А.И. Жданов

ВВЕДЕНИЕ В МЕТОДЫ РЕШЕНИЯ НЕКОРРЕКТНЫХ ЗАДАЧ

УДК 519.6(075)+512.64(075)
ББК 22.19+22.143
Ж 422

Жданов А.И.
Введение в методы решения некорректных задач: Учеб. пособие. – Изд-во Самарского гос. аэрокосмического ун-та, 2006. – 87 с.

ISBN 5-7883-0472-6

Рассмотрены основные понятия теории некорректных задач. В основном изложение ведется для случая конечномерных задач. Это дает возможность получить важнейшие начальные понятия о неустойчивости вычислительных задач, а также научиться решать наиболее практически важные некорректные задачи на компьютере. Для решения неустойчивых вычислительных задач данного класса рассмотрены современные наиболее эффективные вычислительные алгоритмы. Эти алгоритмы являются универсальными и могут быть использованы для решения широкого класса неустойчивых конечномерных линейных задач.

Предназначено для студентов обучающихся по специальностям "Прикладная математика и информатика", "Прикладные математика и физика"и др., а также для специалистов, применяющих в своей деятельности идеи и методы решения некорректных задач на компьютерах.

Оглавление

Предисловие 5

1 Вспомогательные сведения **7**
1.1 Арифметические пространства 7
1.2 Матричная алгебра 11
1.3 Нормы векторов и матриц 17
1.4 Сингулярное разложение матриц 22

2 Нормальные решения и псевдорешения **27**
2.1 Псевдорешения линейных систем 27
2.2 Линейная задача наименьших квадратов 30
2.3 Псевдообращение 32
2.4 Вычисление псевдообратных матриц 40
2.5 Типовые примеры 42

3 Вычисление псевдорешений **45**
3.1 Определение множества чисел с плавающей точкой 45
3.2 Обусловленность и числа обусловленности 46
3.2.1 Обусловленность задачи 47
3.2.2 Абсолютное число обусловленности 47
3.2.3 Относительное число обусловленности 48
3.2.4 Обусловленность матрично-векторного умножения 49
3.2.5 Число обусловленности матрицы 51
3.3 Теория возмущений 52
3.3.1 Системы с квадратными невырожденными матрицами 52
3.3.2 Теория относительных возмущений 56
3.3.3 Теория возмущений для задачи наименьших квадратов 60
3.4 Вычисление нормальных псевдорешений линейных систем 62

4 Вычисление решений приближенных систем **65**
4.1 Корректность вычислительной задачи 65
4.2 Постановка задачи 67
4.3 Регуляризация систем на основе расширенных систем . . 69
4.4 Обусловленность вычислительной задачи 71
4.5 Метод мнимого сдвига спектра 73
4.6 Численные примеры 74

5 Идентификация нестационарных AR-моделей **77**
5.1 Стохастические непрерывные модели 77
5.2 Стохастические дискретные модели 80
5.3 Постановка задачи идентификации 82
Литература 85

Предисловие

На простейших примерах в учебном пособии подробно исследуется феномен неустойчивости вычислительных задач и даются основные подходы по преодолению численных "катастроф". При решении некоторых математических задач встречаются ситуации, когда казалось бы, незначительные возмущения (погрешности) исходных данных (например, при записи их в компьютер) влекут катастрофические последствия в результате. Это явление связано с неустойчивостью (некорректностью) задачи – малые изменения данных задачи вызывают большие изменения в решении. Для так называемых хорошо поставленных проблем традиционные вычислительные методы работают вполне надежно, однако их использование для решения неустойчивых (некорректно поставленных) задач не безопасно.

Основная задача, которая рассматривается в данном учебном пособии, – это решение приближенных *систем линейных алгебраических уравнений* (СЛАУ).

Наличие неизбежных погрешностей (неточностей) в задании коэффициентов как в правой так и в левой (матричном операторе) её частях, порождённых конечной точностью представления чисел в ЭВМ приводит к неопределённости искомого решения.

Как было указано А.Н. Тихоновым, при построении решения СЛАУ принципиальным фактором является наличие погрешности задания правой части и матрицы. Классические алгоритмы решения СЛАУ, основанные на концепции абсолютной точности, при наличии погрешностей не могут быть положены в основу универсальных вычислительных программ для ЭВМ в силу неустойчивости к погрешностям.

В учебном пособии рассматриваются два близких, но в тоже время различных, класса проблем: плохо обусловленные и некорректные задачи, которые с точки зрения применимости численных методов, являются аномальными. Любой "нормальный" метод (алгоритм), предназначенный для решения "нормальной" (корректной) задачи, как правило, для них работать не будет: либо результат будет неправильным, либо произойдет останов ЭВМ в случае, когда выполнить требуемое программой действие невозможно (разделить на 0 и т.п.).

Подобные задачи возникают во многих областях науки и техники: геофизике, радиоастрономии, спектроскопии, экономике, медицинской и технической диагностике, обработке и интерпретации данных физических экспериментов. Тем не менее долгое время считалось, что подоб-

ного рода задачи не могут описывать реальных явлений и находятся вне рамок вычислительной математики. Приводимая при этом аргументация сводилась к указанию на их аномальные свойства, что якобы делает невозможным их решение. Дело в том, что для некорректных задач малые возмущения в исходных данных могут приводить к сколь угодно большим изменениям решения (неустойчивость задачи). Так как при действии с вещественными числами избежать внесения возмущений нельзя (из-за операции округления), то полученное решение практически бесполезно, поскольку нет гарантии его близости к искомому решению. Следовательно, в рамках традиционной концепции приближенного решения, отвечающего приближенным данным, некорректные задачи не могут быть решены.

Благодаря основополагающим работам А.Н. Тихонова, В.К. Иванова, М.М. Лаврентьева и их последователей (особенно следует отметить работы В.А. Морозова) разработаны регулярные (устойчивые) методы решения некорректно поставленных задач. Таким образом, они получили статус "законных" задач. Сама природа не позволяет сделать их решение чисто технической процедурой. Тем не менее возможность получения достоверных результатов при наличии неустойчивости сейчас никем не оспаривается.

Центральным понятием, объединяющим все методы решения некорректных задач, является понятие регуляризирующего алгоритма [17]. На первый взгляд оно достаточно парадоксально: прежде чем решать задачу, нужно должным образом "исказить" исходное уравнение, выбрав специальным образом уровень "искажения". Дело в том, что регуляризованная задача с точки зрения численного решения оказывается существенно лучше: она корректна, устойчива.

В учебном пособии использован ряд методических приемов из учебников, написанных А.Н. Малышевым [14], Д.В. Беклемишевым [3], Ф.Р. Гантмахером [6], Г.С. Шевцовым [20], Дж. Деммелем [9], Дж. Голубом и Ч. Ван Лоуном [8], Л.Н. Трефезеном и Д. Бау [24].

Автор благодарит свою коллегу по кафедре прикладной математики СГАУ доц. С.Ю. Гоголеву, оказавшей помощь в подготовке книги к изданию.

Глава 1

Вспомогательные сведения из линейной алгебры

1.1. Арифметические пространства

В вычислительных методах линейной алгебры под *линейным (векторным) пространством над полем вещественных чисел* понимается *арифметическое n-мерное линейное пространство* $\mathbb{R}^n$, элементами которого являются вектор-столбцы вида $x = (x_1, \ldots, x_n)^\top$, составленные из фиксированного числа n вещественных компонент (координат) $x_i \in \mathbb{R}$, $1 \le i \le n$, где $\top$ – знак транспонирования. Аналогично под *линейным (векторным) пространством над полем комплексных чисел* понимается *арифметическое n-мерное линейное пространство* $\mathbb{C}^n$, элементами которого являются вектор-столбцы вида $x = (x_1, \ldots, x_n)^\top$, составленные из фиксированного числа n комплексных компонент (координат) $x_i \in \mathbb{C}$, $1 \le i \le n$.

Если $x = (x_1, \ldots, x_n)^\top$ и $y = (y_1, \ldots, y_n)^\top$ – элементы $\mathbb{R}^n$ (или $\mathbb{C}^n$), а $\alpha \in \mathbb{R}$ (или $\alpha \in \mathbb{C}$), то сумма $x + y = (x_1 + y_1, \ldots, x_n + y_n)^\top$ и произведение на скаляр $\alpha x = (\alpha x_1, \ldots, \alpha x_n)^\top$ принадлежит пространству $\mathbb{R}^n(\mathbb{C}^n)$. Условия замкнутости относительно сложения векторов и умножения на скаляр характеризуют абстрактное понятие линейного пространства. При этом в линейном пространстве выполняется система аксиом:

$$x + y = y + x, \quad x + (y + z) = (x + y) + z,$$

$$x + 0 = x, \quad x + (-x) = 0,$$

$$\alpha(x + y) = \alpha x + \alpha y, \quad (\alpha + \beta)x = \alpha x + \beta x, \quad 1 \cdot x = x,$$

где $x, y, z \in \mathbb{R}^n\,(\mathbb{C}^n)$, $0 = (0, \ldots, 0)^\top$ – нулевой вектор пространства $\mathbb{R}^n\,(\mathbb{C}^n)$, $\alpha, \beta \in \mathbb{R}\,(\mathbb{C})$ и 1 – вещественные скаляры, $-x = (-1)x$. Как правило, вместо $x + (-y)$ пишут $x - y$. Символ 0, как обычно, будет обозначать нулевой скаляр, нулевой вектор или нулевую матрицу, т.е. структуры, состоящие только из нулевых элементов; тип структуры и ее размеры, если не указаны явно, определяются контекстом.

