ИТЕРАЦИОННОЕ УТОЧНЕНИЕ РЕШЕНИЯ ПО МЕТОДУ НАИМЕНЬШИХ КВАДРАТОВ ДЛЯ СИСТЕМ ЛИНЕЙНЫХ УРАВНЕНИЙ*)

А. Н. Малышев

Одной из фундаментальных проблем вычислительной линейной алгебры является решение общих систем линейных уравнений методом наименьших квадратов (см. [1–4]). Итерационное уточнение решений по методу наименьших квадратов обсуждалось во многих работах (см., например, [5–8]). Однако интенсивно изучались только системы с матрицами полного ранга, в то время как системам с матрицами неполного ранга практически не уделялось внимания. В настоящей работе предлагается метод уточнения решения именно для матриц неполного ранга или для близких к ним. Описывается алгоритм и обсуждаются числа обусловленности соответствующих матриц. Исчерпывающий априорный анализ ошибок округления содержится в [9, 10]. Там же выводятся важные оценки сходимости итерационного решения уравнений Риккати, возникающих при уточнении сингулярных подпространств, отвечающих малым сингулярным числам. Мы приводим эти оценки без доказательства.

§ 1. Алгоритм решения по методу наименьших квадратов

Прежде чем ввести обозначения и понятие решения по методу наименьших квадратов, изложим стандартный алгоритм решения общих систем линейных уравнений. Строгий анализ вычислительных погрешностей при реализации этого метода в арифметике чисел с плавающей точкой можно найти в [2, 11].

Пусть A — $(M \times N)$-матрица ранга $r = \operatorname{rank} A$, $r \leqslant \min\{M, N\}$. Система линейных уравнений $Ax = f$ решается следующим образом.

1. Рассматриваем сингулярное разложение матрицы A, т. е. $A = U\Sigma V^T$, где $UU^T = I$, $VV^T = I$,

$$\Sigma = \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & 0 \\ & & \sigma_r & & \\ & & & \sigma_{r+1} & \\ 0 & & & & \ddots \\ & & & & & \sigma_{\min\{M,N\}} \end{bmatrix},$$

*) Работа выполнена за период пребывания автора в GMD (Германия) с декабря 1991 г. по май 1992 г.

$$\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_r > \sigma_{r+1} \geqslant \ldots \geqslant \sigma_{\min\{M,N\}} \geqslant 0.$$

2. Вычисляем решение диагональной системы $\Sigma y = g$, где $g = U^T f$, по формулам

$$y_1 = \frac{g_1}{\sigma_1},\ y_2 = \frac{g_2}{\sigma_2},\ \ldots,\ y_r = \frac{g_r}{\sigma_r},\ y_{r+1} = 0,\ \ldots,\ y_N = 0.$$

3. Полагаем $x = Vy$.

Нетрудно показать, что так определенное решение x системы $Ax = f$ не зависит от выбора ортогональных матриц U и V в сингулярном разложении матрицы A. При решении систем линейных уравнений методом наименьших квадратов нет необходимости вычислять полное сингулярное разложение. Можно использовать неполное сингулярное разложение — частный случай так называемого полного ортогонального разложения матрицы A. Чтобы определить понятие неполного сингулярного разложения и указать способы его применения для нахождения решений по методу наименьших квадратов, опишем алгоритм вычисления r-решения (решения ранга r) для системы $Ax = f$.

Алгоритм

1. Выполняется «двухдиагонализация» матрицы A с помощью преобразований отражения Хаусхолдера, т. е. матрица A записывается в виде $A = Q_l B Q_r^T$, где B — двухдиагональная, а Q_l, Q_r — ортогональные матрицы. Обычно матрица B верхняя двухдиагональная при $M \geqslant N$ и нижняя двухдиагональная при $M < N$.

2. При помощи вращений Якоби — Гивенса двухдиагональная матрица B приводится к виду $B = \hat{Q}_l \hat{B} \hat{Q}_r^T$, где матрицы $\hat{Q}_l$ и $\hat{Q}_r$ ортогональны,

$$\hat{B} = \begin{cases} \begin{bmatrix} C & & \\ & \sigma_{r+1} & \\ & & \ddots \\ 0 & & \\ & & \sigma_N \\ & 0 & \end{bmatrix} & \text{при } M \geqslant N, \\ \begin{bmatrix} C & & & \\ & \sigma_{r+1} & & \\ & & \ddots & \\ 0 & & & \\ & & \sigma_M & \end{bmatrix} & \text{при } M < N, \end{cases}$$

C — квадратная двухдиагональная матрица с сингулярными числами $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_r > 0$, удовлетворяющими условию $\sigma_r > \sigma_{r+1} \geqslant \ldots \geqslant \sigma_{\min\{M,N\}} \geqslant 0$.

3. На основе окончательного разложения

$$A = Q_l \hat{Q}_l \hat{B} \hat{Q}_r Q_r \tag{1.1}$$

система $Ax = f$ приводится к двухдиагональной системе $C\hat{y} = \hat{g}$, где

$$(Q_l\hat{Q}_l)^T f = \begin{bmatrix} \hat{g} \\ g_1 \end{bmatrix} \begin{matrix} \} \ r \text{ строк}, \\ \} \ M - r \text{ строк}, \end{matrix} \qquad \hat{Q}_r Q_r x = \begin{bmatrix} \hat{y} \\ y_1 \end{bmatrix} \begin{matrix} \} \ r \text{ строк}, \\ \} \ N - r \text{ строк}. \end{matrix}$$

4. Решение x системы $Ax = f$ дается формулой

$$x = Q_r^T \hat{Q}_r^T \begin{bmatrix} C^{-1}\hat{g} \\ 0 \end{bmatrix}.$$

Матрица C не приводится к диагональному виду, поэтому в полученном «неполном» разложении явно присутствуют только наименьшие сингулярные числа $\sigma_r, \ldots, \sigma_{\min\{M,N\}}$.

В [2] показано, что приближенное r-решение $\tilde{x}$ системы $Ax = f$, вычисленное согласно описанной процедуре, удовлетворяет оценке вида

$$\|\tilde{x} - x\| \leqslant f(M, N, \varepsilon_0, d, \theta)\varepsilon_1\|x\|, \tag{1.2}$$

где x — точное r-решение, f — полиномиальная функция умеренного роста, $d = \sigma_1/(\sigma_r - \sigma_{r+1})$ характеризует «зазор» между «нулевой» и «ненулевой» частями сингулярного спектра, а $\theta = \|f - Ax\|/\|Ax\|$ — несовместность системы. Константа ε_1 (соответственно ε_0) является относительной точностью арифметики чисел с плавающей точкой (соответственно абсолютной точностью нуля; в терминологии [12] — underflow threshold). Оценка (1.2) справедлива только тогда, когда параметры обусловленности d и θ не слишком большие.

