5. — М.: Наука, 1987. — С. 174—179.
41. Соловьев В. Н. Обобщение теоремы Гершгорина // Изв. АН СССР. Сер. мат. — 1983. — Т. 47, № 6. — С. 1285—1302.
42. Уилкинсон Дж. Х. Алгебраическая проблема собственных значений. — М.: Наука, 1970.
43. Уилкинсон Дж., Райнш К. Справочник алгоритмов на языке АЛГОЛ. Линейная алгебра. — М.: Машиностроение, 1976.
44. Уонэм М. Линейные многомерные системы управления. — М.: Наука, 1980.
45. Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры. — М.; Л.: Физматгиз, 1963.
46. Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры. Зап. научн. семин. — Л.: ЛОМИ АН СССР, 1975.
47. Файбусович Е. Л. Обобщенные потоки Тода, уравнения Риккати на грассманиане и QR-алгоритм // Функц. анализ и его прилож. — 1987. — Т. 21, № 2.
48. Форсайт Дж., Малькольм М., Моулер К. Машинные методы математических вычислений. — М.: Мир, 1980.
49. Хорн Р. А., Джонсон Ч. Матричный анализ. — М.: Мир, 1989.
50. Aasland L., Bjørstad P. The generalized eigenvalue problem in shipdesign and offshore industry—a comparison of traditional methods with the Lanczos process // Lecture Notes Math. — 1983. — V. 973. — P. 146—155.
51. Ammar G., Martin C. The geometry of matrix eigenvalue methods // Acta appl. math. — 1986. — 5, № 3. — P. 239—278.
52. Andrew A. L. Eigenvectors of certain matrices // Linear Algebra Appl. — 1973. — 7. — P. 151—162.

53. Andrew A. L. Further comments on «On the eigenvectors of symmetric Toeplitz matrices» // IEEE Trans. Acoust. Speech Signal Process.— 1985.— 33, № 4.— P. 1013.
54. Arnoldi W. E. The principle of minimized iterations in the solution of the matrix eigenvalue problem // Quart. Appl. Math.— 1951.— 9, № 1.
55. Bartels R. H., Stewart G. W. Solution of the matrix equation $AX+XB=C$ // Commun. ACM.— 1971.— 15.— P. 820—826.
56. Bartle R. G. Elements of real analysis.— N. Y.: Wiley, 1964.
57. Bauer F. L. Das Verfahren der Treppeniteration und verwandte Verfahren zur Lösung algebraischer Eigenwertprobleme // ZAMM.— 1957.— 8.
58. Bauer F. L. On modern matrix iteration processes of Bernoulli and Graeffe type // J. ACM.— 1958.— 5.— P. 246—257.
59. Bauer F. L. Fields of values and Gershgorin discs // Numer. Math.— 1968.— 12.— P. 91—95.
60. Bauer F. L., Fike C. T. Norms and exclusion theorems // Numer. Math.— 1960.— 2.— P. 137—141.
61. Bhatia R., Friedland S. Variation of Grassmann powers and spectra // Linear Algebra Appl.— 1981.— 40.— P. 1—18.
62. Björck Å. Solving linear least squares problems by Gram-Schmidt orthogonalization // BIT.— 7.— P. 1—21.
63. Björck Å., Hammarling S. A Schur method for the square root of a matrix // Linear Algebra Appl.— 1983.— 52/53.— P. 127—140.
64. Borwein J. M., Richmond B. How many matrices have roots? // Can. J. Math.— 1984.— XXXVI, № 2.— P. 286—299.
65. Brauer A. Limits for the characteristic roots of a matrix // Duke Math. J.— I: 1946.— 13.— P. 387—395; II: 1947.— 14.— P. 21—26; III: 1952.— 19.
66. Brualdi R. Matrices, eigenvalues and directed graphs // Lin. Multilin. Algebra.— 1982.— 11.— P. 143—165.
67. Bunse W., Bunse-Gerstner A. Numerische lineare Algebra.— Stuttgart: B. G. Teubner, 1985.
68. Bunse-Gerstner A., Byers R., Mehrmann V. A quaternion QR algorithm // Numer. Math.— 1989.— V. 55, № 1.— C. 83—95.
69. Bunse-Gerstner A., Mehrmann V. A symplectic QR like algorithm for the solution of the real algebraic Riccati equation // IEEE Trans. Automat. Contr.— 1986.— 31, № 12.— P. 1104—1113.