Аналогичным образом можно рассмотреть пространство $\mathbb{R}^n$ (или $\mathbb{C}^n$), образованное вектор-столбцами. В дальнейшем, если не оговорено иное, предполагается, что линейное пространство $\mathbb{R}^n$ (или $\mathbb{C}^n$) образовано вектор-столбцами.

Введем теперь важное понятие подпространства линейного пространства $\mathbb{R}^n$. В дальнейшем в этом разделе будет рассматриваться лишь пространство $\mathbb{R}^n$, хотя все результаты распространяются и на $\mathbb{C}^n$. Множество $\mathcal{L}$ векторов из $\mathbb{R}^n$ называется *подпространством* пространства $\mathbb{R}^n$, если оно замкнуто относительно сложения и умножения вектора на число (соответственно из $\mathbb{R}$ или $\mathbb{C}$), т.е. если

1° вместе с каждым x множество $\mathcal{L}$ содержит и все векторы вида αx $(\alpha x \in \mathcal{L})$,

2° вместе с векторами x, y множество $\mathcal{L}$ содержит их сумму $x + y$ $(x + y \in \mathcal{L})$.

Подпространство $\mathcal{L}$ может состоять лишь из одного вектора - нулевого. Такое подпространство назовем *тривиальным*. Нетривиальное подпространство, не совпадающее с $\mathbb{R}^n$, называется *собственным*. Например, все векторы из $\mathbb{R}^3$ вида $(x_1, x_2, x_3) = (\xi + 2\eta, 2\xi + \eta, 3\xi + 3\eta)$ образуют собственное подпространство $\mathbb{R}^3$.

Введем понятие базиса подпространства $\mathcal{L}$ пространства $\mathbb{R}^n$.

Линейной комбинацией векторов $x^{(1)}, x^{(2)}, \ldots, x^{(k)} \in \mathcal{L} \subset \mathbb{R}^n$ *с коэффициентами* α_k назовем вектор вида

$$x = \alpha_1 x^{(1)} + \alpha_2 x^{(2)} + \ldots + \alpha_k x^{(k)} = \sum_{j=1}^{k} \alpha_j x^{(j)}.$$

Система векторов $\{x^{(1)}, x^{(2)}, \ldots, x^{(k)}\}$ называется *линейно зависимой*, если можно подобрать такие коэффициенты $\alpha_1, \alpha_2, \ldots, \alpha_k$, среди которых хотя бы один отличен от нуля, что линейная комбинация $x = \sum_{j=1}^{k} \alpha_j x^{(j)}$ является нулевым вектором. Бесконечная система векторов из $\mathbb{R}^n$ называется *линейно зависимой*, если линейно зависима одна из ее конечных подсистем.

Система векторов $\{x^{(1)}, x^{(2)}, \ldots, x^{(k)}\}$ *линейно независима*, если равенство $\sum_{j=1}^{k} \alpha_j x^{(j)} = 0$ выполняется лишь при $\alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$.

П р и м е р. Векторы из $\mathbb{R}^3$

$$x^{(1)} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad x^{(2)} = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}, \quad x^{(3)} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$$

линейно зависимы, так как $x^{(1)} + x^{(2)} - 3x^{(3)} = 0$. В то же время $x^{(1)}$, $x^{(2)}$ линейно независимы, так как из векторного равенства $\alpha_1 x^{(1)} + \alpha_2 x^{(2)} = 0$ следует, что $\alpha_1 = \alpha_2 = 0$.

Легко убедиться, что система ненулевых попарно ортогональных векторов пространства $\mathbb{R}^n$ всегда линейно независима. Действительно, пусть $\sum_{j=1}^{m} \alpha_k x^{(k)} = 0$. Умножим это равенство скалярно на $x^{(j)}$, $1 \le j \le m$. Так как $(x^{(i)}, x^{(j)}) = 0$ при $i \ne j$, то получаем в результате уравнение $\alpha_j (x^{(j)}, x^{(j)}) = 0$, $1 \le j \le m$. Следовательно, $\alpha_j = 0$, $1 \le j \le m$.

Максимальная система линейно независимых векторов подпространства $\mathcal{L}$ называется базисом $\mathcal{L}$; другими словами, система линейно независимых векторов $\{x^{(1)}, x^{(2)}, \ldots, x^{(k)}\} \subset \mathcal{L}$ является базисом подпространства $\mathcal{L}$, если любой вектор $x \in \mathcal{L}$ представим в виде $x = \alpha_1 x^{(1)} + {} + \alpha_2 x^{(2)} + \ldots + \alpha_k x^{(k)}$.

Можно показать, что каждое нетривиальное подпространство $\mathcal{L}$ пространства $\mathbb{R}^n$ имеет базис и все базисы одного подпространства $\mathcal{L}$ состоят из одинакового числа векторов. Это число называется *размерностью* подпространства $\mathcal{L}$ и обозначается $\dim \mathcal{L}$. Размерность тривиального (нулевого) подпространства $\{0\}$ полагаем равной нулю.

Важно отметить, что каждое нетривиальное подпространство $\mathcal{L}$ имеет ортонормированный базис и что любая ортонормированная система векторов подпространства $\mathcal{L}$ может быть дополнена до ортонормированного базиса этого подпространства.

П р и м е р. В качестве $\mathcal{L}$ рассмотрим $\mathbb{R}^n$. Стандартный (ортонормированный) базис $\mathbb{R}^n$ - это набор n векторов $e^{(i)} = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$, содержащих 1 в i-позиции и 0 в остальных. Поэтому $\dim \mathbb{R}^n = n$.

В силу определения базиса подпространства $\mathcal{L}$ пространства $\mathbb{R}^n$ любой вектор $x \in \mathcal{L}$ представим в виде $x = \sum_{j=1}^{k} \alpha_j x^{(j)}$, где $x^{(1)}, x^{(2)}, \ldots, x^{(k)}$ - базис подпространства $\mathcal{L}$. Коэффициенты α_j определены однозначно. Действительно, пусть $x = \sum_{j=1}^{k} \beta_j x^{(j)}$ - другое представление вектора x через базис $x^{(1)}, x^{(2)}, \ldots, x^{(k)}$. Тогда, очевидно, что

$$(\alpha_1 - \beta_1) x^{(1)} + (\alpha_2 - \beta_2) x^{(2)} + \ldots + (\alpha_k - \beta_k) x^{(k)} = 0.$$

Из линейной независимости векторов базиса $x^{(k)}$, $1 \le j \le k$, следует, что $\alpha_j - \beta_j = 0$ для всех $j = 1, 2, \ldots, k$. Коэффициенты α_j в разложении вектора $x = \sum_{j=1}^{k} \alpha_j x^{(j)}$ по базису $x^{(1)}, x^{(2)}, \ldots, x^{(k)}$ подпространства $\mathcal{L}$ называются *координатами вектора x в базисе* $x^{(1)}, x^{(2)}, \ldots, x^{(k)}$.

С помощью координат векторов в базисе $x^{(1)}, x^{(2)}, \ldots, x^{(k)}$ подпространство $\mathcal{L} \in \mathbb{R}^n$ можно отождествлять с пространством $\mathbb{R}^k$, т.е. любой вектор $x \in \mathcal{L} \in \mathbb{R}^n$ можно представить в виде $x = (x_1, \ldots, x_k)^\top$, где x_i - координаты x в базисе $x^{(1)}, x^{(2)}, \ldots, x^{(k)}$. Фактически и арифметическое n-мерное пространство $\mathbb{R}^n$ является множеством векторов с компонентами (координатами) разложения по стандартному базису $e^{(1)}, e^{(2)}, \ldots, e^{(n)}$.

Базис подпространства $\mathcal{L}$ пространства $\mathbb{R}^n$ дает возможность определить это подпространство конструктивно: множество векторов из $\mathcal{L}$ состоит из всех линейных комбинаций векторов базиса, другими словами, является *линейной оболочкой базиса*. Линейная оболочка любой системы векторов $x^{(1)}, x^{(2)}, \ldots, x^{(k)}$, обозначаемая через $\mathrm{span}(x^{(1)}, x^{(2)}, \ldots, x^{(k)})$, образует подпространство размерности m, $m \le k$. Преимущество базиса перед другими системами векторов, которые имеют одинаковые линейные оболочки, заключается в том, что в разложении векторов по базису коэффициенты определяются однозначно.

Некоторые подпространства возникают естественным образом в связи с матрицами. Так, с $m \times n$-матрицей A связано *пространство столбцов* $\mathrm{im}\, A$ (или *образ матрицы*), т.е. линейная оболочка столбцов матрицы A,

$$\mathrm{im}\, A = \mathrm{span}(a_1, a_2, \ldots, a_n) = \{y \in \mathbb{R}^m : Ax = y, \forall\, x \in \mathbb{R}^n\},$$

где $A = (a_1, a_2, \ldots, a_n)$. *Пространство строк* матрицы A это $\mathrm{im}\, A^\top$.

Пространства строк и столбцов матрицы A имеют одинаковую размерность, называемую *рангом матрицы* и обозначаемую через $\mathrm{rank}\, A$. Говорят, что *матрица A размера $m \times n$ имеет неполный ранг*, если $\mathrm{rank}\, A < \min(m, n)$, и *полный ранг*, если $\mathrm{rank}\, A = \min(m, n)$. Заметим, что $\mathrm{rank}\, PAQ = \mathrm{rank}\, A$ для любых невырожденных матриц P и Q.

Можно показать, что вырожденность $n \times n$-матрицы A эквивалентна существованию такого ненулевого вектора x, что $Ax = 0$. Следовательно, квадратная матрица A порядка n невырождена, если $\mathrm{rank}\, A = n$, и вырождена, если $\mathrm{rank}\, A < n$.