Возникает естественный вопрос: как улучшить точность приближенного r-решения $\tilde{x}$, используя уже вычисленное разложение (1.1), которое, в свою очередь, тоже приближенное. Конечно, можно вычислить $\tilde{x}$ с двойной точностью ε_1^2 и таким образом существенно увеличить точность решения. Однако это может быть очень «дорого» с точки зрения машинных ресурсов (например, если одинарная точность компьютера использует 64-битовое представление вещественных чисел). Следует использовать вычисления с двойной точностью как можно реже.

В настоящей статье предлагается вариант итерационного уточнения r-решения $\tilde{x}$, основанный на методе расширенных систем. При этом основное внимание уделяется деталям алгоритма, хотя анализ обусловленности проводится полностью. Анализ ошибок округления не приводится, для этого рекомендуются статьи [9, 10].

§ 2. Расширенные системы линейных уравнений

Пусть дана система $Ax = f$ линейных уравнений с $(M \times N)$-матрицей A неполного ранга, $\operatorname{rank} A = r$. Более точно, имеют место неравенства $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_r > \sigma_{r+1} \geqslant \ldots \geqslant \sigma_{\min\{M,N\}} \geqslant 0$, т. е. наибольшие r сингулярных чисел $\sigma_1, \sigma_2, \ldots, \sigma_r$ отделены от остальных. Запишем сингулярное разложение матрицы A в виде

$$A = [U_1 U_2] \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ & & \sigma_r \\ 0 & & & \ddots \end{bmatrix} [V_1 V_2]^T. \tag{2.1}$$

где U_1 — $(M \times r)$-матрица, V_1 — $(N \times r)$-матрица. Рассмотрим расширенную систему линейных уравнений $Hz = h$ с матрицей $H = \begin{bmatrix} \alpha I & A & 0 \\ 0 & \beta V_2^T & 0 \\ \gamma I & 0 & \gamma U_2 \end{bmatrix}$ или, в подробной записи,

$$\begin{bmatrix} \alpha I & A & 0 \\ 0 & \beta V_2^T & 0 \\ \gamma I & 0 & \gamma U_2 \end{bmatrix} \begin{bmatrix} \xi \\ x \\ \eta \end{bmatrix} = \begin{bmatrix} f \\ 0 \\ 0 \end{bmatrix} \tag{2.2}$$

с некоторыми положительными параметрами α, β, γ. В векторных базисах с $U = I$, $V = I$ система (2.2) сводится к системе

$$\begin{bmatrix} \alpha I & 0 & \Sigma_1 & 0 & 0 \\ 0 & \alpha I & 0 & \Sigma_2 & 0 \\ 0 & 0 & 0 & \beta I & 0 \\ \gamma I & 0 & 0 & 0 & 0 \\ 0 & \gamma I & 0 & 0 & \gamma I \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ x_1 \\ x_2 \\ \eta \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{2.3}$$

решение которой имеет вид

$$\xi_1 = 0, \quad \xi_2 = f_2/\alpha, \quad x_1 = \Sigma_1^{-1} f_1, \quad x_2 = 0, \quad \eta = -f_2/\alpha.$$

В первоначальных векторных базисах это означает, что x — r-решение системы $Ax = f$. Таким образом, расширенная система (2.2) эквивалентна системе $Ax = f$. Однако (2.2) является системой с квадратной матрицей H полного ранга. Следовательно, к ней можно применить стандартный процесс итерационного уточнения согласно формулам

$$z_0 = 0, \qquad z_{k+1} = z_k - \widetilde{H}^{-1}(Hz_k - h). \tag{2.4}$$

Здесь невязка $Hz_k - h$ вычисляется с высокой точностью, а $\widetilde{H}^{-1}$ является оператором приближенного решения системы $\widetilde{H}z = Hz_k - h$, которое вычислено с одинарной точностью. Хорошо известно, что скорость сходимости процесса (2.4) зависит от числа обусловленности матрицы H. Поэтому исследуем поведение числа обусловленности в зависимости от параметров α, β, γ. Для этого воспользуемся представлением (2.3).

Сингулярные числа матрицы H совпадают с сингулярными числами матриц

$$H_1 = \begin{bmatrix} \alpha & \sigma' \\ \gamma & 0 \end{bmatrix}, \qquad H_2 = \begin{bmatrix} \alpha & \sigma'' & 0 \\ 0 & \beta & 0 \\ \gamma & 0 & \gamma \end{bmatrix},$$

где σ' и σ'' «пробегают» диагонали Σ_1 и Σ_2 соответственно. Чтобы убедиться в этом, нужно в матрице системы (2.3) поменять второй и третий столбцы, а затем перенести четвертую строку в позицию после первой строки. Очевидно, что

$$\|H\| = \max\{\|H_1\|, \|H_2\|\} \leqslant \max\{\|H_1\|_F, \|H_2\|_F\}$$
$$\leqslant \max\{\sqrt{\alpha^2 + \gamma^2 + (\sigma')^2}, \sqrt{\alpha^2 + 2\gamma^2 + \beta^2 + (\sigma'')^2}\}.$$

Можно проверить, что

$$H_1^{-1} = \begin{bmatrix} 0 & \frac{1}{\gamma} \\ \frac{1}{\sigma'} & -\frac{\alpha}{\gamma\sigma'} \end{bmatrix}, \qquad H_2^{-1} = \begin{bmatrix} \frac{1}{\alpha} & -\frac{\sigma''}{\alpha\beta} & 0 \\ 0 & \frac{1}{\beta} & 0 \\ -\frac{1}{\alpha} & \frac{\sigma''}{\alpha\beta} & \frac{1}{\gamma} \end{bmatrix}.$$

Следовательно,

$$\|H^{-1}\| \leqslant \max\left\{\sqrt{\frac{1}{(\sigma')^2}+\frac{1}{\gamma^2}+\frac{\alpha^2}{\gamma^2(\sigma')^2}},\ \sqrt{\frac{2}{\alpha^2}+\frac{1}{\beta^2}+\frac{1}{\gamma^2}+\frac{2(\sigma'')^2}{\alpha^2\beta^2}}\right\},$$

$$\|H\|\,\|H^{-1}\| \leqslant \max\{\sqrt{\alpha^2+\gamma^2+\sigma_1^2},\ \sqrt{\alpha^2+\beta^2+2\gamma^2+\sigma_{r+1}^2}\}$$
$$\times \max\left\{\sqrt{\frac{1}{\sigma_r^2}+\frac{1}{\gamma^2}+\frac{\alpha^2}{\gamma^2\sigma_r^2}},\ \sqrt{\frac{2}{\alpha^2}+\frac{1}{\beta^2}+\frac{1}{\gamma^2}+\frac{2\sigma_{r+1}^2}{\alpha^2\beta^2}}\right\}.$$

Выбирая $\alpha=\beta=\gamma=\sigma_1$, получаем

$$\|H\|\,\|H^{-1}\| \leqslant \sqrt{4\sigma_1^2+\sigma_{r+1}^2}\sqrt{2/\sigma_r^2+\ 4/\sigma_1^2}$$
$$\leqslant \sqrt{8(\sigma_1/\sigma_r)^2+22} < \sqrt{8}(\sigma_1/\sigma_r+1).$$

