Записки научных
семинаров ПОМИ
Том 346, 2007 г.

А. Н. Борзых

О СХОДИМОСТИ ОДНОГО ОПТИМИЗАЦИОННОГО АЛГОРИТМА ВЫЧИСЛЕНИЯ НАИБОЛЬШЕГО СОБСТВЕННОГО ЗНАЧЕНИЯ СИММЕТРИЧНОЙ МАТРИЦЫ

В ряде инженерных приложений требуется вычисление только наибольшего собственного значения заданной матрицы (частичная проблема собственных значений), что требует меньших вычислительных затрат, нежели вычисление всех ее собственных значений (полная проблема). Для решения данной задачи возможно применение степенного метода (см., например, [1, 2, 3, 4]), методов релаксации отношения Релея (см., например, [2]), метода бисекции (см., например, [5]). Степенной метод является простейшим и универсальным (применим к матрицам любой структуры), однако при реальных вычислениях он обычно неэффективен (его скорость сходимости существенно зависит от близости двух старших собственных значений). Группа методов релаксации отношения Релея включает различные стратегии выбора поправок для перехода от предыдущего вектора к последующему. Какая из стратегий более эффективна – вопрос практического эксперимента. Сфера применимости метода релаксации отношения Релея – симметричные матрицы, метода бисекции – трехдиагональные матрицы.

В работе [6] автором был предложен новый эффективный алгоритм вычисления наибольшего собственного значения. В данной статье предлагается несколько иная модификация алгоритма и доказывается линейная скорость сходимости.

Предлагаемый алгоритм вычисляет наибольшее собственное значение произвольной вещественной симметричной матрицы. Алгоритм основывается на последовательных преобразованиях подобия, обеспечивающих итерационное увеличение суммы элементов матрицы. Показывается, что при осуществляемых преобразованиях сумма элементов матрицы является величиной ограниченной и сходится к своему максимальному значению линейно. С другой стороны, рассматривается поведение строковых сумм. Показывается, что на интервале, ограниченном наименьшей и наибольшей строковыми суммами, существует хотя бы одно собственное число, а в пределе выполнения итерационного процесса данный интервал сужается до точки.

В работе используются следующие обозначения:

$$p_i(A) = \sum_k a_{ik}, \quad p(A) = \frac{1}{n}\sum_i p_i(A);$$

$$d_i(A) = p_i(A) - p(A), \quad d(A) = \sqrt{\frac{1}{n}\sum_i d_i(A)^2};$$

$T_{ij}(\alpha)$ – ортогональная матрица вращения:

$$T_{ij}(\alpha) = \begin{array}{cc} \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & c & & \dots & & -s & \\ & & & \ddots & & & & \\ & & \vdots & & 1 & & \vdots & \\ & & & & & \ddots & & \\ & & s & & \dots & & c & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix} & \begin{array}{c} \\ \\ \longleftarrow i \\ \\ \\ \\ \longleftarrow j \\ \\ \\ \end{array} \end{array},$$

где $s = \sin\alpha, c = \cos\alpha$;

$e = (1, 1, \dots, 1)^T$ – вектор из единиц;

I – единичная матрица;

$\|A\| = \|A\|_2 = \sqrt{\lambda_{\max}(A^*A)}$ – спектральная норма матрицы;

$\|A\|_\infty = \max\limits_i \sum\limits_j |a_{ij}|$ – ∞-норма матрицы;

$\|A\|_E = \sqrt{\sum\limits_{ij} a_{ij}^2}$ – евклидова норма матрицы.

Сперва представим вычислительную схему предлагаемого алгоритма, а затем рассмотрим его математическое обоснование.

Вычислительная схема

1. Выберем строки с наименьшей и наибольшей строковыми суммами:

$$i = \arg\min_k(p_k(A)), \quad j = \arg\max_k(p_k(A)).$$

2. Выполним преобразование подобия

$$B = T_{ij}^{-1} A T_{ij} = T_{ij}^{T} A T_{ij}. \qquad (1)$$

Параметр α в матрице вращения T_{ij} выберем так, чтобы сумма элементов матрицы B была наибольшей, т.е. решим задачу на поиск экстремума:

$$f(\alpha) = \sum_{k,m} b_{km} \xrightarrow{\alpha} \max.$$

Для этого выпишем в явном виде элементы матрицы B:

$$\begin{aligned}
b_{km} &= b_{mk} = a_{km} = a_{mk}, \quad k \neq i,j; \quad m \neq i,j;\\
b_{ik} &= b_{ki} = c a_{ik} + s a_{jk}, \quad k \neq i,j;\\
b_{jk} &= b_{kj} = c a_{jk} - s a_{ik}, \quad k \neq i,j;\\
b_{ii} &= a_{ii} c^2 + 2 a_{ij} sc + a_{jj} s^2;\\
b_{jj} &= a_{ii} s^2 - 2 a_{ij} sc + a_{jj} c^2;\\
b_{ij} &= b_{ji} = a_{ij}(c^2 - s^2) + (a_{jj} - a_{ii}) sc.
\end{aligned} \qquad (2)$$