Определим *правое нуль-пространство* $m \times n$-матрицы A (или *ядро матрицы*) как множество

$$\ker A = \{x \in \mathbb{R}^n : Ax = 0\}$$

и соответственно левое нуль-пространство матрицы A как $\ker A^\top = \{y \in \mathbb{R}^m : A^\top y = 0\}$. Несложно показать, что

$$\operatorname{rank} A = n - \dim \ker A = m - \dim \ker A^\top.$$

1.2. Матричная алгебра

1.2.1. Определения и обозначения. Приведем определения и обозначения из матричной алгебры используемые в данном учебном пособии.

$\mathbb{R}^{m\times n}$ ($\mathbb{C}^{m\times n}$) - множество вещественных (комплексных) $m \times n$-матриц, имеющих m строк и n столбцов;

a_{ij} $(i = 1, \dots, m,\ j = 1, \dots, n)$ - элементы матрицы $A \in \mathbb{R}^{m\times n}$, $A = (a_{ij})$;

$0 \in \mathbb{R}^{m\times n}$ - матрица, состоящая из нулей (или $0_{m\times n}$);

$\operatorname{diag}(d_{11}, \dots, d_{nn}) \in \mathbb{R}^{n\times n}$ - диагональная матрица (ее диагональные элементы равны d_{ii}, а внедиагональные - нулю);

$E_n \in \mathbb{R}^{n\times n}$ - единичная матрица, $E_n = \operatorname{diag}(1, \dots, 1)$ (если из контекста очевиден порядок единичной матрицы, то матрица обозначается без индекса - E);

$\lambda_1(B), \dots, \lambda_n(B)$ - *собственные числа* матрицы $B \in \mathbb{R}^{n\times n}$, т.е. корни характеристического уравнения

$$\det(B - \lambda E_n) = 0;$$

$$\lambda_{\min}(B) = \min_{1\le i\le n} \lambda_i(B), \qquad \lambda_{\max}(B) = \max_{1\le i\le n} \lambda_i(B);$$

$$\lambda_i = \lambda_i(A), \quad i = 1, \dots, n, \quad A \in \mathbb{R}^{n\times n}.$$

Квадратная матрица $U = (u_{ij}) \in \mathbb{R}^{n\times n}$ называется *верхней треугольной*, если все ее элементы ниже главной диагонали равны нулю, т.е. $u_{ij} = 0\ \forall\, i > j$ (соответственно, квадратная матрица $L = (l_{ij}) \in \mathbb{R}^{n\times n}$ называется *нижней треугольной*, если все ее элементы ниже главной диагонали равны нулю, т.е. $l_{ij} = 0\ \forall\, i < j$.

Следом квадратной матрицы $A \in \mathbb{R}^{n\times n}$ называется сумма ее диагональных элементов $\operatorname{tr} A = \sum_{i=1}^n a_{ii}$.

Матрица $A \in \mathbb{R}^{n\times n}$ называется *невырожденной*, если $\det A \neq 0$.

Обратной для невырожденной матрицы $A \in \mathbb{R}^{n\times n}$ называется такая матрица $A^{-1} \in \mathbb{R}^{n\times n}$, что $A^{-1}A = AA^{-1} = E_n$.

Квадратная матрица A называется *симметричной*, если $A = A^\top$. Квадратная матрица A называется *эрмитовой*, если $A = A^*$, где A^* – матрица эрмитово сопряженная к матрице A, т.е. $A^* = \bar{A}^\top$.

Симметричная матрица $A \in \mathbb{R}^{n\times n}$ называется *положительно (неотрицательно* определенной, если $\forall\, a \in \mathbb{R}^n \backslash \{0\}$ выполнено $a^\top A a > 0$ (соответственно, $a^\top A a \geq 0$).

$\mathbb{R}^{\geq}_n$ - множество неотрицательно определенных квадратных матриц порядка n.

$\mathbb{R}^{>}_n$ - множество положительно определенных квадратных матриц порядка n.

Для любых матриц $A, B \in \mathbb{R}^{\geq}_n$ запись $A > B$ $(A \geq B)$ означает, что $A - B \in \mathbb{R}^{>}_n$ (соответственно, $A - B \in \mathbb{R}^{\geq}_n$).

Матрицы называются *согласованными* относительно некоторой операции, если эта операция определена.

1.2.2. Симметричные, положительно и неотрицательно определенные матрицы.

Теорема 1.1 (теорема о спектральном разложении). *Если* $A = A^\top \in \mathbb{R}^{n\times n}$, *то выполнено*

$$P^\top A P = \Lambda, \qquad A = P \Lambda P^\top. \tag{1.1}$$

где P *– ортогональная* $n \times n$*-матрица (т.е.* $P^\top P = P P^\top = E_n$*), столбцами которой являются ортонормированные собственные векторы матрицы* A, $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, *где* λ_i *– собственные числа матрицы* A.

Доказательство имеется в [4, 6, 16].

Теорема 1.2. *Матрица* $A \in \mathbb{R}^{n\times n}$ *неотрицательно определена тогда и только тогда, когда существует матрица* $F \in \mathbb{R}^{n\times n}$, *такая, что* $A = F F^\top$.

Теорема 1.3. *Матрица* $A \in \mathbb{R}^{n\times n}$ *положительно определена тогда и только тогда, когда существует невырожденная матрица* $F \in \mathbb{R}^{n\times n}$, *такая, что* $A = F^\top F$.

В теоремах 1.2 и 1.3 необходимость вытекает из теоремы 1.1, а достаточность – из определения неотрицательной (положительной) определенности.

Теорема 1.4. *Любую матрицу* $A \in \mathbb{R}^{>}_n$ *можно представить в виде* $A = F F^\top$, *где матрица* $F \in \mathbb{R}^{n\times n}$ *невырождена и является верхней треугольной.*

Доказательство имеется в [6].

Теорема 1.5. *Если* $A \in \mathbb{R}_n^>$, *то*

$$\lambda_i \geq 0, \quad a_{ii} \geq 0, \quad \det A = \prod_{i=1}^{n} \lambda_i,$$

$$a_{ij} \leq (a_{ii} + a_{jj})/2, \tag{1.2}$$

где $i, j = 1, \ldots, n$.

Если $A \in \mathbb{R}_n^>$, *то*

$$\lambda_i > 0, \quad a_{ii} > 0, \quad \det A > 0, \quad a_{ij} < (a_{ii} + a_{jj})/2, \quad i, j = 1, \ldots, n.$$

Теорема 1.6. *Если* $A \in \mathbb{R}_n^>$, $B \in \mathbb{R}_n^>$, $A + B \in \mathbb{R}_n^>$.

Теорема 1.7. *Если* $A \in \mathbb{R}_n^>$, *то* $A^{-1} \in \mathbb{R}_n^>$.

Теорема 1.8. *Пусть* $A \in \mathbb{R}_n^>$, $B \in \mathbb{R}^{n \times m}$, $\operatorname{rank} B = m \leq n$. *Тогда* $B^\top AB \in \mathbb{R}_n^>$. *В частности,* $B^\top B \in \mathbb{R}_n^>$.

Доказательство теорем 1.5 – 1.8 вытекают из определений неотрицательной и положительной определенности матриц и теорем 1.1 – 1.3. Комментария требует только вывод формулы 1.2. Эта формула следует из того, что

$$(e_i - e_j)^\top A(e_i - e_j) = e_i^\top Aa_i + e_j^\top Aa_j - 2e_i^\top Aa_j = a_{ii} + a_{jj} - 2a_{ij} \geq 0,$$

где $e_i \in \mathbb{R}^n$ – вектор, все компоненты которого равны нулю кроме i-й, равной единице.

Теорема 1.9. *Пусть матрица* $A \in \mathbb{R}_n^>$ *симметрична,* $\lambda_1 \geq \ldots \geq \lambda_n$ – *ее собственные числа и* $p_1, \ldots, p_n$ – *соответствующие им ортонормированные собственные векторы. Тогда*

$$\sup_{a \in \mathbb{R}^n \setminus \{0\}} \left\{ \frac{a^\top Aa}{a^\top a} \right\} = \lambda_1, \tag{1.3}$$

$$\inf_{a \in \mathbb{R}^n \setminus \{0\}} \left\{ \frac{a^\top Aa}{a^\top a} \right\} = \lambda_n, \tag{1.4}$$

причем экстремумы достигаются соответственно на p_1 *и* p_n.

Д о к а з а т е л ь с т в о. Формулу (1.1) перепишем в виде:

$$A = \sum_{i=1}^{n} \lambda_i p_i p_i^\top, \qquad \sum_{i=1}^{n} p_i p_i^\top = E_n.$$

В силу того, что $\{p_i\}_{i=1}^n$ – базис в $\mathbb{R}^n$ любой вектор $a \in \mathbb{R}^n$ представим в виде

$$a = \sum_{i=1}^{n} c_i p_i.$$

Поэтому

$$\frac{a^\top A a}{a^\top a} = \sum_{i=1}^{n} c_i^2 \lambda_i \left(\sum_{i=1}^{n} c_i^2 \right)^{-1}.$$

Очевидно, что супремум и инфимум этого выражения относительно векторов $(c_1, \ldots, c_n)^\top$ равны соответственно λ_1 и λ_n, причем супремум достигается при $a = p_1$, а инфимум – при $a = p_n$. □

1.2.3. След.