70. Buurema H. J. A geometric proof of convergence for the QR method.— Groningen, Math. inst., Rep. TW-62, 1968.
71. Buurema H. J. A geometric proof of convergence for the QR method.— Thesis, Rijksuniversiteit Groningen, 1970, 51 pp.
72. Businger P. Reducing a matrix to Hessenberg form // Math. Comput.— 1969.— 23.— P. 819—821.
73. Byers R. A Hamiltonian QR algorithm / SIAM J. Sci. Stat. Comput. 1986.— 7, № 1.— P. 212—229.
74. Byers R. Solving the algebraic Riccati equation with the matrix sign function // Linear Algebra Appl.— 1987.— 85.— P. 267—279.
75. Cantoni A., Butler P. Eigenvalues and eigenvectors of symmetric centrosymmetric matrices // Linear Algebra Appl.— 1976.— 13.— P. 275—288.
76. Carnoy E. G., Geradin M. On the practical use of the Lanczos algorithm in finite element applications to vibration and bifurcation problems // Lecture Notes Math.— 1983.— 873.— P. 156—176.
77. Chan S. P., Feldman R., Parlett B. N. Algorithm 517. A program for computing the condition numbers of matrix eigenvalues without computing eigenvectors [F2] // ACM Trans. Math. Software.— 1977.— 3, № 2.— P. 186—203.
78. Chatelin F. Simultaneous Newton's iteration for the eigenproblem // Computing, Suppl. 5.— 1984.— P. 67—74.
79. Chen Nai-fu. Inverse iteration on defective matrices // Math. Comput.— 1977.— 31, № 139.— P. 726—732.
80. Chu M. T. On the global convergence of the Toda lattice for real normal matrices and its applications to the eigenvalue problem // SIAM J. Math. Anal.— 1984.— 15, № 1.— P. 98—104.

81. Chu M. T. The generalized Toda flow, the QR algorithm and the center manifold theory // SIAM J. Algebr. Discrete Math.— 1984.— 5, № 2.
82. Chu M. T. Asymptotic analysis of Toda lattice on diagonalizable matrices // Nonlinear Anal., Theory, Math. and Appl.— 1985.— 9, № 2.
83. Clint M., Jennings A. A simultaneous iteration method for the unsymmetric eigenvalue problem // J. Inst. Math. and Appl.— 1971.— 8, № 1.— P. 111—121.
84. Colpa J. H. P. Diagonalization of the quadratic boson Hamiltonian with zero modes I. Mathematical // Physica.— 1986.— A134, № 2.— P. 377—419.
85. Crandall S. H. Iterative procedures related to relaxation methods for eigenvalue problems // Proc. Roy. Soc. London, Ser. A.— 1951.— 207.
86. Cullum J. K., Willoughby R. A. A Lanczos procedure for the modal analysis of very large nonsymmetric matrices.— Proc. 23rd IEEE Conf. Dec. and Contr., Las Vegas, Nev. 1984. N. Y.: IEEE, 1984.— P. 1758—1761.
87. Cullum J. K., Willoughby R. A. Lanczos algorithms for large symmetric eigenvalue computations. Vol. 1. Theory. Vol. 2. Programs.— Basel: Birkhauser, 1985.
88. Cullum J. K., Willoughby R. A. A practical procedure for computing eigenvalues of large sparse nonsymmetric matrices. In: Large scale eigenvalue problems.— Amsterdam: North-Holland, 1986, p. 193—240.
89. Davis C., Kahan W. M. The rotation of eigenvectors by a perturbation. III // SIAM J. Numer. Anal.— 1970.— 7.— P. 1—46.
90. Dax A., Kaniel S. The ELR method for computing the eigenvalues of a general matrix // SIAM J. Numer. Anal.— 1981.— 18, № 4.— P. 597—605.
91. Deift P., Nanda T., Tomei C. Ordinary differential equations and the symmetric eigenvalue problem // SIAM J. Numer. Anal.— 1983.— 20.