Введем обозначения:

$$\tilde{p}_i = p_i(A) - a_{ii} - a_{ij} = \sum_{k \neq i,j} a_{ik},$$

$$\tilde{p}_j = p_j(A) - a_{ji} - a_{jj} = \sum_{k \neq i,j} a_{jk}.$$

Тогда имеем:

$$\begin{aligned}
f(\alpha) = 2(\tilde{p}_j - \tilde{p}_i)\sin\alpha + 2(\tilde{p}_i + \tilde{p}_j)\cos\alpha + a_{ii} + a_{jj} + 2a_{ij}\cos(2\alpha)\\
+ (a_{jj} - a_{ii})\sin(2\alpha) + \sum_{k,m \neq i,j} a_{km};
\end{aligned}$$

$$\begin{aligned}
f'(\alpha) = -4a_{ij}\sin(2\alpha) + 2(a_{jj} - a_{ii})\cos(2\alpha)\\
-2(\tilde{p}_i + \tilde{p}_j)\sin\alpha + 2(\tilde{p}_j - \tilde{p}_i)\cos\alpha.
\end{aligned}$$

Решая уравнение $f'(\alpha) = 0$, получаем:

$$c_1 \sin(2\alpha) + c_2 \cos(2\alpha) + c_3 \sin\alpha + c_4 \cos\alpha = 0, \qquad (3)$$

где

$$c_1 = 2a_{ij}, \quad c_2 = a_{ii} - a_{jj}, \quad c_3 = \tilde{p}_i + \tilde{p}_j, \quad c_4 = \tilde{p}_i - \tilde{p}_j.$$

Очевидно, что уравнение (3) имеет 2, 3 или 4 вещественных корня. Их приближенные значения можно вычислить каким-либо численным методом за время $O(1)$. Среди нескольких найденных корней в качестве α возьмем тот, при котором значение $f(\alpha)$ будет наибольшим (т. е. тот, при котором будет обеспечен глобальный максимум суммы элементов матрицы B).

3. Положим $A = B$.

4. Пусть ε – требуемая точность вычислений. Если $d(A) < \varepsilon$, то прекратить вычисления и в качестве приближенного значения наибольшего собственного числа взять

$$\lambda_{\max} \approx p(A);$$

при этом $\lambda_{\max} \in [p(A) - d(A), p(A) + d(A)]$.

5. Иначе перейти к п. 1.

Рассмотрим математическое обоснование представленного алгоритма.

Утверждение 1. *На каждом шаге алгоритма хотя бы одно собственное значение лежит в интервале* $[p(A) - d(A), p(A) + d(A)]$.

Доказательство. В силу симметричности исходной матрицы и свойств матриц вращения, можно утверждать, что на каждом шаге алгоритма матрица A сохраняет симметричность. Рассмотрим теорему об оценке невязки ([3, §4.5; 4, §7]).

Пусть матрица A диагонализуема, т.е. существует матрица P такая, что $P^{-1}AP = D$, где $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Тогда для произвольных $\mu \in C$ и $x \in R^n$ существует $\lambda \in \sigma(A)$ такое, что $|\lambda - \mu| \leqslant \operatorname{cond} P \, \frac{\|Ax - \mu x\|}{\|x\|}$.

Здесь и ниже через $\sigma(A)$ обозначается спектр матрицы A.

Выберем μ и x следующим образом: $\mu = p(A)$, $x = e$. Тогда существует $\lambda \in \sigma(A)$ такое, что

$$|\lambda - \mu| \leqslant \underbrace{\operatorname{cond} P}_{=1} \frac{\|Ax - \mu x\|}{\|x\|}$$

$$= \frac{\left\| \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix} - \begin{pmatrix} p \\ \vdots \\ p \end{pmatrix} \right\|}{\|e\|} = \frac{\left\| \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix} \right\|}{\sqrt{n}} = \frac{\sqrt{\sum\limits_i d_i^2}}{\sqrt{n}} = \sqrt{\frac{1}{n} \sum_i d_i^2} = d.$$

Следовательно, $\lambda \in [p(A) - d(A), p(A) + d(A)]$. $\square$

Утверждение 2. *На каждом шаге алгоритма преобразование подобия* (1) *обеспечивает равенство сумм элементов строк i и j: $p_i(B) = p_j(B)$.*

Доказательство. Из формул (2) имеем:

$$p_i(B) = (c\tilde{p}_i + s\tilde{p}_j) + (a_{ii}c^2 + 2a_{ij}sc + a_{jj}s^2) + b_{ij},$$

$$p_j(B) = (c\tilde{p}_j - s\tilde{p}_i) + (a_{ii}s^2 - 2a_{ij}sc + a_{jj}c^2) + b_{ij}.$$