Теорема 1.10. *а) Если* $A \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}$, *то* $\operatorname{tr} A = \operatorname{tr} A^\top$, $\operatorname{tr} cA = c \operatorname{tr} A$;
б) если $A, B \in \mathbb{R}^{n\times n}$, *то*

$$\operatorname{tr}(A + B) = \operatorname{tr} A + \operatorname{tr} B; \tag{1.5}$$

в) если $A \in \mathbb{R}^{n\times m}$, $B \in \mathbb{R}^{m\times n}$, *то*

$$\operatorname{tr} AB = \operatorname{tr} BA; \tag{1.6}$$

г) если $a, b \in \mathbb{R}^n$, *то*

$$\operatorname{tr} ab^\top = \operatorname{tr} a^\top b;$$

д) если $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, *то*

$$\operatorname{tr}(Abb^\top) = \operatorname{tr}(bb^\top A) = b^\top A b;$$

е) если $A, P \in \mathbb{R}^{n\times n}$, $P^\top P = E_n$, *то*

$$\operatorname{tr} PAP^\top = \operatorname{tr} A. \tag{1.7}$$

Все утверждения теоремы легко следуют из определения операции trace (след).

Теорема 1.11. *Если A – симметричная матрица из $\mathbb{R}^{n\times n}$, то*

$$\operatorname{tr} A = \sum_{i=1}^{n} \lambda_i, \tag{1.8}$$

$$\operatorname{tr} A^s = \sum_{i=1}^{n} \lambda_i^s, \qquad s = -1, 0, 1, 2, \ldots.$$

Доказательство следует из теоремы 1.1 и (1.7) с учетом того, что при $PP^\top = E_n$ выполнено $(P^\top AP)^s = P^\top A^s P$.

Теорема 1.12. *Пусть A – симметричная $n \times n$-матрица. Необходимым и достаточным условием ее неотрицательной определенности является выполнение для всех $B \in \mathbb{R}_n^{\geq}$ неравенства $\operatorname{tr} AB \geq 0$.*

Доказательство. По теореме (1.1) имеем

$$A = P\Lambda P = \sum_{i=1}^{n} \lambda p_i p_i^\top,$$

где P_i – ортонормированные собственные векторы матрицы A, соответствующие собственным числам λ_i. Отсюда следует:

$$\operatorname{tr} AB = \operatorname{tr} \sum_{i=1}^{n} \lambda_i p_i p_i^\top B = \sum_{i=1}^{n} \lambda_i p_i^\top B p_i.$$

Поскольку $B \in \mathbb{R}_n^{\geq}$, величины $p_i^\top B p_i$ $(i = 1, \ldots, n)$ неотрицательны.

Если $A \in \mathbb{R}_n^{\geq}$, то по теореме 1.5 $\lambda_i \geq 0$, поэтому $\operatorname{tr} AB \geq 0$. С другой стороны, если $\operatorname{tr} AB \geq 0$ для всех $B \geq 0$, то это справедливо и для матрицы $B = p_i p_i^\top$. Следовательно,

$$\operatorname{tr} A p_i p_i^\top = \operatorname{tr} \left(\sum_{j=1}^{n} \lambda_j p_j p_j^\top \right) p_i p_i^\top = \lambda_i \geq 0.$$

Отсюда с учетом теоремы 1.1 следует, что $A \geq 0$. □

1.2.4. Ранг.

Теорема 1.13. *Для любых согласованных матриц A, B выполнено*

а) $\operatorname{rank} AB \leq \min(\operatorname{rank} A, \operatorname{rank} B)$;

б) $\operatorname{rank}(A + B) \leq \operatorname{rank} A + \operatorname{rank} B$.

Д о к а з а т е л ь с т в о. а) Столбцы матрицы AB являются линейными комбинациями столбцов матрицы A, поэтому число линейно независимых столбцов в матрице AB не больше, чем в матрице A. Следовательно, $\operatorname{rank} AB \le \operatorname{rank} A$. Аналогично $\operatorname{rank} AB \le \operatorname{rank} B$.

б) Пусть матрицы A и B имеют размер $p \times q$. Обозначим через $a_1, \ldots, a_q$ и $b_1, \ldots, b_q$ столбцы матриц A и B соответственно, и пусть

$$D = (A, B) = (a_1, \ldots, a_q, b_1, \ldots, b_q)$$

есть блочная матрица, составленная из матриц A и B.

Запишем матрицу $A + B$ в виде

$$A + B = (a_1 + b_1, \ldots, a_q + b_q).$$

Поскольку размерность пространства, порожденного набором векторов $(a_1, \ldots, a_q, b_1, \ldots, b_q)$ не меньше, чем размерность пространства для векторов $(a_1 + b_1, \ldots, a_q + b_q)$, то $\operatorname{rank}(A + B) \le \operatorname{rank} D$.

Покажем теперь, что $\operatorname{rank} D \le \operatorname{rank} A + \operatorname{rank} B$. Для этого удалим из набора $(a_1, \ldots, a_q, b_1, \ldots, b_q)$ все векторы b_i, линейно зависящие от векторов a_j $(j = 1, \ldots, q)$. Матрицу, составленную из оставшихся векторов b_i , обозначим через B_*. Имеем:

$$\begin{aligned} \operatorname{rank} D &= \operatorname{rank} A + \operatorname{rank} B_*, \\ \operatorname{rank} B_* &\le \operatorname{rank} B, \end{aligned}$$

откуда и вытекает требуемое. □

Теорема 1.14. *Пусть* $A \in \mathbb{R}^{n\times m}$, $B \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{m\times m}$, $\det B \ne 0$, $\det C \ne 0$. *Тогда*

$$\operatorname{rank} BAC = \operatorname{rank} A.$$

Д о к а з а т е л ь с т в о. В силу предыдущей теоремы

$$\operatorname{rank} A \ge \operatorname{rank} AC \ge \operatorname{rank} ACC^{-1} = \operatorname{rank} A.$$

Поэтому $\operatorname{rank} A = \operatorname{rank} AC$. Аналогично $\operatorname{rank} A = \operatorname{rank} BAC$. □

Теорема 1.15. *Если матрица симметрична, то ее ранг равен числу ненулевых ее собственных значений.*

Доказательство следует из теорем 1.1 и 1.14.

1.3. Нормы векторов и матриц

1.3.1. Векторные нормы. Пусть V – линейное пространство, вещественное или комплексное. В дальнейшем под линейным пространством V будем понимать $\mathbb{R}^n$ или $\mathbb{C}^n$. *Нормой* в линейном пространстве V называется отображение $\|\cdot\| : V \to \mathbb{R}$, ставящее в соответствие каждому вектору $x \in V$ число $\|x\| \in \mathbb{R}$ и удовлетворяющее аксиомам: $\forall\, x, y \in V$, $\alpha \in \mathbb{R}(\mathbb{C})$

1) $\|x\| \geq 0, \quad \|x\| = 0 \Leftrightarrow x = 0$ (неотрицательность),

2) $\|\alpha x\| = |\alpha| \cdot \|x\|$ (однородность),

3) $\|x + y\| \leq \|x\| + \|y\|$ (неравенство треугольника).

Линейное пространство V с заданной на нем нормой $\|\cdot\|$ называется *линейным нормированным пространством*. Число $\|x\|$ называется *нормой вектора* x.

Наиболее употребительными в арифметических пространствах являются:

1. *Октаэдрическая норма вектора* (1-норма)

$$\|x\|_1 = \sum_{k=1}^{n} |x_k|.$$

2. *Евклидова или сферическая норма вектора* (2-норма)

$$\|x\|_2 = \|x\|_E = \sqrt{\sum_{k=1}^{n} |x_k|^2}.$$

3. *Кубическая норма вектора* (3-норма)

$$\|x\|_\infty = \max_{1 \leq k \leq n} |x_k|.$$

4. *Норма Гёльдера* (p-норма)

$$\|x\|_p = \left(\sum_{k=1}^{n} |x_k|^p \right)^{1/p}, \qquad 1 \leq p \leq \infty.$$

Нетрудно заметить, что первые три векторные нормы являются частным случаем нормы Гёльдера, соответственно для $p = 1, 2, \infty$.

1.3.2. Эквивалентность норм в конечномерном пространстве. Две нормы $\|x\|_1$ и $\|x\|_2$ в линейном пространстве V называются

эквивалентными, если существуют такие числа $c_1 > 0$, $c_2 > 0$, что для любого вектора $x \in V$ выполняются неравенства

$$\|x\|_1 \le c_1\|x\|_2 \quad \text{и} \quad \|x\|_2 \le c_2\|x\|_1.$$

Теорема 1.16. *В конечномерном пространстве любые две нормы эквивалентны.*

Доказательство имеется в [11].

1.3.3. Нормы матриц. Под *нормой матрицы* A с действительными или комплексными элементами понимают действительное число $\|A\|$, т.е. отображение $\|A\| : \mathbb{R}^{m\times n}(\mathbb{C}^{m\times n}) \to \mathbb{R}$, и удовлетворяющее аксиомам:

1) $\|A\| \ge 0, \quad \|A\| = 0 \Leftrightarrow A = 0$ (неотрицательность),

2) $\|\alpha A\| = |\alpha| \cdot \|A\|$ (однородность),

3) $\|A + B\| \le \|A\| + \|B\|$ (неравенство треугольника),

4) $\|AB\| \le \|A\| \cdot \|B\|$ (мультипликативность)

$\forall\, A, B \in \mathbb{R}^{m\times n}(\mathbb{C}^{m\times n})$ (согласованных относительно указанных операций матриц) и $\forall\, \alpha \in \mathbb{R}(\mathbb{C})$.

Иногда в определении нормы матрицы ограничиваются лишь первыми тремя аксиомами. В таком случае норму матрицы называют *обобщенной*.

Примерами матричных норм матрицы $A = (a_{ij})$ являются:

1) $\|A\| = \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|$;

2) $\|A\|_1 = \max_{1\le j\le n}\sum_{i=1}^{m}|a_{ij}|$ – *максимально столбцовая норма*;

3) $\|A\|_\infty = \max_{1\le i\le m}\sum_{j=1}^{n}|a_{ij}|$ – *максимально строковая норма*;

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

Норма $\|A\|_E$ называется *евклидовой* (*сферической, Фробениуса, Шура*) матричной нормой.