Сделаем важное для дальнейшего изложения замечание. Результат итерационного уточнения будет тем же самым, если в системе (2.2) матрицы U_2, V_2 заменить матрицами $\hat{U}_2$, $\hat{V}_2$ при условии

$$\mathrm{Image}\,(\hat{U}_2) \approx \mathrm{Image}\,(U_2), \qquad \mathrm{Image}\,(\hat{V}_2) \approx \mathrm{Image}\,(V_2),$$
$$\text{величины } \mathrm{cond}\,(\hat{U}_2),\ \mathrm{cond}\,(\hat{V}_2) \text{ малы},$$

где $\mathrm{Image}(U_2)$ — линейная оболочка столбцов матрицы U_2. Введем обозначения $\check{U}_2=\hat{U}_2(\hat{U}_2^T\hat{U}_2)^{-1/2}$, $\check{V}_2=\hat{V}_2(\hat{V}_2^T\hat{V}_2)^{-1/2}$. Столбцы матриц $\check{U}_2$, $\check{V}_2$ образуют ортонормальные векторные системы. Более того, линейные оболочки столбцов матриц $\check{U}_2$, $\hat{U}_2$ совпадают. То же справедливо для матриц $\check{V}_2$, $\hat{V}_2$. Предположим, что $\mathrm{cond}(\hat{U}_2)\leqslant a_1$, $\mathrm{cond}(\hat{V}_2)\leqslant a_2$ с достаточно малыми a_1, a_2. Ввиду условия $\mathrm{Image}\,(\hat{U}_2)\approx\mathrm{Image}\,(U_2)$ норма $\|U_2^T\check{U}_2-I\|$ должна быть малой. Аналогично норма $\|V_2^T\check{V}_2-I\|$ также должна быть малой. Тем не менее обычно невозможно провести эффективный анализ ошибок, используя оценки норм $\|U_2^T\check{U}_2-I\|$ и $\|V_2^T\check{V}_2-I\|$. Вместо них применяют другие невязки.

Пусть $\check{U}_1$, $\check{V}_1$ — ортогональные дополнения матриц $\check{U}_2$, $\check{V}_2$, т. е. квадратные матрицы $(\check{U}_1\check{U}_2)$, $(\check{V}_1\check{V}_2)$ ортогональны. При соответствующих условиях справедливы равенства $\|\check{U}_1^TA\check{V}_2\|=\delta_1\|A\|$, $\|\check{U}_2^TA\check{V}_1\|=\delta_2\|A\|$, где δ_1, δ_2 малы. В этом случае матрица

$$\check{A}=A-[\check{U}_1\check{U}_2]\begin{bmatrix}0 & \check{U}_1^TA\check{V}_2\\ \check{U}_2^TA\check{V}_1 & 0\end{bmatrix}[\check{V}_1\check{V}_2]^T$$

отличается от матрицы $\check{A}$ по норме на величину $\max\{\delta_1,\delta_2\}\|A\|$.

Рассмотрим матрицу

$$\hat{H} = \begin{bmatrix} \alpha I & A & 0 \\ 0 & \beta \hat{V}_2^T & 0 \\ \gamma I & 0 & \gamma \hat{U}_2 \end{bmatrix}$$

$$= \begin{bmatrix} I & 0 & 0 \\ 0 & (\hat{V}_2^T \hat{V}_2)^{1/2} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \alpha I & A & 0 \\ 0 & \beta \check{V}_2 & 0 \\ \gamma I & 0 & \gamma \check{U}_2 \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & (\hat{U}_2^T \hat{U}_2)^{1/2} \end{bmatrix}.$$

Система $\hat{H}\hat{z} = (f^T\, 0\, 0)^T$ эквивалентна следующей:

$$\begin{bmatrix} \alpha I & A & 0 \\ 0 & \beta \check{V}_2^T & 0 \\ \gamma I & 0 & \gamma \check{U}_2 \end{bmatrix} \begin{bmatrix} \hat{\xi} \\ \hat{x} \\ (\hat{U}_2^T \hat{U}_2)^{1/2} \hat{\eta} \end{bmatrix} = \begin{bmatrix} f \\ 0 \\ 0 \end{bmatrix}. \tag{2.5}$$

Заменим матрицу A в (2.5) матрицей $\check{A}$. Классические результаты теории возмущений [2–4] позволяют вывести оценку

$$\left\| \begin{bmatrix} \hat{\xi} \\ \hat{x} \\ (\hat{U}_2^T \hat{U}_2)^{1/2} \hat{\eta} \end{bmatrix} - \begin{bmatrix} \check{\xi} \\ \check{x} \\ \check{\eta} \end{bmatrix} \right\| \leqslant \frac{\operatorname{cond}(\check{H})\delta}{1 - \operatorname{cond}(\check{H})\delta} \left\| \begin{bmatrix} \check{\xi} \\ \check{x} \\ \check{\eta} \end{bmatrix} \right\|, \tag{2.6}$$

где $\delta = \max\{\delta_1, \delta_2\}$,

$$\check{H} = \begin{bmatrix} \alpha I & \check{A} & 0 \\ 0 & \beta \check{V}_2^T & 0 \\ \gamma I & 0 & \gamma \check{U}_2 \end{bmatrix}, \qquad \check{H} \begin{bmatrix} \check{\xi} \\ \check{x} \\ \check{\eta} \end{bmatrix} = \begin{bmatrix} f \\ 0 \\ 0 \end{bmatrix}.$$

Заметим, что столбцы $\check{U}_2$ и $\check{V}_2$ образуют базисы точных сингулярных подпространств матрицы $\check{A}$. Следовательно, вектор $\check{x}$ является точным r-решением системы $\check{A}\check{x} = f$ при условии

$$\sigma_r(A) - \sigma_{r+1}(A) > 2\delta\|A\|, \tag{2.7}$$

которое в дальнейшем предполагаем выполненным. Неравенство (2.7) следует из оценки $|\sigma_j(A) - \sigma_j(\check{A})| \leqslant \delta\|A\|$, так как $\|A - \check{A}\| \leqslant \delta\|A\|$. Если $\delta \leqslant \min\{\sqrt{2} - 1, 1/\mu\}$, то

$$\operatorname{cond}(\check{M}) \leqslant \sqrt{4\sigma_1^2 + \sigma_r^2}\sqrt{4/\sigma_1^2 + 2/(\sigma_r - \delta\sigma_1)^2} \leqslant \sqrt{8}(\mu + 1)/(1 - \delta\mu), \tag{2.8}$$

где $\mu = \sigma_1/\sigma_r$, а σ_1, σ_r — сингулярные числа матрицы A. Справедливо неравенство (см. [11])

$$\frac{\|\check{x} - x\|}{\|x\|} \leqslant \frac{\delta\mu(1 + \theta d) + \delta d}{1 - \delta\mu - \delta d}.$$