Следовательно,

$$p_i(B) - p_j(B) = (a_{ii} - a_{jj})(c^2 - s^2) + 4a_{ij}sc + (\tilde{p}_i + \tilde{p}_j)s + (\tilde{p}_i - \tilde{p}_j)c$$

$$= \underbrace{2a_{ij}}_{c_1}\sin(2\alpha) + \underbrace{(a_{ii} - a_{jj})}_{c_2}\cos(2\alpha) + \underbrace{(\tilde{p}_i + \tilde{p}_j)}_{c_3}\sin\alpha + \underbrace{(\tilde{p}_i - \tilde{p}_j)}_{c_4}\cos\alpha$$

$$= c_1\sin(2\alpha) + c_2\cos(2\alpha) + c_3\sin\alpha + c_4\cos\alpha = 0. \quad \square$$

Замечание 1. Отметим, что в описании предложенного алгоритма параметр α выбирается как решение экстремальной задачи $e^T Be \xrightarrow{\alpha} \max$. С другой стороны, максимизировав по α сумму всех элементов B, получаем равенство $p_i(B) = p_j(B)$. Таким образом, задача максимизации суммы всех элементов и задача "уравнивания" строковых сумм оказываются эквивалентными (обе приводят к уравнению (3)).

Утверждение 3. *На каждом шаге алгоритма после преобразования подобия* (1) *сумма всех элементов матрицы не уменьшается, т.е. $\sum\limits_{k,m} b_{km} \geqslant \sum\limits_{k,m} a_{km}$.*

Доказательство. Так как α является точкой глобального максимума $f(\alpha)$, то для любого α_0 имеем $f(\alpha) \geqslant f(\alpha_0)$. В частности, при $\alpha_0 = 0$ мы имеем

$$\sum_{k,m} b_{km} = f(\alpha) \geqslant f(0) = \sum_{k,m} a_{km}. \quad \square$$

Утверждение 4. *В ходе выполнения алгоритма сумма всех элементов матрицы стремится к некоторому предельному значению:* $\sum\limits_{k,m} a_{km} \to s.$

Доказательство. Ранее было показано, что сумма всех элементов матрицы является величиной неубывающей. Осталось показать, что данная сумма остается ограниченной сверху.

Обозначим через A_k и T_k матрицы A и T, полученные на k-ом шаге алгоритма; $A_0 = A$. Имеем:

$$A_k = T_k^T A_{k-1} T_k = T_k^T T_{k-1}^T \cdots T_1^T A_0 T_1 T_2 \cdots T_k$$

Введем обозначение: $e_k = T_1 T_2 \cdots T_k e$.

Отметим, что $(e_k, e_k) = (e, e) = n$.

Сумма всех элементов A_k есть

$$e^T A_k e = \underbrace{e^T T_k^T \cdots T_1^T}_{e_k^T} A_0 \underbrace{T_1 \cdots T_k e}_{e_k} = e_k^T A_0 e_k = (A_0 e_k, e_k)$$

$$= n \frac{(A_0 e_k, e_k)}{(e_k, e_k)} \leqslant n \lambda_{\max}(A_0). \quad \square$$

Замечание 2. Заметим, что задача максимизации суммы всех элементов оказывается тождественной задаче максимизации отношения Релея. Действительно, взглянем на алгоритм с иной стороны: A_0 есть неизменная исходная матрица, T_k – порождаемые алгоритмом элементарные матрицы вращения, $e_k = T_1 T_2 \cdots T_k e$ – порождаемые векторы. Параметры матриц вращения T_k выбираются как решения экстремальных задач $(Ae_k, e_k) \to \max$.

Методы вычисления наибольшего собственного числа, основанные на максимизации отношения Релея, называются методами релаксации [2, §61]. Выделяют методы релаксации по одной из координат, по группе координат (групповая координатная релаксация), а также по направлению градиента (см. [2, §74]). Предлагаемый алгоритм сходен с методом координатной релаксации, однако релаксация на каждом шаге ведется не по координате, а по углу в двумерном координатном пространстве.

Лемма 1. *Пусть векторы x и y удовлетворяют соотношению $\|x\| = \|y\|$. Пусть θ – острый угол между векторами x и y:*

$$\cos\theta = \frac{(x,y)}{\|x\|\,\|y\|}, \quad 0 < \theta < \frac{\pi}{2}.$$

Тогда существует такая ортогональная матрица U, что $Ux = y$ и выполнено соотношение $\|U - I\| = 2\sin(\theta/2)$.

Доказательство. Можно считать, что $\|x\| = \|y\| = 1$. Действительно, если лемма 1 справедлива для нормированных x и y, то она также справедлива и для произвольных x^* и y^* таких, что $\|x^*\| = \|y^*\|$.

Построим вектор z, удовлетворяющий следующим условиям:

1) $z = a_1 x + a_2 y$ (z – линейная комбинация x, y);

2) $(z, x) = 0$ (z ортогонален x);

3) $\|z\| = 1$ (z имеет единичную норму).