Для матрицы

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix}$$

все эти нормы будут иметь соответственно значения:

1. $\|A\| = \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}| = 1 + 2 + \ldots + 12 = 78.$

2. $\|A\|_1 = \max_{1\le j\le n}\sum_{i=1}^{m}|a_{ij}| = \max(1 + 4 + 7 + 10, 2 + 5 + 8 + 11, 3 + 6 + 9 + 12) = \max(22, 26, 30) = 30.$

3. $\|A\|_\infty = \max_{1\le i\le m}\sum_{j=1}^{n}|a_{ij}| = \max(1+2+3, 4+5+6, 7+8+9, 10+11+12) = \max(6, 15, 24, 33) = 33.$

4. $\|A\|_E = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\right)^{1/2} = \sqrt{1^2+2^2+\ldots+12^2} = \sqrt{650}.$

Норму матрицы $A \in \mathbb{R}^{m\times n}$ ($\mathbb{C}^{m\times n}$) называют *согласованной с векторной нормой*, если для любого вектора $x \in \mathbb{R}^n(\mathbb{C}^n)$ выполняется условие

$$\|Ax\| \le \|A\| \cdot \|x\|.$$

Часто норму матрицы A вводят через нормы векторов, полагая

$$\|A\| = \sup_{x\ne 0}\frac{\|Ax\|}{\|x\|} = \sup_{\|x\|}\|Ax\|.$$

Такую норму матрицы называют *матричной нормой, подчиненной векторной норме*, или *матричной нормой, индуцированной векторной нормой.*

Нетрудно показазать, что любая матричная норма единичной матрицы E_n, подчиненная векторной норме равна 1. Из этого факта непосредственно следует, что сферическая (евклидова) матричная норма не подчинена никакой векторной норме.

Приведем примеры подчиненных матричных норм.

1. Для октаэдрической нормы вектора $\|x\|_1$, подчиненной нормой матрицы A является

$$\|A\|_1 = \max_{1\le j\le n}\sum_{i=1}^{m}|a_{ij}|.$$

2. Для евклидовой (сферической) нормы вектора $\|x\|_2$, подчиненной матричной нормой является *спектральная матричная норма*

$$\|A\|_2 = \sup_{x\ne 0}\frac{\|Ax\|_2}{\|x\|_2} = \sup_{\|x\|_2}\|Ax\|_2 = \sqrt{\max_{1\le i\le n}|\lambda_i|},$$

где $\lambda_i = \lambda_i(A^*A)$, $i = 1, 2, \ldots, n$.

3. Для кубической нормы вектора $\|x\|_\infty$, подчиненной матричной нормой является

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

Между различными матричными нормами устанавливаются определенные соотношения. Особенно много таких соотношений приведено в [19].

1.3.4. Сходимость по норме.

Утверждение 1.1. *В нормированном пространстве V отображение $\rho : V \times V \to \mathbb{R}$, определенное равенством*

$$\rho(x, y) = \|x - y\|, \qquad \forall\, x, y \in V,$$

является метрикой.

Аксиомы метрики непосредственно вытекают из аксиом нормы. □

Таким образом, в нормированном пространстве можно ввести расстояние между векторами и, значит, пользоваться предельным переходом. Последовательность векторов $\{x^{(k)}\}$ в нормированном пространстве V называется *сходящейся по норме* к вектору $a \in V$, если $\lim\limits_{k\to\infty} \|x^{(k)} - a\| = 0$, при этом вектор a называется *пределом последовательности* $\{x^{(k)}\}$ *по норме* $\|\cdot\|$.

О б о з н а ч е н и е: $\lim\limits_{k\to\infty} x^{(k)} = a$ или $x^{(k)} \to a$.

Утверждение 1.2. *Сходящаяся по норме последовательность имеет единственный предел.*

Д о к а з а т е л ь с т в о повторяет доказательство аналогичной теоремы для числовой последовательности [12] и основано на аксиоме треугольника: $\|a - b\| = \|a - x^{(k)} + x^{(k)} - b\| \le \|x^{(k)} - a\| + \|x^{(k)} - b\|$, где a и b – два предела последовательности $x^{(k)}$. □

Пусть $x_0 \in V$ и $r > 0$. Множество $S(x_0, r) = \{x \in V : \|x - x_0\| = r\}$ называется *сферой радиуса r с центром x_0 по норме* $\|\cdot\|$, а множество $B(x_0, r) = \{x \in V : \|x - x_0\| \le r\}$ – *замкнутым шаром радиуса r с центром x_0 по норме* $\|\cdot\|$.

В дальнейшем сферы и шары по евклидовой норме $\|\cdot\|_2$ будут обозначаться символами $S_E(x_0, r)$ и $B_E(x_0, r)$.

Утверждение 1.3. *Из любой последовательности векторов $x^{(k)} \in B_E(x_0, r)$ (или $S_E(x_0, r)$) можно выделить подпоследовательность, сходящуюся по норме $\|\cdot\|_2$ к вектору $a \in B_E(x_0, r)$ ($S_E(x_0, r)$ соответственно).*

Д о к а з а т е л ь с т в о. Без ограничения общности можно считать, что $x_0 = 0$. Доказательство проведем для сферы $S_E(r) = \{x \in V : \|x\|_2 = r\}$. Пусть $e_1, \ldots, e_n$ – ортонормированный базис пространства V и $x^{(k)} = \sum_{i=1}^{n} x^{(k)} e_i \in S_E(r)$. Тогда

$$\|x^{(k)}\|_2 = \left(\sum_{i=1}^{n} |x_i^{(k)}|^2 \right)^{1/2} = r.$$

Это означает ограниченность координат векторов $x^{(k)}$ рассматриваемой последовательности. Согласно теореме Больцано-Вейерштрасса [12] из этой последовательности можно выделить сходящуюся (покоординатно) подпоследовательность $\{x^{(k_m)}\}$. Пусть $x^{(k_m)}$ имеет координаты $x_1^{(k_m)}, \ldots, x_n^{(k_m)}$, сходящиеся соответственно к $a_1, \ldots, a_n$. Положим $a = \sum_{i=1}^{n} a_i e_i$. Тогда

$$\|x^{(k_m)} - a\|_2 = \left(\sum_{i=1}^{n} \left| x_i^{(k_m)} - a_i \right|^2 \right)^{1/2} \to 0,$$

следовательно, подпоследовательность $\{x^{(k_m)}\}$ сходится к вектору a по евклидовой норме.

Покажем, что $a \in S_E(r)$. Действительно, в очевидном неравенстве $|\|x\| - \|y\|| \le \|x-y\|$ положим $x = x^{(k_m)}$, $y = a$. Тогда $\left| \|x^{(k_m)}\|_2 - \|a\|_2 \right| \le \|x^{(k_m)} - a\|_2$, откуда следует, что

$$\|x^{(k_m)}\|_2 - \|x^{(k_m)} - a\|_2 \le \|a\|_2 \le \|x^{(k_m)}\|_2 + \|x^{(k_m)} - a\|_2$$

или, с учетом того, что $\|x^{(k_m)} - a\|_2 \to 0$,

$$\|x^{(k_m)}\|_2 - \varepsilon \le \|a\|_2 \le \|x^{(k_m)}\|_2 + \varepsilon, \quad \forall \varepsilon > 0,$$

если m достаточно велико. Следовательно, $\|a\|_2 = r$ и $a \in S_E(r)$. □

Пусть $\|\cdot\|_1$ и $\|\cdot\|_2$ две произвольные векторные нормы в пространстве V. Тогда из теоремы 1.16 об эквивалентности норм в конечномерном пространстве непосредственно следует, что *в конечномерном пространстве из сходимости по одной норме следует сходимость по любой другой норме*, так как

$$\|x^{(k)} - a\|_1 \le c_1 \|x^{(k)} - a\|_2.$$

1.4. Сингулярное разложение матриц

1.4.1. Сингулярные числа и векторы матриц. Возможность построения для симметричной (эрмитовой) матрицы канонического разложения с ортогональной (унитарной) трансформирующей матрицей позволяет для произвольной $(m \times n)$-матрицы получить аналог такого разложения. В дальнейшем все изложение ведется для матриц из $\mathbb{C}^{m\times n}$, поэтому все результаты справедливы также для матриц из $\mathbb{R}^{m\times n}$. При этом символ " $*$ " эрмитова сопряжения необходимо заменить на знак транспонирования "$\top$". Прежде всего заметим, что для произвольной матрицы $A \in \mathbb{C}^{m\times n}$ ранга r матрицы A^*A и AA^* являются симметричными (эрмитовыми) матрицами ранга r и порядков соответственно n и m. Причем они неотрицательны. Поэтому собственные числа таких матриц являются действительными неотрицательными числами.