Так как $\|\alpha\check{\xi}\| = \|\alpha\check{\eta}\| = \theta\|\check{A}\check{x}\|$, имеем

$$\begin{aligned}\|x-\hat{x}\| &\leqslant \|x-\check{x}\| + \|\check{x}-\hat{x}\| \\ &\leqslant \frac{\delta\sqrt{8}(\mu+1)/(1-\delta\mu)}{1-\delta\sqrt{8}(\mu+1)/(1-\delta\mu)} \left\| \begin{bmatrix} \check{\xi} \\ \check{x} \\ \check{\eta} \end{bmatrix} \right\| + \frac{\delta\mu(1+\theta d)+\delta d}{1-\delta\mu-\delta d}\|x\| \\ &\leqslant \frac{\sqrt{8}\delta(\mu+1)}{1-\delta(3+4\mu)}\sqrt{2\theta^2\|\check{x}\|^2+\|\check{x}\|^2} + \frac{\delta\mu(1+\theta d)+\delta d}{1-\delta\mu-\delta d}\|x\| \\ &\leqslant \left[\frac{\sqrt{8}\delta(\mu+1)}{1-\delta(3+4\mu)}\sqrt{2\theta^2+1}\Big(1+\frac{\delta\mu(1+\theta d)}{1-\delta\mu-\delta d}\Big) + \frac{\delta\mu(1+\theta d)+\delta d}{1-\delta\mu-\delta d}\right]\|x\| \\ &\leqslant \left[\frac{\sqrt{8}\delta(\mu+1)}{1-\delta(3+4\mu)}\frac{\sqrt{2\theta^2+1}[1+\delta d(\mu\theta-1)]}{1-\delta\mu-\delta d} + \frac{\delta\mu(1+\theta d)+\delta d}{1-\delta\mu-\delta d}\right]\|x\|.\end{aligned}$$

Оценка

$$\frac{\|\hat{x}-x\|}{\|x\|} \leqslant \frac{\delta(\mu+1)(4\theta+3)[1+\delta d(\mu\theta-1)]+\delta\mu(1+\theta d)+\delta d}{1-\delta(3+4\mu+d)}$$

показывает, что число δd должно быть порядка ε_1, а число θ — достаточно малым, чтобы гарантировать улучшение точности r-решения x уравнения $Ax = f$. В силу (2.8)

$$\operatorname{cond}(\hat{H}) \leqslant \operatorname{cond}(\check{H})a_1a_2 \leqslant a_1a_2\sqrt{8}(\mu+1)/(1-\delta\mu).$$

Следовательно, a_1a_2 должно быть порядка единицы, чтобы итерационное уточнение имело примерно ту же скорость сходимости, что и уточнение для системы (2.2).

§ 3. Уточнение сингулярных подпространств

Рассмотрим сначала случай полного сингулярного разложения. Пусть дано разложение (2.1), вычисленное с одинарной точностью. Требуется уточнить вычисленные матрицы $\widetilde{U}_2$ и $\widetilde{V}_2$ с тем, чтобы столбцы уточненных матриц $\hat{U}_2$, $\hat{V}_2$ давали базисы соответствующих подпространств с высокой точностью. Как уже отмечалось, столбцы матриц $\hat{U}_2$ и $\hat{V}_2$ не обязательно должны быть ортонормализованными с высокой точностью. Достаточно, например, чтобы матрицы $\hat{U}_2$, $\hat{V}_2$ удовлетворяли оценкам

$$\|\hat{U}_2^T\hat{U}_2 - I\| \leqslant 1/2, \qquad \|\hat{V}_2^T\hat{V}_2 - I\| \leqslant 1/2.$$

Сформулируем точно поставленную задачу: дано полное сингулярное разложение $A \approx \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T$ матрицы A с оценками

$$\|\widetilde{U}^TA\widetilde{V} - \widetilde{\Sigma}\| \leqslant \delta_1\|A\| + \delta_0, \quad \|\widetilde{U}^T\widetilde{U} - I\| \leqslant \delta_2, \quad \|\widetilde{V}^T\widetilde{V} - I\| \leqslant \delta_3, \quad (3.1)$$

где δ_1, δ_0, δ_2, δ_3 — некоторые малые константы, а прямоугольная матрица $\widetilde{\Sigma}$ диагональна ($\widetilde{\Sigma} = \operatorname{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_{\min\{M,N\}})$). Обычно $\delta_1 = f_1(M,N)\varepsilon_1$, $\delta_0 = f_0(M,N)\varepsilon_0$, $\delta_2 = f_2(M,N)\varepsilon_1$, $\delta_3 = f_3(M,N)\varepsilon_1$, где f_1, f_0, f_2, f_3 — полиномы малой степени с умеренными коэффициентами, ε_1 — относительная точность вычислений с числами с плавающей точкой, ε_0 — абсолютная точность нуля [12].

1. Следуя работам [9, 10], ортонормализуем матрицы $\widetilde{U}$, $\widetilde{V}$. Обычно применяемый метод — вычисление матриц $\widetilde{U}(\widetilde{U}^T\widetilde{U})^{-1/2}$, $\widetilde{V}(\widetilde{V}^T\widetilde{V})^{-1/2}$ — в данном случае не так легко осуществить. Для этого вычисления можно использовать, например, модифицированный метод Грама — Шмидта [3], что является хорошо отработанной устойчивой процедурой. В нашей ситуации воспользуемся другим методом. Именно, аппроксимируем матрицы $\widetilde{U}(\widetilde{U}^T\widetilde{U})^{-1/2}$, $\widetilde{V}(\widetilde{V}^T\widetilde{V})^{-1/2}$ матрицами

$$\bar{U} = \widetilde{U}\left[I - \tfrac{1}{2}(\widetilde{U}^T\widetilde{U} - I)\right], \quad \bar{V} = \widetilde{V}\left[I - \tfrac{1}{2}(\widetilde{V}^T\widetilde{V} - I)\right].$$

Имеют место следующие оценки [9, 10]:

$$\|\bar{U}^T\bar{U} - I\| \leqslant \delta_2^2\frac{3+\delta_2}{4}, \quad \|\bar{V}^T\bar{V} - I\| \leqslant \delta_3^2\frac{3+\delta_3}{4}.$$

Следовательно, при $\delta_2, \delta_3 < \sqrt{\varepsilon_1}$, матрицы $\bar{U}, \bar{V}$ являются хорошими уточнениями ортонормализаций $\widetilde{U}, \widetilde{V}$.