92. Della Dora J. Numerical linear algorithms and group theory // Linear Algebra Appl.— 1975.— 10, № 3.— P. 267—283.
93. Deutsch E., Zenger C. On Bauer's generalized Gershgorin discs // Numer. Math.— 1975.— 24.— P. 63—70.
94. Dongarra J. J., Gabriel J. R., Koelling D. D., Wilkinson J. H. The eigenvalue problem for Hermitian matrices with time reversal symmetry // Linear Algebra Appl.— 1984.— 60.— P. 27—42.
95. Duff I. S., Reid J. K. On the reduction of sparse matrices to condensed form by similarity transformation // J. Inst. Math. and Appl.— 1975.— 15, № 2.— P. 217—224.
96. Eberlein P. J. On measures of non-normality for matrices // Amer. Math. Monthly.— 1965.— 72.— P. 995—996.
97. Eberlein P. J., Huang C. P. Global convergence of the QR algorithm for unitary matrices with some results for normal matrices // SIAM J. Numer. Anal.— 1975.— 12, № 1.— P. 97—104.
98. Elman H. C., Saad Y., Saylor P. E. A hybrid Chebyshev Krylov subspace algorithm for solving nonsymmetric systems of linear equations // SIAM J. Sci. Stat. Comput.— 1986.— 7, № 3.— P. 840—855.
99. Elsner L. On some algebraic problems in connection with general eigenvalue algorithms // Linear Algebra Appl.— 1979.— 26.— P. 123—138.
100. Elsner L. Neuere Verfahren zur Bestimmung der Eigenwerte von Matrizen. Proc. Numer. Math. Symp., Hamburg, 1979.— Basel: Birkhauser, 1979.
101. Elsner L. On the variation of the spectra of matrices // Linear Algebra Appl.— 1982.— 47.— P. 127—138.
102. Elsner L. An optimal bound for the spectral variation of two matrices // Linear Algebra Appl.— 1985.— 71.— P. 77—80.
103. Elsner L. Matrix decompositions, symmetries and eigenvalue algorithms. In: Current trends in matrix theory.— Amsterdam: North-Holland, 1987.
104. Elsner L., Paardekooper M. H. C. On measures of nonnormality of matrices // Linear Algebra Appl.— 1987.— 92.— P. 107—124.
105. Ericsson T., Ruhe A. The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems // Math. Comput.— 1980.— 35, № 152.— P. 1251—1268.
106. Feingold D. G., Varga R. S. Block diagonally dominant matrices and generalizations of the Gershgorin circle theorem // Pacif. J. Math.— 1962.— 12.— P. 1241—1250.

107. Filipponi P. An algorithm for computing functions of traingular matrices // Computing.— 1981.— 26.— P. 67—71.
108. Flamm D. S., Walker R. A. Remark on Algorithm 506. HQR3 and EXCHNG: Fortran subroutines for calculating and ordering the eigenvalues of a real upper Hessenberg matrix // ACM Trans. Math. Software.— 1982.— 8. № 2.— P. 219—220.
109. Francis J. G. F. The QR-transformation—a unitary analogue to the LR-transformation // Comput. J.— I: 1961.— 4.— P. 265—271; II: 1962.— 4.
110. Frank W. L. Computing eigenvalues of complex matrices by determinant evaluation and by methods of Danilewski and Wielandt // J. SIAM.— 1958.— 6, № 4.— 378—392.
111. Gambolati G., Perdon A. Minimal eigenvalue of large sparse matrices by an efficient reverse power-conjugate gradient scheme // Comput. Meth. Appl. Mech. and Eng.— 1983.— 41, № 1.— P. 1—10.
112. Garbow B. S., Boyle J. M., Dongarra J. J., Moler C. B. Matrix eigensystem routines— EISPACK guide extension // Lecture Notes Comput. Sci.— 1977.— 51.
113. Geurts A. J. A contribution to the theory of condition // Numer. Math.— 1982.— 39.— P. 85—96.
114. Goldstein M. J. Reduction of the eigenproblem for Hermitian persymmetric matrices // Math. Comput.— 1974.— 28, № 125.— P. 237—238.
115. Golub G. H., Nash S., Van Loan Ch. F. A Hessenberg—Schur method for the problem $AX+XB=C$ // IEEE Trans. Automat. Contr. 1979.— 24, № 6.— P. 909—913.