Из условия (2) следует, что

$$(z,x) = a_1(x,x) + a_2(y,x) = a_1 + a_2\cos\theta = 0,$$

$$a_1 = -a_2\cos\theta.$$

Из условия (3) имеем:

$$(z,z) = a_1^2 + 2a_1a_2\cos\theta + a_2^2 = 1,$$

и так как $a_1 = -a_2\cos\theta$, то $a_2 = \frac{1}{\sin\theta}$. Тогда $a_1 = -\operatorname{ctg}\theta$.

Построим ортогональную матрицу B, у которой первая строка – вектор x, вторая строка – вектор z (ортогонален x), а остальные строки ортогональны векторам x, z и друг другу. Из линейной алгебры известно, что такое построение всегда возможно.

Отметим следующие свойства матрицы B:

1) $Bx = x_0 = (1,0,0,\ldots,0)^T$; $\quad B^T x_0 = x$;

2) $Bz = z_0 = (0,1,0,\ldots,0)^T$; $\quad B^T z_0 = z$;

3)

$$By = \left(-\frac{a_1}{a_2}, \frac{1}{a_2}, 0, \ldots, 0\right)^T = (\cos\theta, \sin\theta, 0, \ldots, 0)^T = y_0; \quad B^T y_0 = y.$$

Пусть $\mathfrak{Q}$ – элементарная матрица вращения вида

$$\mathfrak{Q} = \begin{pmatrix} c & -s & 0 & & 0 \\ s & c & 0 & \vdots & 0 \\ 0 & 0 & 1 & & 0 \\ & \dots & & \ddots & \\ 0 & 0 & 0 & & 1 \end{pmatrix}, \quad \text{где} \quad s = \sin\theta,\ c = \cos\theta.$$

Тогда

$$\mathfrak{Q}x_0 = \mathfrak{Q}(1,0,0,\dots,0)^T = (c,s,0,\dots,0)^T = y_0$$

и

$$\|\mathfrak{Q} - I\| = \left\| \begin{pmatrix} c-1 & -s \\ s & c-1 \end{pmatrix} \right\| = 2\sin(\theta/2).$$

Мы утверждаем, что $U = B^T \mathfrak{Q} B$ – искомая матрица. Действительно,

$$Ux = B^T \mathfrak{Q} \underbrace{Bx}_{x_0} = B^T \underbrace{\mathfrak{Q}x_0}_{y_0} = B^T y_0 = y,$$

а поскольку матрица $U - I$ ортогонально подобна $\mathfrak{Q} - I$, то

$$\|U - I\| = \|\mathfrak{Q} - I\| = 2\sin(\theta/2). \quad \square$$

Утверждение 5. *Для любого* $\beta \in [0, \pi/8]$ *справедливо неравенство*

$$\sin\beta \leqslant \sin(\pi/8)\sin(4\beta).$$

Доказательство. Определим функцию

$$t(\beta) = \frac{\sin(4\beta)}{\sin\beta} = 4\cos\beta\cos(2\beta)$$

и заметим, что на интервале $[0, \pi/8]$ функция $f(\beta)$ монотонно убывает, так как $\cos\beta$ и $\cos(2\beta)$ – монотонно убывающие функции. Следовательно, для любого $\beta \in [0, \pi/8]$ мы имеем

$$t(\beta) \geqslant t(\pi/8) = [\sin(\pi/8)]^{-1},$$

$$\frac{\sin(4\beta)}{\sin\beta} \geqslant [\sin(\pi/8)]^{-1},$$

$$\sin\beta \leqslant \sin(\pi/8)\sin(4\beta). \quad \square$$

Лемма 2. *Пусть A – симметричная матрица, $\lambda_1, \dots, \lambda_n$ – собственные числа A, $x_1, \dots, x_n$ – соответствующие собственные векторы, $d(A) = \varepsilon$. Согласно утверждению 1, на интервале $[p(A) - \varepsilon,\ p(A) + \varepsilon]$ существует хотя бы одно собственное число A. Пусть ε мало настолько, что на интервале $[p(A) - \varepsilon\sqrt{n},\ p(A) + \varepsilon\sqrt{n}]$ существует ровно одно собственное число λ_k, и пусть оно простое.*

Тогда существуют ортогональная U и симметричная A^* такие, что

$$UAU^T = A^*, \quad d(A^*) = 0, \quad \|U - I\| \leqslant C_1\varepsilon,$$

$$C_1 = \frac{4\sqrt{n}}{\min_{i \neq k} |\lambda_i - \lambda_k|} \sin(\pi/8).$$

Доказательство. Возьмем $D = \operatorname{diag}(d_1, \dots, d_n)$, $d_i = p_i(A) - p(A)$. Тогда

$$\|D\| = \max_i |d_i| = \sqrt{(\max_i |d_i|)^2}$$

$$\leqslant \sqrt{\sum_i d_i^2} = \sqrt{n}\sqrt{\frac{1}{n}\sum_i d_i^2} = \sqrt{n}d(A) = \varepsilon\sqrt{n}.$$

Для $\tilde{A} = A - D$ мы имеем

$$\tilde{A}e = Ae - De = \begin{pmatrix} p_1(A) \\ \vdots \\ p_n(A) \end{pmatrix} - \begin{pmatrix} p_1(A) - p(A) \\ \vdots \\ p_n(A) - p(A) \end{pmatrix} = \begin{pmatrix} p(A) \\ \vdots \\ p(A) \end{pmatrix} = p(A)e,$$

так что $\{p(A), e\}$ – собственная пара $\tilde{A}$.