Легко получить аналог первого неравенства в (3.1) для матриц $\bar{U}, \bar{V}$:

$$\begin{aligned}
\|\bar{U}^TA\bar{V} - \widetilde{\Sigma}\| &= \|\left[I - \tfrac{1}{2}(\widetilde{U}^T\widetilde{U} - I)\right]\widetilde{U}^TA\widetilde{V}\left[I - \tfrac{1}{2}(\widetilde{V}^T\widetilde{V} - I)\right] - \widetilde{\Sigma}\| \\
&\leqslant \|\widetilde{U}^TA\widetilde{V} - \widetilde{\Sigma}\| + \|\tfrac{1}{2}(\widetilde{U}^T\widetilde{U} - I)\widetilde{U}^TA\widetilde{V}\left[I - \tfrac{1}{2}(\widetilde{V}^T\widetilde{V} - I)\right]\| \\
&\quad + \|\left[I - \tfrac{1}{2}(\widetilde{U}^T\widetilde{U} - I)\right]\widetilde{U}^TA\widetilde{V}\tfrac{1}{2}(\widetilde{V}^T\widetilde{V} - I)\| \\
&\quad + \|\tfrac{1}{2}(\widetilde{U}^T\widetilde{U} - I)\widetilde{U}^TA\widetilde{V}\tfrac{1}{2}(\widetilde{V}^T\widetilde{V} - I)\| \\
&\leqslant \delta_1\|A\| + \delta_0 + \tfrac{1}{2}\|A\|\left[\|(\widetilde{U}^T\widetilde{U} - I)\widetilde{U}^T\| + \|\widetilde{V}(\widetilde{V}^T\widetilde{V} - I)\|\right] \\
&\quad + \tfrac{1}{4}\|A\|\,\|(\widetilde{U}^T\widetilde{U} - I)\widetilde{U}^T\|\,\|\widetilde{V}(\widetilde{V}^T\widetilde{V} - I)\| \\
&\leqslant \delta_0 + \|A\|\left[\delta_1 + \tfrac{1}{2}(\delta_2\sqrt{1+\delta_2} + \delta_3\sqrt{1+\delta_3})\right. \\
&\quad \left. + \tfrac{1}{4}\delta_2\delta_3\sqrt{1+\delta_2}\sqrt{1+\delta_3}\right] \\
&\leqslant \delta_0 + \|A\|\left\{\delta_1 + \left(1 + \frac{\delta_2}{2-\delta_2}\right)\left(1 + \frac{\delta_3}{2-\delta_3}\right) - 1\right\} \\
&\leqslant \delta_0 + \|A\|\left(\delta_1 + \frac{\delta_2+\delta_3}{2-\delta_2-\delta_3}\right).
\end{aligned}$$

Таким образом,

$$\|\bar{U}^TA\bar{V} - \widetilde{\Sigma}\| \leqslant \bar{\delta}_1\|A\| + \bar{\delta}_0, \quad \|\bar{U}^T\bar{U} - I\| \leqslant \bar{\delta}_2, \quad \|\bar{V}^T\bar{V} - I\| \leqslant \bar{\delta}_3,$$

где $\bar{\delta}_1 = \delta_1 + (\delta_2+\delta_3)/(2-\delta_2-\delta_3)$, $\bar{\delta}_0 = \delta_0$, $\bar{\delta}_2 = \delta_2^2(3+\delta_2)/4$, $\bar{\delta}_3 = \delta_3^2(3+\delta_3)/4$.

2. Аппроксимируем с высокой точностью матричное произведение $\bar{U}^TA\bar{V}$, используя для этого матрицу

$$A_1 = \widetilde{\Sigma} + \tfrac{1}{2}\widetilde{U}^T(A\widetilde{V} - \widetilde{U}\widetilde{\Sigma}) + \tfrac{1}{2}(\widetilde{U}^TA - \widetilde{\Sigma}\widetilde{V}^T)\widetilde{V}. \tag{3.2}$$

Так как

$$\begin{aligned}
\bar{U}^TA\bar{V} &= \left[I - \tfrac{1}{2}(\widetilde{U}^T\widetilde{U} - I)\right]\widetilde{U}^TA\widetilde{V}\left[I - \tfrac{1}{2}(\widetilde{V}^T\widetilde{V} - I)\right] \\
&= \widetilde{U}^TA\widetilde{V} + \tfrac{1}{2}(I - \widetilde{U}^T\widetilde{U})\widetilde{U}^TA\widetilde{V} + \tfrac{1}{2}\widetilde{U}^TA\widetilde{V}(I - \widetilde{V}^T\widetilde{V}) \\
&\quad + \tfrac{1}{4}(I - \widetilde{U}^T\widetilde{U})\widetilde{U}^TA\widetilde{V}(I - \widetilde{V}^T\widetilde{V}), \\
A_1 &= \widetilde{U}^TA\widetilde{V}\tfrac{1}{2}(I - \widetilde{U}^T\widetilde{U})\widetilde{\Sigma} + \tfrac{1}{2}\widetilde{\Sigma}(I - \widetilde{V}^T\widetilde{V}),
\end{aligned}$$

справедливо равенство

$$\bar{U}^T A\bar{V} - A_1 = \tfrac{1}{2}(I - \tilde{U}^T\tilde{U})(\tilde{U}^T A\tilde{V} - \tilde{\Sigma}) + \tfrac{1}{2}(\tilde{U}^T A\tilde{V} - \tilde{\Sigma})(I - \tilde{V}^T\tilde{V}) + \tfrac{1}{4}(I - \tilde{U}^T\tilde{U})\tilde{U}^T A\tilde{V}(I - \tilde{V}^T\tilde{V}),$$

из которого получаем оценку аппроксимации $\tilde{U}^T A\bar{V}$ матрицы A_1:

$$\|\bar{U}^T A\bar{V} - A_1\| \leqslant \tfrac{1}{2}(\delta_2 + \delta_3)(\delta_1\|A\| + \delta_0) + \|A\|\frac{\delta_2}{2-\delta_2}\frac{\delta_3}{2-\delta_3} \leqslant \frac{\delta_0}{2}(\delta_2 + \delta_3) + \frac{\|A\|}{2}\Big[\delta_1(\delta_2 + \delta_3) + \frac{\delta_3\delta_3}{2 - \delta_2 - \delta_3}\Big].$$

Укажем два способа использования матрицы A_1. При первом способе следует вычислить матрицу A_1 с высокой точностью, соблюдая порядок, указанный скобками в (3.2). На самом деле только матричные произведения $A\tilde{V}$ и $\tilde{U}^T A$ должны вычисляться с двойной точностью. Произведения $\tilde{U}\tilde{\Sigma}$, $\tilde{\Sigma}\tilde{V}^T$ также нужно вычислить с двойной точностью, но это «не дорогие» операции ввиду диагональности матрицы $\tilde{\Sigma}$. Произведения $\tilde{U}^T \times (A\tilde{V} - \tilde{U}\tilde{\Sigma})$, $(\tilde{U}^T A - \tilde{\Sigma}\tilde{V}^T) \times \tilde{V}$ можно вычислять с одинарной точностью, так как выражения в скобках достаточно малы. Таким образом, «цена вычисления» A_1 равна двум матричным произведениям с двойной точностью. При втором способе при дальнейших вычислениях используется явное представление (3.2) матрицы A_1. Первый способ удобен, когда число столбцов матриц U_2, V_2 не мало. Если требуется улучшить только сингулярные векторы, второй способ выглядит более привлекательным.