Обозначим собственные числа матрицы A^*A через $\sigma_1^2, \sigma_1^2, \ldots, \sigma_n^2$ и будем считать, что $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_n^2$ ($\sigma_i \neq 0$ при $i = 1, \ldots, r$). Оператор с симметричной (эрмитовой) матрицей A^*A имеет ортонормированную систему собственных векторов $e_1, e_2, \ldots, e_n$ соответственно по $\sigma_1^2, \sigma_1^2, \ldots, \sigma_n^2$, т.е. таких векторов $e_1, e_2, \ldots, e_n$, что

$$A^*Ae_i = \sigma_i^2 e_i, \quad (e_i, e_j) = \begin{cases} 1, & \text{при } i = j; \\ 0, & \text{при } i \neq j,\ i, j = \overline{1, n}. \end{cases} \tag{1.9}$$

Эта система векторов переводится оператором с матрицей A в некоторую ортогональную систему векторов $Ae_1, Ae_2, \ldots, Ae_n$, так как

$$(Ae_i, Ae_j) = (A^*Ae_i, e_j) = \sigma_i^2(e_i, e_j) = 0 \quad \text{при} \quad i \neq j.$$

Кроме того, модуль вектора Ae_i равен σ_i, так как

$$|Ae_i| = \sqrt{(A^*Ae_i, e_j)} = \sqrt{\sigma_i^2(e_i, e_i)} = \sigma_i.$$

Поэтому вектор Ae_i отличен от нулевого вектора тогда и только тогда, когда $\sigma_i \neq 0$, т.е. при $i = \overline{1, r}$. Нулевой вектор Ae_i является собственным вектором оператора AA^* по собственному значению σ_i^2, так как

$$AA^*(Ae_i) = A(A^*A)e_i = A(\sigma_i^2 e_i) = \sigma_i^2 Ae_i.$$

Верно и обратное. Следовательно, ненулевые характеристические числа матриц A^*A и AA^* совпадают с учетом их кратностей и их число равно r, а кратности нулевого собственного числа этих матриц равны

соответственно $n-r$ и $m-r$. Общих собственных чисел у матриц A^*A и AA^* будет $s=\min(m,n)$.

Арифметические значения $\sigma_1,\sigma_2,\ldots,\sigma_s$ ($\sigma_i\neq 0$ при $i=\overline{1,r}$) корней квадратных из общих собственных чисел матриц A^*A и AA^* называются *сингулярными* (или *главными*) *числами марицы* A.

В пространстве $\mathbb{C}^n$ примем за базис ортонормированную систему $e_1,e_2,\ldots,e_n$ собственных векторов оператора с матрицей A^*A и построим ортонормированную систему векторов

$$f_1=\frac{Ae_1}{|Ae_1|}=\frac{Ae_1}{\sigma_1},\ldots,f_r=\frac{Ae_r}{|Ae_r|}=\frac{Ae_r}{\sigma_r}.$$

Дополним эту систему любыми векторами $f_{r+1},\ldots,f_m$ до ортонормированного базиса в $\mathbb{C}^m$. По построению векторы $f_1,f_2,\ldots,f_m$ удовлетворяют соотношениям

$$Ae_i=\begin{cases}\sigma_i f_i, & \text{при } i\leq r,\\ 0, & \text{при } i>r.\end{cases}\tag{1.10}$$

Умножая эти равенства слква на A^* и учитывая, что $A^*Ae_i=\sigma_i^2e_i$, получим соотношения

$$A^*f_i=\begin{cases}\sigma_i e_i, & \text{при } i\leq r,\\ 0, & \text{при } i>r.\end{cases}\tag{1.11}$$

Ортонормированные базисы $e_1,e_2,\ldots,e_n$ и $f_1,f_2,\ldots,f_m$ пространств $\mathbb{C}^n$ и $\mathbb{C}^m$, связанные соотношениями (1.10) и (1.11), называют *сингулярными базисами*. Причем векторы $e_1,e_2,\ldots,e_n$ называют *правыми сингулярными векторами матрицы* A, а векторы $f_1,f_2,\ldots,f_m$ – ее *левыми сингулярными векторами*.

Оператор, имеющий в паре исходных базисов пространств $\mathbb{C}^n$ и $\mathbb{C}^m$ матрицу A, в сингулярных базисах $e_1,e_2,\ldots,e_n$ и $f_1,f_2,\ldots,f_m$ этих пространств, в силу определения матрицы оператора и соотношений (1.10), имеет $(m\times n)$-матрицу

$$\Sigma=\begin{pmatrix}\sigma_1 & & & 0\\ & \ddots & & \\ & & \sigma_r & \\ 0 & & & 0\end{pmatrix}.\tag{1.12}$$

При этом из формулы, устанавливающей связь между матрицами одного и того же оператора в разных базисах, получаем

$$A=Q\Sigma P^*,\tag{1.13}$$

где P – ортогональная (унитарная) матрица порядка n, столбцами которой служат столбцы координат векторов $e_1, e_2, \ldots, e_n$ в исходном базисе пространства $\mathbb{C}^n$, Q – ортогональная (унитарная) матрица порядка m, столбцами которой являются столбцы координат векторов $f_1, f_2, \ldots, f_m$ в исходном базисе пространства $\mathbb{C}^m$.

Разложение (1.13) называют *сингулярным разложением матрицы* A или сокращенно *SVD-разложением*, где SVD – сокращение (аббревиатура) английского термина "singular value decomposition".

Любая матрица из $\mathbb{C}^{m\times n}$ ($\mathbb{R}^{m\times n}$) обладает многими различными сингулярными разложениями. Это следует из некоторого произвола при построении векторов $e_1, e_2, \ldots, e_n$ и $f_1, f_2, \ldots, f_m$.

Сингулярному разложению (1.13) можно придать следующий вид:

$$A = U\Sigma_r V^*, \tag{1.14}$$

где

$$\Sigma_r = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{pmatrix} - \text{ квадратная матрица порядка } r,$$

получающаяся из $(m \times m)$-матрицы Σ вычеркиванием $n - r$ нулевых столбцов справа и $m - r$ нулевых строк снизу, U – $(m \times n)$-матрица, состоящая из первых r столбцов матрицы Q, V^* – $(r \times n)$-матрица, состоящая из первых r строк матрицы P^*.

Разложение (1.14) называют *второй формой сингулярного разложения матрицы* A. В него входят матрицы меньших размерностей, чем в первую форму, и, кроме того, в нем матрица Σ_r – квадратная невырожденная. Все это может оказаться существенным, особенно при работе с сингулярным разложением на компьютере.

При $m \geq n$ сингулярному разложению (1.13) иногда придают вид

$$A = U\Sigma_n V^*, \tag{1.15}$$

где

$$\Sigma_n = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix}$$

– квадратная матрица, состоящая из первых n строк и столбцов матрицы Σ, U – $(m \times n)$-матрица, состоящая из первых n столбцов матрицы Q, $V^* = P^*$.

Если действительная (комплексная) матрица A симметричная (унитарная), то можно добиться, чтобы (см. [19]) в сингулярном разложении $A = Q\Sigma P^*$ ортогональные (унитарные) матрицы P и Q удовлетворяли условиям $Q = P$ и $P^* = P^\top$.

При конструировании сингулярного разложения на ЭВМ его обычно получают косвенным путем (см. [14]). Стандартную программу такого метода можно найти в пакете Matlab (см., например, [9]).

Сингулярное разложение находит самое широкое применение в теории и приложениях, которые будут рассмотрены позже: при вычислении псевдообратной матрицы, при отыскании псевдорешений систем линейных алгебраических уравнений (СЛАУ) и их проекций на пространства правых сингулярных векторов, при отыскании решений неустойчивых СЛАУ, при проведении сингулярного анализа модели выравнивающей функции по методу наименьших квадратов (МНК). В [9] приведен пример применения SVD-разложения для решения задачи сжатия изображений.

Пример 1.1. Вычислить сингулярные числа для матрицы

$$A = \begin{pmatrix} -1 & -7 \\ 1 & 7 \end{pmatrix}.$$

Р е ш е н и е. Для матрицы

$$A^*A = \begin{pmatrix} -1 & 1 \\ -7 & 7 \end{pmatrix} \begin{pmatrix} -1 & -7 \\ 1 & 7 \end{pmatrix} = \begin{pmatrix} 2 & 14 \\ 14 & 98 \end{pmatrix}$$

характеристический многочлен $|A^\top A - \lambda E| = \lambda(\lambda - 100)$ имеет корни $\lambda_1 = 100$, $\lambda_2 = 0$. Поэтому $\sigma_1 = \sqrt{\lambda_1} = 10$, $\sigma_2 = 0$.

Пример 1.2. Построить сингулярное разложение матрицы

$$A = \begin{pmatrix} 4 & -3i \\ -3i & 4 \end{pmatrix}.$$

Р е ш е н и е. Характеристический многочлен

$$|A^*A - \lambda E| = \begin{vmatrix} 25 - \lambda & 0 \\ 0 & 25 - \lambda \end{vmatrix} = (25 - \lambda)^2$$

матрицы A^*A имеет корни $\lambda_1 = \lambda_2 = 25$. Поэтому $\sigma_1 = \sigma_2 = \sqrt{25} = 5$. Следовательно, матрица Σ имеет вид

$$\Sigma = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}.$$

При $\lambda = 25$ система $(A^*A - \lambda E)v = 0$, т.е. система

$$\begin{cases} 0 \cdot v_1 + 0 \cdot v_2 &= 0, \\ 0 \cdot v_1 + 0 \cdot v_2 &= 0, \end{cases}$$

имеет фундаментальную систему решений, состоящую из двух решений, например, из решений $b_1 = (1,0)^\top$, $b_2 = (0,1)^\top$. Они уже ортонормированы, поэтому $e'_1 = (1,0)^\top$, $e'_2 = (0,1)^\top$. Из столбцов координат этих векторов построим матрицу

$$P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Далее строим векторы

$$\begin{aligned} f_1 &= \frac{Ae_1}{\sigma_1} = \frac{1}{5}\begin{pmatrix} 4 & -3i \\ -3i & 4 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{5}\begin{pmatrix} 4 \\ -3i \end{pmatrix}, \\ f_2 &= \frac{Ae_2}{\sigma_2} = \frac{1}{5}\begin{pmatrix} 4 & -3i \\ -3i & 4 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{5}\begin{pmatrix} -3i \\ 4 \end{pmatrix}. \end{aligned}$$

Число этих ортонормированных векторов равно размерности пространства $\mathbb{C}^2$. Поэтому из столбцов их координат построим матрицу

$$Q = \begin{pmatrix} 4/5 & -3i/5 \\ -3i/5 & 4/5 \end{pmatrix}$$

и запишем искомое сингулярное разложение

$$A = Q\Sigma P^* = \begin{pmatrix} 4/5 & -3i/5 \\ -3i/5 & 4/5 \end{pmatrix}\begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Глава 2

Нормальные решения и псевдорешения

2.1. Псевдорешения линейных систем

2.1.1. Нормальные решения. Рассмотрим произвольную СЛАУ общего вида:

$$Au = f, \qquad A \in \mathbb{C}^{m\times n}, \qquad f \in \mathbb{C}^m. \tag{2.1}$$

В системе (2.1) m обозначает число уравнений, а n – число неизвестных.