Матрицу A по отношению к матрице $\tilde{A}$ считаем “возмущенной”: $A = \tilde{A} + D$, $\|D\| \leqslant \varepsilon\sqrt{n}$.

Из теории возмущений известно, что существует собственная пара $\{\lambda_m, x_m\}$ матрицы $A = \tilde{A} + D$, удовлетворяющая следующим оценкам (см. [5, §5.2]):

1) $|\lambda_m - p(A)| \leqslant \|D\|$;

2) $\frac{1}{2}\sin(2\theta) \leqslant \frac{\|D\|}{\min\limits_{i \neq m} |\lambda_i - \lambda_m|}$, где θ – угол между векторами e и x_m.

Из первой оценки имеем:

$$|\lambda_m - p(A)| \leqslant \|D\| \leqslant \varepsilon\sqrt{n} \ \Rightarrow\ \lambda_m \in [p(A) - \varepsilon\sqrt{n},\ p(A) + \varepsilon\sqrt{n}].$$

Заметим, что, по условиям леммы 2, на интервале $[p(A) - \varepsilon\sqrt{n},\ p(A) + \varepsilon\sqrt{n}]$ существует только одно собственное число λ_k. Следовательно, λ_m есть λ_k, $m = k$.

Из второй оценки вытекает, что

$$\sin(2\theta) \leqslant 2\frac{\|D\|}{\min\limits_{i\neq m}|\lambda_i-\lambda_m|} \leqslant \frac{2\varepsilon\sqrt{n}}{\min\limits_{i\neq k}|\lambda_i-\lambda_k|}.$$

Полагаем, что $\|x_m\| = \|e\| = \sqrt{n}$. По лемме 1, существует ортогональная матрица U такая, что $Ux_m = e$, $\|U-I\| = 2\sin(\theta/2)$.

В силу утверждения 5, $\sin(\theta/2) \leqslant \sin(\pi/8)\sin(2\theta)$. Следовательно,

$$\|U-I\| \leqslant 2\sin(\pi/8)\sin(2\theta) \leqslant 2\sin(\pi/8)\frac{2\sqrt{n}}{\min\limits_{i\neq k}|\lambda_i-\lambda_k|}\varepsilon = C_1\varepsilon.$$

Представим матрицу A в виде $A = P\Lambda P^T$, где $\Lambda = \operatorname{diag}(\lambda_1,\ldots,\lambda_n)$, P – матрица собственных векторов, x_m – один из столбцов P.

Обозначим $P_2 = UP$. Тогда e – один из столбцов P_2 (так как x_m – есть столбец P, $Ux_m = e$).

Получаем: $A^* = UAU^T = UP\Lambda P^T U^T = P_2\Lambda P_2^T$, e – собственный вектор A^* (так как является столбцом P_2).

Тогда $A^*e = \lambda e$, $p_1(A^*) = \cdots = p_n(A^*) = \lambda$, $d(A^*) = 0$. □

Утверждение 6.

$$e^TBe - e^TAe \geqslant C_2 d(A)^2, \quad C_2 = \frac{3}{28\sqrt{n}\|A\|_E}.$$

Доказательство. Ранее было показано, что сумма элементов матрицы B определяется выражением

$$e^TBe = 2(\tilde{p}_j-\tilde{p}_i)\sin\alpha^* + 2(\tilde{p}_i+\tilde{p}_j)\cos\alpha^* + a_{ii} + a_{jj}$$
$$+2a_{ij}\cos(2\alpha^*) + (a_{jj}-a_{ii})\sin(2\alpha^*) + \sum_{k,m\neq i,j} a_{km},$$

где α^* – решение задачи $e^TBe \xrightarrow{\alpha^*} \max$. Следовательно, для любого α справедливо неравенство

$$e^TBe \geqslant 2(\tilde{p}_j-\tilde{p}_i)\sin\alpha + 2(\tilde{p}_i+\tilde{p}_j)\cos\alpha + a_{ii} + a_{jj}$$
$$+2a_{ij}\cos(2\alpha) + (a_{jj}-a_{ii})\sin(2\alpha) + \sum_{k,m\neq i,j} a_{km}.$$