Отметим, что A_1 является возмущением диагональной матрицы $\tilde{\Sigma}$.

3. Теперь остается вычислить с высокой точностью базисы сингулярных подпространств, которые соответствуют малым сингулярным числам $\sigma_{r+1}, \ldots, \sigma_{\min\{M,N\}}$, при условии, что число $\tau = (\tilde{\sigma}_r - \tilde{\sigma}_{r+1})/\tilde{\sigma}_1$ не очень мало. Пусть $A_1 = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ — блочная матрица, где A_{11} — $(r \times r)$-матрица, отвечающая сингулярным числам $\tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_r$ матрицы $\tilde{\Sigma}$, т. е., A_{11} — возмущение диагональной матрицы $\mathrm{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_r)$. Будем искать требуемые подпространства в следующем виде. Пусть правое сингулярное подпространство порождается столбцами $\begin{bmatrix} R \\ I \end{bmatrix}$, а левое — столбцами $\begin{bmatrix} L \\ I \end{bmatrix}$. Тогда $(r \times (N-r))$-матрица R и $(r \times (M-r))$-матрица L должны удовлетворять уравнению

$$\begin{bmatrix} I & -L \\ L^T & I \end{bmatrix}\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} I & R \\ -R^T & I \end{bmatrix} = \begin{bmatrix} A'_{11} & 0 \\ 0 & A'_{22} \end{bmatrix} \qquad (3.3)$$

с подходящими матрицами A'_{11}, A'_{22}, где A'_{11} — квадратная матрица по-

рядка r. В этом случае матрицы

$$V_1 = \begin{bmatrix} I & R \\ -R^T & I \end{bmatrix} \begin{bmatrix} (I+RR^T)^{-1/2} & 0 \\ 0 & (I+R^TR)^{-1/2} \end{bmatrix},$$

$$U_1 = \begin{bmatrix} I & -L \\ L^T & I \end{bmatrix} \begin{bmatrix} (I+LL^T)^{-1/2} & 0 \\ 0 & (I+L^TL)^{-1/2} \end{bmatrix}$$

ортогональные, а матрица $U_1^T A V_1$ имеет вид

$$U_1^T A V_1 = \begin{bmatrix} (I+LL^T)^{-1/2} A'_{11} (I+RR^T)^{-1/2} & 0 \\ 0 & (I+L^TL)^{-1/2} A'_{22} (I+R^TR)^{-1/2} \end{bmatrix}.$$

Очевидно, что столбцы матриц $\begin{bmatrix} R \\ I \end{bmatrix}(I+R^TR)^{-1/2}$, $\begin{bmatrix} L \\ I \end{bmatrix}(I+L^TL)^{-1/2}$ образуют ортонормальные базисы искомых сингулярных подпространств, отвечающих малым сингулярным числам.

В силу (3.3) матрицы L, R должны удовлетворять системе матричных уравнений

$$(L^T I)\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} I \\ -R^T \end{bmatrix} = 0, \qquad (I-L)\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} R \\ I \end{bmatrix} = 0. \quad (3.4)$$

Поскольку $A_{11} = \Lambda_1 + \Delta_{11}$, $A_{22} = \Lambda_2 + \Delta_{22}$, где матрицы Λ_1, Λ_2 диагональные, а нормы матриц Δ_{11}, Δ_{22} малы, систему (3.4) можно преобразовать следующим образом:

$$(L^T I)\begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}\begin{bmatrix} I \\ -R^T \end{bmatrix} = -(L^T I)\begin{bmatrix} \Delta_{11} & A_{12} \\ A_{21} & \Delta_{22} \end{bmatrix}\begin{bmatrix} I \\ -R^T \end{bmatrix},$$

$$(I-L)\begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}\begin{bmatrix} R \\ I \end{bmatrix} = -(I-L)\begin{bmatrix} \Delta_{11} & A_{12} \\ A_{21} & \Delta_{22} \end{bmatrix}\begin{bmatrix} R \\ I \end{bmatrix}.$$

Так как

$$(L^T I)\begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}\begin{bmatrix} I \\ -R^T \end{bmatrix} = L^T\Lambda_1 - \Lambda_2 R^T,$$

$$(I-L)\begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} = \Lambda_1 R - L\Lambda_2,$$

имеем

$$\Lambda_1 R - L\Lambda_2 = -F_1, \qquad \Lambda_1^T L - R\Lambda_2^T = -F_2, \quad (3.5)$$

где

$$F_1(L,R) = (I-L)\begin{bmatrix} \Delta_{11} & A_{12} \\ A_{21} & \Delta_{22} \end{bmatrix}\begin{bmatrix} R \\ I \end{bmatrix}, \quad (3.6)$$

$$F_2(L,R) = (L^T I)\begin{bmatrix} \Delta_{11} & A_{12} \\ A_{21} & \Delta_{22} \end{bmatrix}\begin{bmatrix} I \\ -R^T \end{bmatrix}. \quad (3.7)$$

- Система матричных уравнений (3.5) с фиксированными F_1, F_2, т. е. когда F_1, F_2 не зависят от L, R, называется *обобщенным уравнением Сильвестра.*

Переписав обобщенное уравнение Сильвестра (3.5) в виде

$$\begin{bmatrix} 0 & \Lambda_1 \\ \Lambda_1^T & 0 \end{bmatrix} \begin{bmatrix} L & 0 \\ 0 & R \end{bmatrix} - \begin{bmatrix} L & 0 \\ 0 & R \end{bmatrix} \begin{bmatrix} 0 & \Lambda_2 \\ \Lambda_2^T & 0 \end{bmatrix} = - \begin{bmatrix} F_1 \\ F_2 \end{bmatrix},$$

получим обычное уравнение Сильвестра, которое имеет единственное решение тогда и только тогда, когда у матриц Λ_1, Λ_2 нет кратных сингулярных значений.

Так как матрицы Λ_1, Λ_2 диагональные, формулы для решения (L, R) системы (3.5) достаточно просты. Например, пусть $M \geqslant N$ и

$$\Lambda_1 = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ & & \lambda_r \\ 0 & & & \ddots \end{bmatrix}, \qquad \Lambda_2 = \begin{bmatrix} \mu_1 & & 0 \\ & \ddots & \\ & & \mu_r \\ 0 & & & \ddots \end{bmatrix}.$$

Тогда

$$\begin{aligned} &\lambda_i r_{ij} - l_{ij}\mu_j = -(F_1)_{ij}, \\ &\lambda_i l_{ij} - r_{ij}\mu_j = -(F_2)_{ij}, \end{aligned} \qquad 1 \leqslant i \leqslant r, \quad 1 \leqslant j \leqslant N - r,$$

$$\lambda_i l_{ij} = -(F_2)_{ij}, \quad 1 \leqslant i \leqslant r, \quad N - r + 1 \leqslant j \leqslant M - r,$$