116. Golub G. H., Van Loan Ch. F. Matrix computations.— Baltimore: The John Hopkins University Press, 1983.
117. Gragg W. B. The QR algorithm for unitary Hessenberg matrices // J. Comput. Appl. Math.— 1986.— 16, № 1.— P. 1—8.
118. Grünbaum F. A. A remark on Hilbert's matrix // Linear Algebra Appl.— 1982.— 43.— P. 119—124.
119. Hake J. Fr. A remark on Frank matrices // Computing.— 1985.— 35.
120. Henrici P. Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices // Numer. Math.— 1962.— 4.— P. 24—40.
121. Higham N. J. Computing real square roots of a real matrix // Linear Algebra Appl.— 1987.— 88/89.— P. 239—270.
122. Hille E. Analytic function theory. Vol. II.— Boston: Ginn, 1962.
123. Hoffman A. J., Wielandt H. W. The variation of the spectrum of a normal matrix // Duke Math. J.— 1953.— 20.— P. 37—39.
124. Huang C. P. On the convergence of the QR algorithm with origin shifts for normal matrices // IMA J. Numer. Anal.— 1981.— 1, N 1.— P. 127—133.
125. Ikebe Y., Inagaki T., Miyamoto S. Perturbation theorems for matrix eigenvalues // J. Inform. Processing.— 1983.— 6, N 2.— P. 92—94.
126. Jennings A. A direct iteration method of obtaining latent roots and vectors of a symmetric matrix // Proc. Cambridge Phil. Soc.— 1967.— 63, N 3.— P. 755—765.
127. Jennings A., Stewart W. J. Simultaneous iteration for partial eigensolution of real matrices // J. Inst. Math. Appl.— 1975.— 15.— P. 351—361.
128. Jiang E. On spectral variation of a nonnormal matrix // Linear Algebra Appl.— 1982.— 42.— P. 223—241.
129. Jiang E. The convergence rate of the QL-algorithm with shifts for symmetric tridiagonal matrix // Numer. Math. J. Chin. Univ.— 1985.— 7, N 1.— P. 16—30.
130. Jiang E., Zhang Z. A new shift of the QL-algorithm for irreducible symmetric tridiagonal matrices // Linear Algebra Appl.— 1985.— 65.— P. 261—272.
131. Kahan W. Inclusion theorems for clusters of eigenvalues of Hermitian matrices. Stanford University, Technical Report CS42, 1967.
132. Kahan W., Parlett B. N., Jiang E. Residual bounds on approximate eigensystems of nonnormal matrices // SIAM J. Numer. Anal.— 1982.— 19, N 3.— P. 470—484.
133. Kaniel S. Estimates for some computational techniques in linear algebra // Math. Comput.— 1966.— 20, N 95.— P. 369—378.
134. Kovarik Z. V., Olesky D. D. Sharpness of generalized Gershgorin discs // Linear Algebra Appl.— 1974.— 8.— P. 477—482.

135. Lanczos C. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators // J. Res. Nat. Bur. Stand. B.— 1950.— 45, N 4.— P. 255—281.
136. Lanczos C. Solution of systems of linear equations by minimized iterations // J. Res. Nat. Bur. Stand. B.— 1952.— 49, N 1.— P. 33—53.
137. Laub A. J. A Schur method for solving algebraic Riccati equations // IEEE Trans. Automat. Contr.— 1979.— 24, N 6.— P. 913—921.
138. Lebaud C. L'algorithme double QR avec «shift» // Numer. Math.— 1970.— 16.
139. Lee A. Centrohermitian and skew-centrohermitian matrices // Linear Algebra Appl.— 1980.— 29.— P. 205—210.
140. Lewis J. G., Grimes R. G. Practical Lanczos algorithms for solving structural engineering eigenvalue problems. In: Sparse matrices and their uses. New York: Academic Press, 1981, p. 349—355.
141. Lin P. On spectral variation of approximate invariant subspaces of a nonnormal matrices // J. Nanjing Univ., Math. Biq. 1986.— 3.— P. 127—134.
142. Manteuffel T. A. The Tchebychev iteration for nonsymmetric linear systems // Numer. Math.— 1977.— 28.— P. 307—327.
143. Manteuffel T. A. Adaptive procedure for estimating parameters for the nonsymmetric Tchebychev iteration.— 1978.— 31.— P. 183—208.