Сделаем замены по формулам Тейлора:

$$\sin\alpha = \alpha + R_3(\alpha), \quad |R_3(\alpha)| \leqslant \frac{1}{6}\alpha^3,$$

$$\cos\alpha = 1 + R_2(\alpha), \quad |R_2(a)| \leqslant \frac{1}{2}\alpha^2.$$

Тогда получим

$$e^T Be \geqslant e^T Ae + 2(\tilde{p}_j - \tilde{p}_i)\alpha + 2\alpha(a_{jj} - a_{ii}) + R(\alpha),$$

где

$$R(\alpha) = 2R_3(\alpha)(\tilde{p}_j - \tilde{p}_i) + 2R_2(\alpha)(\tilde{p}_i + \tilde{p}_j) \\ +R_3(2\alpha)(a_{jj} - a_{ii}) + 2a_{ij}R_2(2\alpha),$$

и

$$e^T Be - e^T Ae \geqslant 2\alpha(\tilde{p}_j - \tilde{p}_i + a_{jj} - a_{ii}) + R(\alpha) = 2(p_j - p_i)\alpha + R(\alpha). \quad (4)$$

Далее предполагаем, что $|\alpha| < 1$. Поскольку

$$|R(\alpha)| \leqslant |R_3(\alpha)|\,|2(\tilde{p}_j - \tilde{p}_i)| + |R_2(\alpha)|\,|2(\tilde{p}_i + \tilde{p}_j)| \\ +|R_3(2\alpha)|\,|a_{jj} - a_{ii}| + |R_2(2\alpha)|\,|2a_{ij}| \\ \leqslant \frac{1}{6}\alpha^2 \cdot 4\|A\|_\infty + \frac{1}{2}\alpha^2 \cdot 4\|A\|_\infty + \frac{1}{6}\cdot 8\alpha^2 \cdot 2\|A\|_\infty + \frac{1}{2}\cdot 4\alpha^2 \cdot 2\|A\|_\infty \\ = \frac{28}{3}\|A\|_\infty \alpha^2 \leqslant \frac{28}{3}\sqrt{n}\|A\|_E \alpha^2 = \frac{1}{C_2}\alpha^2,$$

то $R(\alpha) \geqslant -\frac{1}{C_2}\alpha^2$, и неравенство (4) принимает вид

$$e^T Be - e^T Ae \geqslant 2(p_j - p_i)\alpha + R(\alpha) \geqslant 2(p_j - p_i)\alpha - \frac{1}{C_2}\alpha^2. \quad (5)$$

Отметим, что

$$p_j - p_i \leqslant 2\|A\|_\infty \leqslant 2\sqrt{n}\|A\|_E < \frac{1}{C_2}.$$

Положим в неравенстве (5) $\alpha = C_2(p_j - p_i) < 1$. Тогда

$$e^T Be - e^T Ae \geqslant 2\alpha(p_j - p_i) - \frac{1}{C_2}\alpha^2 \\ = 2(p_j - p_i)C_2(p_j - p_i) - \frac{1}{C_2}(C_2(p_j - p_i))^2 = C_2(p_j - p_i)^2.$$

Легко проверить, что $d(A) \leqslant p_j - p_i$. Следовательно,

$$e^T Be - e^T Ae \geqslant C_2(p_j - p_i)^2 \geqslant C_2 d(A)^2. \quad \square$$

Утверждение 7. *Пусть U – ортогональная матрица такая, что $\|U-I\| = \chi \leqslant \frac{1}{2}$. Тогда U и U^T представляются в виде*

$$U = I + S + P, \quad U^T = I - S + P^T,$$

причем S – кососимметричная матрица, $\|S\| \leqslant 2\chi$, $\|P\| \leqslant 2\chi^2$.

Доказательство. Определим $S = \log U$, где логарифм задается рядом по степеням $U - I$:

$$S = \log U = \log(I + U - I) = (U-I) - \frac{1}{2}(U-I)^2 + \frac{1}{3}(U-I)^3 - \cdots .$$

Заметим, что собственные числа S определяются таким же рядом, который дает главное значение логарифма, откуда следует, что $\log U^{-1} = -\log U$. Тогда

$$S^T = (\log U)^T = \log(U^T) = \log(U^{-1}) = -\log U = -S,$$

так что S – кососимметричная матрица.

Далее,

$$P = U - I - S = (U-I) - \log U = \frac{1}{2}(U-I)^2 - \frac{1}{3}(U-I)^3 + \cdots .$$

Остается заметить, что

$$\|P\| \leqslant \sum_{k=2}^{\infty} \frac{1}{k}\|(U-I)^k\| \leqslant \sum_{k=2}^{\infty} \chi^k = \frac{\chi^2}{1-\chi} \leqslant 2\chi^2;$$

$$\|S\| = \|U - I - P\| \leqslant \|U-I\| + \|P\| \leqslant \chi + 2\chi^2 \leqslant 2\chi;$$

$$U^T = (I + S + P)^T = I - S + P^T. \quad \square$$

Утверждение 8. *Начиная с некоторого шага алгоритма, имеет место неравенство*

$$d(A)^2 \geqslant C_3(\lambda_1 - p(A)), \quad C_3 = \frac{1}{8n^2C_1^2\|A\|}.$$

Доказательство. Согласно утверждению 6, прирост суммы всех элементов матрицы оценивается снизу величиной $C_2 d(A)^2$. Сумма всех элементов матрицы – величина ограниченная. Следовательно, $e^TBe - e^TAe \to 0$, $d(A) \to 0$.