где r_{ij} и l_{ij} — элементы неизвестных матриц R и L соответственно. Следовательно, решение $r_{ij} l_{ij}$ системы (3.5) можно выписать явно в виде

$$\begin{bmatrix} r_{ij} \\ l_{ij} \end{bmatrix} = \frac{1}{\lambda_i^2 - \mu_j^2} \begin{bmatrix} \lambda_i & \mu_j \\ \mu_j & \lambda_i \end{bmatrix} \begin{bmatrix} -(F_1)_{ij} \\ -(F_2)_{ij} \end{bmatrix}, \quad 1 \leqslant i \leqslant r, \quad 1 \leqslant j \leqslant N - r, \tag{3.8}$$

$$l_{ij} = \frac{1}{\lambda_i}[-(F_2)_{ij}], \qquad 1 \leqslant i \leqslant r, \qquad N - r + 1 \leqslant j \leqslant M - r. \tag{3.9}$$

Норма линейного оператора (оператора решения) в формулах (3.8), (3.9) ограничена сверху величиной $\max\{1/(\lambda_i - \mu_j), 1/\lambda_k\} = 1/(\tilde{\sigma}_r - \tilde{\sigma}_{r+1})$. Случай $M < N$ разбирается аналогично.

- Система (3.5) с правыми частями (3.6), (3.7) называется *обобщенным матричным уравнением Риккати.*

Чтобы решить систему (3.5) с высокой точностью, используется итерационный метод:

$$\begin{aligned} \Lambda_1 R^{(i+1)} - L^{(i+1)}\Lambda_2 &= -F_1(L^{(i)}, R^{(i)}), \\ \Lambda_1^T L^{(i+1)} - R^{(i+1)}\Lambda_2^T &= -F_2(L^{(i)}, R^{(i)}), \end{aligned} \tag{3.10}$$

где правые части F_1, F_2 вычисляются с высокой точностью. Поясним, как вычислить F_1, F_2 с высокой точностью. Выше уже обсуждалось, что матрица $\begin{bmatrix} \Delta_{11} & A_{12} \\ A_{21} & \Delta_{22} \end{bmatrix}$ аппроксимируется с повышенной точностью матрицей

$$A_2 = \tfrac{1}{2}\tilde{U}^T(A\tilde{V} - \tilde{U}\tilde{\Sigma}) + \tfrac{1}{2}(\tilde{U}^T A - \tilde{\Sigma}\tilde{V}^T)\tilde{V}, \tag{3.11}$$

если δ_1, δ_0, δ_2, δ_3 в (3.1) достаточно малы, например такие, что

$$\tfrac{1}{2}\delta_0(\delta_2+\delta_3)+\tfrac{1}{2}[\delta_1(\delta_2+\delta_3)+(\delta_2\delta_3)/(2-\delta_2-\delta_3)]\|A\| \leqslant \varepsilon_1\|A\|. \quad (3.12)$$

Напомним, что F_1, F_2 даются формулами

$$F_1(L,R)=(I-L)A_2\begin{bmatrix}R\\ I\end{bmatrix}, \qquad F_2(L,R)=(L^T I)A_2\begin{bmatrix}I\\ R^T\end{bmatrix}. \quad (3.13)$$

Как отмечалось выше, можно сначала вычислить с высокой точностью (используя двойную точность подходящим образом!) матрицу A_2, а затем сохранить ее для последующего использования в итерационном процессе (3.10). Функции F_1 и F_2 вычисляются с использованием A_2 и, конечно, с учетом структуры (3.13) с единичными матрицами.

Другой способ состоит в том, чтобы вычислять правые части (3.10) по формулам (3.13), используя всякий раз представление (3.11) матрицы A_2. По мнению автора, такой способ эффективен когда либо $M-r$ и $N-r$ малы, либо r мало. В этих случаях вместо матричных умножений с двойной точностью можно обходиться только операциями умножения матриц на «почти векторы», т. е. на матрицы с малым числом столбцов или строк. Например, если $N-r=1$, то функция F_1 может быть вычислена как результат следующей цепочки матричных и векторных операций:

$$w_1=\widetilde{V}\begin{bmatrix}R\\ I\end{bmatrix}, \quad w_2=Aw_1, \quad w_3=\widetilde{\Sigma}\begin{bmatrix}R\\ I\end{bmatrix}, \quad w_4=\widetilde{U}w_3,$$
$$w_5=w_2-w_4, \quad w_6=\widetilde{U}^T w_5, \quad w_7=\widetilde{U}^T w_2, \quad w_8=\widetilde{V}^T w_1,$$
$$w_9=\widetilde{\Sigma}w_8, \quad w_{10}=w_7-w_9, \quad w_{11}=(w_6+w_{10})/2, \quad w_{12}=(I-L)w_{10}.$$

Величины $w_1,\dots,w_{10}$ следует вычислять с двойной точностью, при этом только w_1, w_2, w_4, w_7, w_8 являются наиболее «дорогими» для вычислений. Если дополнительно норма $\|R\|$ достаточно мала, то вычисление w_1 и w_4 «дешевое».

Аккуратный анализ ошибок округления проводился для некоторых вариантов вычисления F_1 и F_2 в [9, 10]. Другие варианты могут быть проанализированы по аналогии с [9, 10].

Сходимость метода вычисления по формулам (3.10) и оценки вычисляемых решений доказаны, по существу, в [9, 10] и здесь не приводятся. Тем не менее, чтобы дать некоторое представление читателю о характере этих результатов, одна из финальных оценок, полученных в [9, 10], будет выписана ниже.

При помощи итерационного процесса (3.10) можно получить следующую оценку δ (см. (2.6)), при выводе которой учитывались ошибки округления во время вычисления $[\widetilde{U}(\widetilde{U}^T\widetilde{U})^{-1/2}]^T A[\widetilde{V}(\widetilde{V}^T\widetilde{V})^{-1/2}]$:

$$\delta \leqslant \frac{c_1}{\|A\|}\Big(1+\tau\|A\|\frac{c_2+c_3}{1-c_2-c_3}\Big), \quad (3.14)$$

где c_1, c_2, c_3 — некоторые положительные константы, $\tau=(\tilde{\sigma}_r-\tilde{\sigma}_{r+1})/\tilde{\sigma}_1$. Эта оценка справедлива при определенных не очень ограничительных условиях. Так как точные выражения констант c_1, c_2, c_3 громоздки, мы приведем их приближенные значения:

$$c_1=O(\varepsilon_1\|A_2\|), \qquad c_2=O\Big(\frac{\|A_2\|}{\tau\|A\|}\Big), \qquad c_3=O\Big(\frac{\|A_2\|}{\tau\|A\|}\Big),$$

где коэффициентов, как правило, меньше десяти. При выводе оценки (3.14) величина $\max\{M, N\}$ считается малой. Таким образом, в силу (3.12) и (3.14)

$$\delta = O(1)\varepsilon_1\|A_2\|/\|A\|, \tag{3.15}$$

где $\|A_2\| \ll \tau\|A\|$. Напомним, что обычно $\|A_2\| = O(N^{5/2})\varepsilon_1\|A\|$. Поэтому в (2.6) $\operatorname{cond}(\tilde{M})\delta = O(N^{5/2})\varepsilon_1^2\sigma_1/\sigma_r$.