Если $m < n$, то система (2.1) называется *недоопределенной*, если $m = n$, то в системе (2.1) число уравнений равно числу неизвестных и ее называют системой с *квадратной матрицей* коэффициентов, а в случае $m > n$ система называется *переопределенной*. Последний случай наиболее часто встречается в задачах обработки экспериментальных данных.

Условия совместности, т.е. существования решения, системы (2.1) определяются известной теоремой Кронекера - Капелли (см., например, [11]):

$$\operatorname{rank}(A\vdots f) = \operatorname{rank}(A).$$

При этом возможны два различных варианта:

1. $\operatorname{rank}(A\vdots f) = \operatorname{rank}(A) = n$ – система (2.1) имеет *единственное* решение u_0 (в случае $m = n$ единственность решения эквивалентна условию $\det A \neq 0$);

2. $\operatorname{rank}(A\vdots f) = \operatorname{rank}(A) \neq n$ – система (2.1) имеет бесчисленное множество решений (*неединственность*) $U = \{u : Au = f\}$. В этом

случае вводится понятие *нормального* решения u_*, т.е. решения с минимальной евклидовой нормой $\|u\|_2$:

$$u_* = \operatorname*{argmin}_{u \in U} \|u\|_2.$$

Утверждение 2.1. *Нормальное решение u_* совместной системы линейных алгебраических уравнений, когда последняя имеет бесчисленное множество решений, определяется единственным образом.*

Доказательство. Если система $Au = f$ имеет неединственное решение, то множество всех ее решений $U \subset \mathbb{C}^n$ есть выпуклое множество в $\mathbb{C}^n$. В самом деле, пусть u_1, u_2 – два решения системы $Au = f$. Тогда $u = tu_1 + (1-t)u_2$ при $0 \le t \le 1$ будет тоже решением:

$$Au = tAu_1 + (1-t)Au_2 = tf + (1-t)f = f.$$

Здесь использована линейность $A : \mathbb{C}^n \to \mathbb{C}^m$. Далее, рассмотрим строго выпуклый функционал на $U \quad g(u) = \|u\|_2$, ограниченный снизу $g(u) \ge 0$. Всякая минимизирующая последовательность $u^k \in U$ будет ограниченной в $\mathbb{C}^n$, так как $\|u^1\|_2 > \|u^2\|_2 > \ldots > \|u^k\|_2 > \ldots$ и, следовательно, компактной в $\mathbb{C}^n$. Поэтому существует предел $u_* = \lim_{k\to\infty} u^k$ и в силу строгой выпуклости $g(u)$ предел единственный. □

2.1.2. Псевдорешения. В случае, когда $\operatorname{rank}(A\vdots f) > \operatorname{rank}(A)$, система (2.1) не имеет решений (несовместность). Тогда вводится понятие *псевдорешения* системы (2.1), под которым понимается решение системы

$$Au = f_{\text{пр}}, \tag{2.2}$$

где $f_{\text{пр}}$ есть проекция вектора f на $\operatorname{im} A$.

Утверждение 2.2. *Для любой матрицы $A \in \mathbb{C}^{m\times n}$ и вектора $f \in \mathbb{C}^m$ ортогональная проекция $f_{пр}$ вектора f на множество значений матрицы A, т.е.* $\operatorname{im} A$, *определяется единственным образом.*

Доказательство. Образ матрицы $\operatorname{im} A$ является подпространством $\mathbb{C}^m$ в силу линейности отображения $A : \mathbb{C}^n \to \mathbb{C}^m$. Пусть $e_1, \ldots, e_q$ – ортонормированный базис в $\operatorname{im} A$, $q \le m$. Тогда, очевидно, многочлен Фурье $f_{\text{пр}} = \sum_{k=1}^{q} (f, e_k)e_k$ будет единственной ортогональной проекцией f на $\operatorname{im} A$. □

Из определения $f_{\text{пр}}$ также непосредственно следует, что

$$\operatorname{rank}(A\vdots f_{\text{пр}}) = \operatorname{rank} A,$$

т.е. система (2.2) всегда *совместна* и имеет единственное или бесконечное множество решений $U' = \{u : Au = f_{\text{пр}}\}$.

При этом также возможны два различных варианта:

1. $\operatorname{rank}(A\vdots f_{\text{пр}}) = \operatorname{rank}(A) = n$ – система (2.2) имеет *единственное* решение u_*, которое и называется *псевдорешением* системы (2.1);

2. $\operatorname{rank}(A\vdots f_{\text{пр}}) = \operatorname{rank}(A) \neq n$ – система (2.2) имеет бесчисленное множество решений (*неединственность*) $U' = \{u : Au = f\}$. В этом случае вводится понятие *нормального псевдорешения* u_*, т.е. псевдорешения с минимальной евклидовой нормой $\|u\|_2$:

$$u_* = \operatorname*{argmin}_{u \in U'} \|u\|_2.$$

Иногда, в случае несовместности систем (2.1) (или (2.2)), среди бесчисленного множества решений (или псевдорешений) ищется решение (псевдорешение), ближайшее к некоторой заданной точке $u^0 = (u_1^0, \ldots, u_n^0)^\top$, называемой *пробным решением*. Такая точка может быть известна из каких-то априорных соображений (например, из физического смысла задачи); в противном случае полагают $u^0 = 0$.

Решение уравнения (2.1) (или (2.2)), ближайшее к пробному решению u^0, называется *нормальным относительно* u^0 *решением* (или соответственно *нормальным относительно* u^0 *псевдорешением*), т.е.

$$u_* = \operatorname*{argmin}_{u \in U} \|u - u^0\|_2 \quad \text{или} \quad u_* = \operatorname*{argmin}_{u \in U'} \|u - u^0\|_2.$$

Если ввести показатель несовместности системы (2.1)

$$\mu = \inf_{u \in \mathbb{C}^n} \|Au - f\|_2,$$

то определению решения (или нормального решения) отвечает $\mu = 0$, а псевдорешению (или нормальному псевдорешению) $\mu > 0$.

2.2. Линейная задача наименьших квадратов

Понятие псевдорешения линейной системы уравнений тесным образом связано с решением *линейной задачи наименьших квадратов*.

Множество решений в смысле наименьших квадратов системы (2.1) определяется как

$$U = \{u \in \mathbb{C}^n : \|Au - f\|_2 = \min\} \tag{2.3}$$

и характеризуется следующей теоремой

Теорема 2.1. *Решение задачи (2.3) эквивалентно следующему условию ортогональности:*

$$u \in U \;\Leftrightarrow\; A^*(f - Au) = 0.$$

Доказательство. Предположим, что u удовлетворяет условию $A^*r_u = 0$, где $r_u = f - Au$. Тогда для любого вектора $v \in \mathbb{C}^n$ $r_v = f - Av = r_u + A(u - v)$. Возводя в квадрат, получаем

$$\|r_v\|_2^2 = \|r_u\|_2^2 + 2(u - v)^* \underbrace{A^*r_u}_{=0} + \|A(u - v)\|_2^2 \geq \|r_u\|_2^2.$$

Теперь предположим, что $A^*r_u = z \neq 0$. Тогда, если $u - v = -\varepsilon z$,

$$\|v\|_2^2 = \|r_u\|_2^2 - 2\varepsilon\|z\|_2^2 + \varepsilon^2\|Az\|_2^2 < \|r_u\|_2^2$$

для достаточно малых ε. □

Вектор $r = f - Au$ обозначает невязку зависящую от u. Теорема 2.1 показывает, что невязка соответствующая решению задачи наименьших квадратов ортогональна подпространству $\operatorname{im} A$.

Таким образом, правая часть системы (2.1) (вектор f) есть декомпозиция двух ортогональных компонент

$$f = Au + r, \qquad r \perp Au,$$

где символ $\perp$ означает $(r, Au) = 0$.

Данная декомпозиция всегда единственна, даже, если решение u линейной задачи наименьших квадратов (2.3) не единственно.

Из теоремы 2.1 следует, что решение линейной задачи наименьших квадратов (2.1) удовлетворяет решению *системы нормальных уравнений (нормального уравнения)*

$$A^*Au = A^*f, \tag{2.4}$$

где матрица A^*A эрмитова и неотрицательно определена, а система (2.4) совместна.

Из теоремы 2.1 также непосредственно следует, что множество псевдорешений, определенных в 2.1.2 совпадает с множеством решений линейной задачи наименьших квадратов (2.3) и соответственно с множеством решений системы нормальных уравнений (2.4), т.е.

$$\begin{aligned} U' &= \{u \in \mathbb{C}^n : Au = f_{\text{пр}}\} = \{u \in \mathbb{C}^n : \|Au - f\|_2 = \min\} \\ &= \{u \in \mathbb{C}^n : A^*Au = A^*f\}. \end{aligned}$$

Теорема 2.2. *Матрица A^*A положительно определена тогда и только тогда, когда столбцы матрицы A линейно независимы.*

Доказательство. Если столбцы матрицы A линейно независимы, тогда из $u \neq 0 \Rightarrow Au \neq 0$ и поэтому

$$u \neq 0 \;\Rightarrow\; u^*A^*Au = \|Au\|_2^2 > 0.$$

Следовательно A^*A положительно определена.