Рассмотрим некоторый шаг алгоритма, на котором $d(A)$ достаточно мало: $d(A)=\varepsilon$, причем $C_1\varepsilon<\frac{1}{2}$.

Согласно лемме 2, существуют ортогональная матрица U и симметричная матрица A^* такие, что $UAU^T=A^*$, $d(A^*)=0$, $\|U-I\|\leqslant C_1\varepsilon$.

По утверждению 7,

$$U=I+S+P,\quad U^T=I-S+P^T,$$

где $\|S\|\leqslant 2C_1\varepsilon$, S – кососимметричная матрица, $\|P\|\leqslant 2(C_1\varepsilon)^2$. Тогда

$$A=U^TA^*U=(I-S+P^T)A^*(I+S+P)=A^*-SA^*+A^*S+P_0,$$

где $P_0=-SA^*S+P^TA^*U+U^TA^*P$.

Подсчитаем величину, равную разности суммы всех элементов A^* и суммы всех элементов A:

$$\begin{aligned}f(A^*)-f(A)=(A^*e,e)-(Ae,e)&=(A^*e,e)-((A^*-SA^*+A^*S+P_0)e,e)\\&=(SA^*e,e)-(A^*Se,e)-(P_0e,e)\\&=\lambda(Se,e)-(Se,A^*e)-(P_0e,e)=-(P_0e,e),\end{aligned}$$

так как $A^*e=\lambda e$.

Оценим $\|P_0\|$:

$$\begin{aligned}\|P_0\|&=\|-SA^*S+P^TA^*U+U^TA^*P\|\\&\leqslant\|S\|\,\|A^*\|\,\|S\|+\|P^T\|\underbrace{\|A^*U\|}_{=\|A\|}+\underbrace{\|U^TA^*\|}_{=\|A\|}\|P\|\\&\leqslant 2C_1\varepsilon\|A\|(2C_1\varepsilon)+2(C_1\varepsilon)^2\|A\|+2(C_1\varepsilon)^2\|A\|\\&=8\|A\|(C_1\varepsilon)^2.\end{aligned}$$

Тогда

$$f(A^*)-f(A)=-(P_0e,e)=-n\frac{(P_0e,e)}{(e,e)}\leqslant n\|P_0\|\leqslant 8n\|A\|(C\varepsilon)^2.$$

Поскольку

$$\frac{f(A^*)-f(A)}{d(A)^2}=\frac{f(A^*)-f(A)}{\varepsilon^2}\leqslant\frac{8n\|A\|(C_1\varepsilon)^2}{\varepsilon^2}=8nC_1^2\|A\|,$$

то

$$d(A)^2 \geqslant \frac{1}{8nC_1^2\|A\|}(f(A^*) - f(A))$$

$$= \frac{1}{8n^2C_1^2\|A\|}(\lambda_1 - p(A)) = C_3(\lambda_1 - p(A)). \quad \square$$

Теорема 1. *Алгоритм имеет линейную скорость сходимости.*

Доказательство. В силу утверждений 6 и 8, начиная с некоторого шага, имеют место неравенства

$$e^TBe - e^TAe \geqslant C_2d(A)^2, \quad d(A)^2 \geqslant C_3(\lambda_1 - p(A)).$$

Получаем:

$$p(B) - p(A) = \frac{1}{n}(e^TBe - e^TAe) \geqslant \frac{1}{n}C_2d(A)^2$$

$$\geqslant \frac{1}{n}C_2C_3(\lambda_1 - p(A)) = C_4(\lambda_1 - p(A)),$$

где $C_4 = C_2C_3/n$,

$\lambda_1 - p(A)$ – погрешность на текущем шаге,

$\lambda_1 - p(B)$ – погрешность на следующем шаге,

$$\frac{\lambda_1 - p(B)}{\lambda_1 - p(A)} = \frac{(\lambda_1 - p(A)) - (p(B) - p(A))}{\lambda_1 - p(A)}$$

$$= 1 - \frac{p(B) - p(A)}{\lambda_1 - p(A)} \leqslant 1 - C_4. \quad \square$$

Таким образом показано, что сумма всех элементов матрицы, максимизируемая на каждом шаге алгоритма, стремится к предельному значению с линейной скоростью. Предельное значение есть $n\lambda_1$, где λ_1 – одно из собственных чисел A (требуется, чтобы оно было простым). Может ли алгоритм сойтись не к наибольшему собственному числу?