144. McAllister D. F., Stewart G. W., Stewart W. J. On a Rayleigh — Ritz refinement technique for nearly uncoupled stochastic matrices // Linear Algebra Appl.— 1984.— 60.— P. 1—25.
145. Meyer A. Die Berechnung von Eigenwerten nichtsymmetrischer Matrizen in gegebenen Kreisen // Wiss. Schriftenr. Techn. Hochsch. Karl-Marx-Stadt.— 1979.— 5.— P. 28—38.
146. Nanda T. Differential equations and the QR algorithm // SIAM J. Numer. Anal.— 1985.— 22, N 2.— P. 310—321.
147. Nour-Omid B., Parlett B. N., Taylor R. L. Lanczos versus subspace iteration for solution of eigenvalue problems // Int. J. Numer. Meth. Eng.— 1983.— 19, N 6.— P. 859—871.
148. Ostrowski A. M. Uber das Nichtverschwinden einer Klasse von Determinanten und die Lokalisierung der charakteristischen Wurzeln von Matrizen // Compositio Math.— 1951.— 9.— P. 209—226.
149. Ostrowski A. M. Über die Stetigkeit von charakteristischen Wurzeln in Abhängigkeit von den Matrizenelementen // Jber. Deutsch. Math. Verein.— 1957.— 60, N 1.— P. 40—42.
150. Ostrowski A. M. On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors // Arch. Ration. Mech. and Analysis.— I: 1958.— 1, N 3.— P. 233—241; II: 1959.— 2, N 5.— P. 423—428; III: 1959.— 3, N 4.— P. 325—340; IV: 1959.— 3, N 4.— P. 341—347; V: 1959.— 33, N 5.— P. 472—481; VI: 1959.— 4, N 2.— P. 153—165.
151. Paige C. C. Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix // J. Inst. Math. Appl.— 1976.— 18, N 3.— P. 341—349.
152. Paige C. C., Van Loan Ch. F. A Schur decomposition for Hamiltonian matrices // Linear Algebra Appl.— 1981.— 41.— P. 11—32.
153. Parlett B. N. Laguerre's method applied to the matrix eigenvalue problem // Math. Comput.— 1964.— 18.— P. 464—487.
154. Parlett B. N. Convergence of the QR algorithm // Numer Math.— 1965.— 7.— P. 187—193. (Correction in [10.— P. 163—164]).
155. Parlett B. N. Singular and invariant matrices under the QR algorithm // Math. Comput.— 1966.— 20.— P. 611—615.
156. Parlett B. N. Global convergence of the basic QR algorithm on Hessenberg matrices // Math. Comput.— 1968.— 22.— P. 803—817.
157. Parlett B. N. Normal Hessenberg and moment matrices // Linear Algebra Appl.— 1973.— 6.— P. 37—43.
158. Parlett B. N. The Rayleigh quotient iteration and some generalizations for nonnormal matrices // Math. Comput.— 1974.— 28.— P. 679—693.
159. Parlett B. N. A recurrence among the elements of functions of triangular matrices // Linear Algebra Appl.— 1976.— 14.— P. 117—121.
160. Parlett B. N. The software scene in the extraction of eigenvalues from sparse matrices // SIAM J. Sci. Stat. Comput.— 1984.— 5, N 3.— P. 590—604.

161. Parlett B. N., Poole W. G. A geometric convergence theory for the QR, LU, and power iterations // SIAM J. Numer. Anal.— 1973.— 10.— P. 389—412.
162. Parlett B. N., Reinsch C. Balancing a matrix for calculation of eigenvalues and eigenvectors // Numer. Math.— 1969.— 13.— P. 292—304.
163. Parlett B. N., Taylor D. R., Liu Zh. A. A look-ahead Lanczos algorithm for unsymmetric matrices // Math. Comput.— 1985.— 44, N 169.— P. 105—124.
164. Peters G., Wilkinson J. H. Inverse iteration, ill-conditioned equations and Newton's method // SIAM Review.— 1979.— 21, N 3.— P. 339—360.
165. Powers D. L. A block Gerśgorin theorem // Linear Algebra Appl.— 1976.— 13.— P. 45—52.
166. Ruhe A. The two-sided Arnoldi algorithm for nonsymmetric eigenvalue problems // Lecture Notes Math.— 1983.— 973.— P. 104—120.