§ 4. Модификация для случая с неполным сингулярным разложением

До сих пор матрица Λ_1 предполагалась диагональной. Но на практике матрица Λ_1 обычно двухдиагональная. Например, если $M \geqslant N$, то Λ_1 должна быть верхней двухдиагональной. Эти обстоятельства не создают больших препятствий, хотя процедура решения обобщенного уравнения Сильвестра становится немного сложнее и, что более важно, аналогичные (3.15) оценки ошибок округления ухудшаются в $2/\tau$ раз.

Итак, требуется разрешить систему линейных уравнений (3.5), где Λ_1 — двухдиагональная $(r \times r)$-матрица с сингулярными числами $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_r$. При этом матрица Λ_1, в общем случае прямоугольная, является диагональной с сингулярными числами $\sigma_r \geqslant \sigma_{r+1} \geqslant \ldots \geqslant \sigma_{\min\{M,N\}}$. Известно, что величина $\sigma_1/(\sigma_r - \sigma_{r+1})$ не слишком велика.

Элементы матриц R и L обозначим через r_{ij} и l_{ij} соответственно. Введем обозначения

$$\Lambda_1 = \begin{bmatrix} d_1 & b_2 & & & \\ & d_2 & b_3 & & 0 \\ & & \ddots & \ddots & \\ & 0 & & d_{r-1} & b_r \\ & & & & d_r \end{bmatrix}, \qquad \begin{gathered} b_1 = b_{r+1} = 0, \\ \mu_1 = \sigma_{r+1}, \ldots, \mu_{N-r} = \sigma_{\min\{M,N\}}. \end{gathered}$$

При $M \geqslant N$ получаем системы

$$\begin{aligned} &d_i r_{ij} + b_{i+1} r_{i+1,j} - l_{ij}\mu_j = -(F_1)_{ij}, \\ &d_i l_{ij} + b_i l_{i-1,j} - r_{ij}\mu_j = -(F_2)_{ij}, \end{aligned} \qquad 1 \leqslant i \leqslant r, \quad 1 \leqslant j \leqslant N - r, \tag{4.1}$$

$$d_i l_{ij} + b_i l_{i-1,j} = -(F_2)_{ij}, \quad 1 \leqslant i \leqslant r, \quad N - r + 1 \leqslant j \leqslant M - r. \tag{4.2}$$

Систему (4.2) решить несложно, так как при фиксированном j она является системой линейных уравнений с двухдиагональной матрицей.

Рассмотрим систему (4.1) при фиксированном j. Введем вектор неизвестных $Z = (r_{1j}, l_{1j}, r_{2j}, l_{2j}, \ldots, r_{ij}, l_{ij}, \ldots)^T$ с перемежающимися компонентами r_{ij}, l_{ij} и вектор правых частей $G = -[(F_2)_{1j}, (F_1)_{1j}, \ldots]^T$. Получим систему линейных уравнений $SZ = G$ с симметричной трехдиагональной матрицей

$$S = \begin{bmatrix} -\mu_j & d_1 & & & \\ d_1 & -\mu_j & b_2 & & 0 \\ & \ddots & \ddots & \ddots & \\ & 0 & b_{r-1} & -\mu_j & d_r \\ & & & d_r & -\mu_j \end{bmatrix},$$

где поддиагональными элементами будут d_1, b_2, d_2, b_3, ..., d_{i-1}, b_i, d_i, b_{i+1}, d_r. Переставляя подходящим образом строки и столбцы матрицы S, можно получить матрицу $\begin{bmatrix} -\mu_j I & \Lambda_1 \\ \Lambda_1^T & -\mu_j I \end{bmatrix}$. Следовательно, $\sigma_1 - \mu_j \geqslant \sigma_2 - \mu_j \geqslant \ldots \geqslant \sigma_r - \mu_j > -\sigma_r - \mu_j \geqslant -\sigma_{r-1} - \mu_j \geqslant \ldots \geqslant -\sigma_1 - \mu_j$ — собственные числа матрицы S, а число обусловленности матрицы S ограничено сверху величиной

$$\frac{\sigma_1 + \mu_j}{\sigma_r - \mu_j} \leqslant \frac{\sigma_1 + \sigma_{r+1}}{\sigma_r - \sigma_{r+1}} < 2\frac{\sigma_1}{\sigma_r - \sigma_{r+1}}.$$

Система $SZ = G$ легко решается ортогональным методом. Вначале к матрице S слева применяется одна цепочка вращений Якоби — Гивенса, для того чтобы получить верхнюю треугольную ленточную матрицу с шириной ленты, равной двум. Это стандартный метод преобразования гессенберговой матрицы к верхней треугольной форме вращениями с одной стороны. После этого ленточная треугольная система решается обычным последовательным исключением. Такая процедура не требует специального анализа ошибок округления, так как является частным случаем метода QR-разложения для решения линейных систем.

ЛИТЕРАТУРА

1. **Лоусон Ч., Хенсон Р.** Численное решение задач метода наименьших квадратов. М.: Наука, 1986.
2. **Годунов С. К., Антонов А. Г., Кирилюк О. П., Костин В. И.** Гарантированная точность решения систем линейных уравнений в евклидовых пространствах. Новосибирск: Наука, 1988.
3. **Golub G., Van Loan Ch.** Matrix computations. Baltimore, Maryland: John Hopkins University Press, 1989.
4. **Stewart G. W., Sun Ji-guang.** Matrix perturbation theory. San-Diego, California: Acad. Press, 1990.
5. **Björck A.** Iterative refinement of linear least squares solutions. I // BIT. 1967. V. 7. P. 257–278.
6. **Björck A.** Iterative refinement of linear least squares solutions. II // BIT. 1968. V. 8. P. 8–30.
7. **Björck A., Golub G.** Iterative refinement of linear least squares solutions by Householder transformations // BIT. 1967. V. 7. P. 322–337.
8. **Кирилюк О. П.** Итерационное уточнение решения систем линейных уравнений // Тр. Ин-та математики / АН СССР. Сиб. отд-ние. Новосибирск: Наука, 1989. Т. 17: Вычислительные методы линейной алгебры. С. 5–18.
9. **Malyshev A. N.** On iterative refinement for the spectral decompositions of symmetric matrices // East-West J. of Numer. Math. 1992. V. 1. P. 27–51.
10. **Malyshev A. N.** Raffinement itérative d'une décomposition spectrale de matrice symétrique. // IRISA–INRIA Publication interne. 1992. N 628.
11. **Малышев А. Н.** Введение в вычислительную линейную алгебру. Новосибирск: Наука, 1991.
12. **Demmel J.** Underflow and the reliability of numerical software // SIAM J. Sci. Statist. Comput. 1984. V. 5. P. 887–919.