Предположим, что так оно и случилось. На некотором шаге алгоритма имеем: для всех i, j, $p_i(A) \approx p_j(A) \approx \lambda_p$, где λ_p отлично от наибольшего собственного числа (знаком “$\approx$” обозначено равенство со сколь угодно малой погрешностью).

Покажем, что, с позиции методов оптимизации, данная точка есть "минимакс", то есть покажем, что сколько угодно малое смещение в пространстве взаимно подобных матриц может вызвать как увеличение, так и уменьшение целевой функции.

Для матрицы A сумма всех элементов есть

$$e^T Ae = (Ae, e) = n\frac{(Ae, e)}{(e, e)} \approx n\lambda_p.$$

Сумма всех элементов чуть возмущенной матрицы $\tilde{A} = U^T AU$, где U – произвольная ортогональная матрица близкая к единичной, есть

$$e^T \tilde{A}e = e^T U^T AUe = \tilde{e}^T A\tilde{e} = (A\tilde{e}, \tilde{e}) = n\frac{(A\tilde{e}, \tilde{e})}{(\tilde{e}, \tilde{e})},$$

где вектор $\tilde{e} = Ue$ близок к e.

Из свойств отношения Релея хорошо известно, что его минимум достигается на наименьшем собственном числе, максимум – на наибольшем, а минимаксы (стационарные точки) – на всех остальных собственных значениях. Так как λ_p не есть экстремальное собственное значение, то разность $e^T \tilde{A}e - e^T Ae$ может иметь любой знак.

В силу вышесказанного, сходимость к λ_p неустойчива. Любое малейшее возмущение (например, вызванное накоплением ошибок округления или неточностью вычисления углов вращений) может привести к тому, что сумма всех элементов матрицы превысит $n\lambda_p$, и алгоритм начнет сходиться к иному собственному числу, пока не сойдется к наибольшему.

Замечание 3. Конечно, можно придумать исключительные примеры, для которых алгоритм "застрянет" на собственном числе, отличном от наибольшего. Например, если еще до начала выполнения алгоритма исходная матрица имеет равные строковые суммы, то алгоритм остановится на первом же шаге (хотя "найденное" собственное число не обязательно будет наибольшим). В качестве практического приема, предотвращающего поиск нежелательного (т.е. не наибольшего) собственного числа, можно предложить выполнение нескольких "случайных"преобразований подобия и последующий повторный запуск алгоритма. Если по результатам двух или более запусков получается один и тот же результат, то с вероятностью 1 можно утверждать, что вычислено именно наибольшее собственное число.

Замечание 4. Очевидно, что предлагаемый алгоритм можно рассматривать как релаксационный алгоритм. Целевой функцией выступает сумма всех элементов текущей матрицы, а пространством, в котором ведется оптимизация, - пространство ортогональных матриц: $F(U) = e^T U^T A U e \xrightarrow{U} \max$.

В теории методов оптимизации известно множество теорем, представляющих достаточные условия линейной скорости сходимости релаксационных методов. Тем не менее, применение какой-либо из существующих теорем к рассматриваемой задаче оказалось чрезвычайно трудоемким (быть может невозможным) по следующим причинам.

1) Точка экстремума является вырожденной. Действительно, если U_1 есть решение $F(U) \xrightarrow{U} \max$, то для любого $\varepsilon > 0$ существует матрица U_2 такая, что $\|U_2 - U_1\| = \varepsilon$, $F(U_2) = F(U_1)$ (без доказательства).

2) Множество, на котором ищется решение, является нелинейным.

Литература

1. Дж. Х. Уилкинсон, *Алгебраическая проблема собственных значений*. Наука, М., 1970.
2. Д. К. Фаддеев, В. Н. Фаддеева, *Вычислительные методы линейной алгебры*. Лань, СПб., 2002.
3. Б. Парлетт, *Симметричная проблема собственных значений*. Мир, М., 1983.
4. Х. Д. Икрамов, *Несимметричная проблема собственных значений*. Наука, М., 1991.
5. Дж. Деммель, *Вычислительная линейная алгебра*. Мир, М., 2001.
6. А. Н. Борзых, *Эффективная локализация собственных значений симметричных матриц высокого порядка*. — Научная сессия МИФИ-2006. Сборник научных трудов. Том 7. Изд. МИФИ, М., 2006, с. 159–160.

Borzykh A. N. Convergence analysis of an optimization algorithm for computing the largest eigenvalue of a symmetric matrix.

A new optimization algorithm for computing the largest eigenvalue of a real symmetric matrix is considered. The algorithm is based on a sequence of plane rotations increasing the sum of the matrix entries. It is proved that the algorithm converges linearly and it is shown that it can be regarded as a relaxation method for the Rayleigh quotient.

С.-Петербургский
государственный университет
E-mail: borzih_alex@mail.ru

Поступило 1 октября 2007 г.