167. Ruhe A. Closest normal matrix finally found // BIT.— 1987.— 27, N 4.— P. 585—598.
168. Rutishauser H. Computational aspects of F. L. Bauer's simultaneous iteration method // Numer. Math.— 1969.— 13.— P. 4—13.
169. Rutishauser H. Simultaneous iteration method for symmetric matrices // Numer. Math.— 1970.— 16.— P. 205—223.
170. Saad Y. Etudes des translations d'origine dans les algorithmes LR et QR // C. r. Acad. sci.— 1974.— A278, N 2.— P. 93—96.
171. Saad Y. Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices // Linear Algebra Appl.— 1980.— 34.— P. 269—295.
172. Saad Y. On the rates of convergence of the Lanczos and the block-Lanczos methods // SIAM J. Numer. Anal.— 1980.— 17, N 5.— P. 687—706.
173. Saad Y. Krylov subspace methods for solving large unsymmetric linear systems // Math. Comput.— 1981.— 37, N 155.— P. 105—126.
174. Saad Y. Projection methods for solving large sparse eigenvalue problems // Lecture Notes Math.— 1983.— 973.— P. 121—144.
175. Saad U. Chebyshev acceleration techniques for solving nonsymmetric eigenvalue problems // Math. Comput.— 1984.— 42, N 166.— P. 567—588.
176. Saad Y., Shultz M. H. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems // SIAM J. Sci. Stat. Comput.— 1986.— 7, N 3.— P. 856—869.
177. Scott D. S. The Lanczos algorithm. In: Sparse matrices and their uses.— N. Y.: Academic Press, 1981, p. 139—159.
178. Shi J., Ziao T. Error bounds on perturbation of eigenvalue of arbitrary matrix // Numer. Math. J. Chin. Univ.— 1987.— 9, № 2.— P. 190—192.
179. Van der Sluis A. Gershgorin domains for partitioned matrices // Linear Algebra Appl.— 1979.— 26.— P. 265—280.
180. Smith B. T., Boyle J. M., Dongarra J. J., Garbow B. S., Ikebe Y., Klema V. C., Moler C. B. Matrix eigensystem routines.— EISPACK guide // Lecture Notes Comput. Sci.— 1976.— 6.
181. Sommer J. P. Eine Modifikation des QR-Algorithmus zur Verbesserung vorgegebener Eigenwertnäherungen bei allgemeinen reellen Matrizen // Wiss. Schriftenr. Techn. Hochsch. Karl-Marx-Stadt.— 1979.— 5.— P. 47—57.
182. Song Y. Zh. Perturbation bounds for eigenvalues of arbitrary matrices // Math. Numer. Sinica.— 1986.— 8.— P. 101—105.
183. Stewart G. W. Error bounds for approximate invariant subspaces of closed linear operators // SIAM J. Numer. Anal.— 1971.— 8.— P. 796—808.
184. Stewart G. W. Introduction to matrix computations.— N. Y.: Academic Press, 1973.
185. Stewart G. W. Error and perturbation bounds for subspaces associated with certain eigenvalue problems // SIAM. Review.— 1973.— 15.— P. 727—764.
186. Stewart G. W. Methods of simultaneous iterations for calculating eigenvectors of matrices. In: Topics in numerical analysis.— L.: Academic Press, 1975, p. 185—196.
187. Stewart G. W. Simultaneous iteration for computing invariant subspaces of non-Hermitian matrices // Numer. Math.— 1976.— 25.— P. 123—126.
188. Stewart G. W. Algorithm 506. HQR3 and EXCHNG: Fortran subroutines for calculating and ordering the eigenvalues of a real upper Hessenberg matrix [F2] // ACM Trans. Math. Software.— 1976.— 2, № 3.— P. 275—280.

189. Stewart W. J., Jennings A. A simultaneous iteration algorithm for real matrices // ACM Trans. Math. Software. — 1981. — 7, № 2. — P. 184 — 198.
190. Strachey G., Francis J. G. F. The reduction of a matrix to codiagonal form by eliminations // Comput. J. — 1961. — 4. — P. 168 — 176.
191. Symes W. W. The QR-algorithm and scattering for the finite nonperiodic Toda lattice // Physica. — 1982. — D4. — P. 275 — 280.
192. Sun Ji-Guang. On the perturbation of the eigenvalues of a normal matrix // Math. Numer. Sinica. — 1984. — 6, № 3. — P. 334 — 336.
193. Taussky O. A recurring theorem on determinants // Amer. Math. Montly. — 1949. — 56. — P. 672 — 676.
194. Temple G. The accuracy of Rayleigh's method of calculating the natural frequencies of vibrating systems // Proc. Roy. Soc. London, Ser. A. — 1952. — 211. — P. 204 — 224.
195. Van Loan Ch. F. A symplectic method for approximating all the eigenvalues of a Hamiltonian matrix // Linear Algebra Appl. — 1984. — 61. — P. 233 — 251.
196. Van Loan Ch. F. On estimating the condition of eigenvalues and eigenvectors // Linear Algebra Appl. — 1987. — 88/89. — P. 715 — 732.
197. Varah J. M. The calculation of the eigenvectors of a general complex matrix by inverse iteration // Math. Comput. — 1968. — 22. — P. 785 — 791.
198. Varah J. M. Rigorous machine bounds for the eigensystem of a general complex matrix // Math. Comput. — 1968. — 22. — P. 793 — 801.
199. Varah J. M. Computing invariant subspaces of a general matrix when the eigensystem is poorly conditioned // Math. Comput. — 1970. — 24.
200. Varah J. M. A generalization of the Frank matrix // SIAM J. Sci. Stat. Comput. — 1986. — 7, № 3. — P. 835 — 839.
201. Varga R. S. A comparison of successive over relaxation and semi-iterative methods using Chebyshev polynomials // SIAM J. Numer. Anal. — 1957. — 5. — P. 39 — 46.
202. Varga R. S. Minimal Gerschgorin sets // Pacific. J. Math. — 1965. — 15.
203. Ward R. C., Gray L. J. Eigensystem computation for skew-symmetric matrices and a class of symmetric matrices // ACM Trans. Math. Software. — 1978. — 4, № 3. — P. 278 — 285.
204. Ward R. C., Gray L. J. Algorithm 530. An algorithm for computing the eigensystem of skew-symmetric matrices and a class of symmetric matrices [F2] // ACM Trans. Math. Software. — 1978. — 4, № 3. — P. 286 — 289.
205. Watkins D. S. Understanding the QR algorithm // SIAM Review. — 1982. — 24, № 4. — P. 427 — 440.
206. Weaver J. R. Centrosymmetric (cross-symmetric) matrices, their basic properties, eigenvalues, and eigenvectors // Amer. Math. Monthly. — 1985. — 92, № 10. — P. 711 — 717.
207. Wielandt H. Das Iterationsverfahren bei nicht selbstadjungierten linearen Eigenwertaufgaben // Math. Z. — 1944. — 60. — P. 93 — 143.
208. Wilkinson J. H. A global convergence of the tridiagonal QR algorithm with origin shifts // Linear Algebra Appl. — 1968. — 1. — P. 409 — 420.
209. Wilkinson J. H. Note on matrices with a very ill-conditioned eigenproblem // Numer. Math. — 1972. — 19. — P. 176 — 178.
210. Wilkinson J. H. On neighbouring matrices with quadratic elementary divisors // Numer. Math. — 1984. — 44. — P. 1 — 21.
211. Wilkinson J. H. On a theorem of Feingold // Linear Algebra Appl. — 1987. — 88/89. — P. 13 — 30.
212. Williams G., Williams D. The power method for finding eigenvalues on a microcomputer // Amer. Math. Monthly. — 1986. — 93, № 7. — P. 562 — 564.
213. Williams G., Williams D. Linear algebra computer companion. — Boston: Allyn and Bacon, 1984.
214. Wrigley H. E. Accelerating the Jacobi method for solving simultaneous equations by Chebyshev extrapolation when the eigenvalues of the iteration matrix are complex // Computer. J. — 1963. — 6. — P. 169 — 176.
215. Zenger C. Positivity in complex spaces and its application to Gershgorin discs // Numer. Math. — 1984. — 44. — P. 67 — 73.
216. Zhang Z. On the perturbations of the eigenvalues of a nondefective matrix // Math. Numer. Sinica. — 1986. — 8, № 1. — P. 106 — 108.
217. Zhang Z. A shift in the QL algorithm // Math. Numer. Sinica. — 1987. — 9, № 1.

ОГЛАВЛЕНИЕ