С другой стороны, если столбцы линейно зависимы, тогда для некоторого $u_0 \neq 0$ должно выполняться $Au_0 = 0$ и $u_0^*A^*Au_0 = 0$ и поэтому A^*A не является положительно определенной. □

Замечание 2.2.1. К системе нормальных уравнений (2.4) можно прийти также, используя методы математического анализа, а именно, приравняв к нулю дифференциал

$$\begin{aligned} dF(u) &= du^* \cdot A^*Au + u^*A^*A \cdot du - f^*Adu - du^* \cdot A^*f = \\ &= du^* \cdot A^*Au + du^* \cdot A^*Au - du^* \cdot A^*f - du^* \cdot A^*f = \\ &= 2du^* \cdot (A^*Au - A^*f) \end{aligned}$$

функции

$$\begin{aligned} F(u) &= \|Au - f\|_2^2 = (Au - f)^*(Au - f) = (u^*A^* - f^*)(Au - f) = \\ &= u^*A^*Au - f^*Au - u^*A^*f - f^*f. \end{aligned}$$

Замечание 2.2.2. К системе нормальных уравнений (2.4) можно также формально прийти умножив слева на матрицу A^* левую и правую части системы (2.1). Такое преобразование называется *первой трансформацией Гаусса.*

Замечание 2.2.3. Систему нормальных уравнений (2.4) и уравнения определяющие вектор невязки можно скомбинировать в совместную систему из $(m+n)$ линейных уравнений

$$\begin{pmatrix} E_m & A \\ A^* & 0 \end{pmatrix} \begin{pmatrix} r \\ u \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}. \tag{2.5}$$

Матрица системы (2.5) квадратная и эрмитова (соответственно для $A \in \mathbb{R}^{m\times n}$ – симметричная), но знаконеопределенная, если $A \neq 0$. Расширенная система (2.5) используется для итерационного уточнения псевдорешений (см., например, [2]) и в некоторых методах, где матрица A является разреженной.

Из теоремы 2.2 следует, что если $\operatorname{rank}(A) = n$, тогда существует единственное псевдорешение, которое может быть записано в виде

$$u_* = (A^*A)^{-1}A^*f,$$

и соответствующий вектор невязки

$$r_* = f - Au = (E_m - P_A)f, \qquad P_A = A(A^*A)^{-1}A^*,$$

где P_A – ортогональный проектор (матрица ортогонального проектирования) на подпространство $\operatorname{im} A$.

Если $\operatorname{rank}(A) < n$, псевдорешение не единственно (бесконечное множество псевдорешений), однако нормальное псевдорешения определяется однозначно.

2.3. Псевдообращение

Если $A \in \mathbb{C}^{n\times n}$ и $\det A \neq 0$, то для нее существует обратная матрица A^{-1}. Если же $A \in \mathbb{C}^{m\times n}$ и $m \neq n$ или $A \in \mathbb{C}^{n\times n}$, но $\det A = 0$, то матрица A не имеет обратной и символ A^{-1} не имеет смысла. Однако как будет показано далее, для произвольной прямоугольной матрицы существует "псевдообратная"матрица A^+, которая обладает некоторым свойствами обратной матрицы и имеет важные применения при решении систем линейных уравнений. В случае, когда $A \in \mathbb{C}^{n\times n}$ и $\det A \neq 0$, псевдообратная матрица A^+ совпадает с обратной A^{-1}. Приведенное в этом разделе

определение псевдообратной матрицы было дано в 1920 г. Муром, указавшим на важные применения этого понятия. Позже независимо от Мура в несколько иной форме псевдообратная матрица определялась и исследовалась в работах Пенроуза (см., например, [6]).

2.3.1. Псевдообратная матрица. Рассмотрим матричное уравнение

$$AXA = A. \tag{2.6}$$

Если $A \in \mathbb{C}^{n\times n}$ и $\det A \neq 0$, то уравнение (2.6) имеет единственное решение $X = A^{-1}$. Если же A – произвольная прямоугольная $m \times n$-матрица, то искомое решение $X \in \mathbb{C}^{n\times m}$, но не определяется однозначно. В общем случае уравнение (2.6) имеет бесчисленное множество решений, которые называются *обобщенной обратной* матрицей и обозначаются $X = A^{-}$. Обобщенная обратная матрица A^{-} обладает тем свойством, что для любого вектора $b \in \mathbb{C}^m$, при котором система уравнений $Au = b$ совместна, вектор $u = A^{-}b$ является ее решением.

Среди этих решений имеется только одно, обладающее тем свойством, что его строки и столбцы являются линейными комбинациями соответственно строк и столбцов сопряженной матрицы A^{+} [6].

Определение. Матрицу $A^{+} \in \mathbb{C}^{n\times m}$ называют *псевдообратной* (или обобщенной обратной матрицей Мура-Пенроуза) к матрице $A \in \mathbb{C}^{m\times n}$, если

$$AA^{+}A = A, \qquad A^{+} = UA^{*} = A^{*}V, \tag{2.7}$$

где U и V – некоторые матрицы.

Утверждение 2.3. *Матрица A^{+}, удовлетворяющая условиям (2.7), существует и единственна.*

Доказательство. Начнем с доказательства единственности. Пусть A_1^{+} и A_2^{+} – две различные псевдообратные матрицы. Тогда

$$AA_1^{+}A = A, \quad A_1^{+} = U_1A^{*} = A^{*}V_1$$

и

$$AA_2^{+}A = A, \quad A_2^{+} = U_2A^{*} = A^{*}V_2$$

с некоторыми матрицами U_1, V_1, U_2 и V_2. Положим $D = A_1^{+} - A_2^{+}$, $U = U_1 - U_2$, $V = V_1 - V_2$. Тогда

$$ADA = 0, \quad D = UA^{*} = A^{*}V.$$

Но $D^* = V^*A$, поэтому

$$(DA)^*(DA) = A^*D^*DA = A^*V^*ADA = 0,$$

и, значит, $DA = 0$. Отсюда, используя формулу $D^* = AU^*$, находим, что

$$DD^* = DAU^* = 0.$$

Следовательно, $A_1^+ - A_2^+ = D = 0$.

Для доказательства существования матрицы A^+ предположим сначала, что $\operatorname{rank} A = n$ ($A \in \mathbb{C}^{m\times n}$ с $m \geq n$). Покажем, что в этом случае матрица

$$A^+ = (A^*A)^{-1}A^* \tag{2.8}$$

удовлетворяет условиям (2.7).

Свойство $AA^+A = A$ из (2.7), очевидно, выполнено, поскольку

$$AA^+A = A((A^*A)^{-1}((A^*A) = A,$$

где $A^*A \in \mathbb{C}_n^>$. Равенство $A^+ = UA^*$ выполнено с $U = (A^*A)^{-1}$. Равенство же $A^+ = A^*V$ выполняется, как легко проверить, если положить $V = A(A^*A)^{-2}A^*$.

Аналогичным образом показывается, что если $\operatorname{rank} A = m$ ($A \in \mathbb{C}^{m\times n}$ с $m \leq n$), то псевдообратной к матрице A является матрица

$$A^+ = A^*(AA^*)^{-1}. \tag{2.9}$$

Для доказательства существования псевдообратной матрицы в общем случае используем тот факт, что всякую матрицу $A \in \mathbb{C}^{m\times n}$ можно представить в виде произведения

$$A = B \cdot C \tag{2.10}$$

с матрицами $B \in \mathbb{C}^{m\times r}$ и $C \in \mathbb{C}^{r\times n}$, где $r = \operatorname{rank} A \leq \min(m, n)$.

Действительно, возьмем в качестве матрицы B матрицу, составленную из r независимых столбцов матрицы A. Тогда все столбцы матрицы A можно выразить через столбцы матрицы B, о чем и свидетельствует формула (2.10), задающая "скелетное" разложение матрицы A.

Положим теперь

$$A^+ = C^+B^+,$$

где согласно (2.8) и (2.9)

$$C^+ = C^*(CC^*)^{-1}, \quad B^+ = (B^*B)^{-1}B^*.$$

Тогда

$$AA^+A = BCC^*(CC^*)^{-1}(B^*B)^{-1}B^*BC = BC = A.$$

Далее, если положить $U = C^*(CC^*)^{-1}(B^*B)^{-1}C$, то легко проверить, что $UA^* = A^+$.

Аналогичным образом проверяется, что $A^+ = A^*V$ с

$$V = B(B^*B)^{-1}(CC^*)^{-1}(B^*B)^{-1}B. \qquad \square$$

Таким образом, для любой матрицы $A \in \mathbb{C}^{m\times n}$ псевдообратная матрица существует и единственная, причем для невырожденной квадратной матрицы A псевдообратная матрица $A^+ = A^{-1}$.

2.3.2. Характеризация Пенроуза. В своей работе 1955 г., которая, по всей вероятности, возродила интерес к обобщенному обращению, Пенроуз [23] характеризовал псевдообратную матрицу как (единственное) решение совокупности матричных уравнений. Псевдообратная матрица A^+, введенная выше, удовлетворяет условиям Пенроуза.

Утверждение 2.4. *Для любой матрицы $A \in \mathbb{C}^{m\times n}$ $X = A^+$, если и только если*

1. $AXA = A$;
2. $XAX = X$;
3. $(AX)^* = AX \quad$ и $\quad (XA)^* = XA$.

Доказательство имеется в [3], [1].

2.3.3. Псевдообращение и псевдорешения линейных систем.

Утверждение 2.5. *Псевдообратная матрица A^+ является наилучшим приближением (по методу наименьших квадратов) матричного уравнения*

$$AX = E_m. \tag{2.11}$$

Утверждение 2.5 можно также сформулировать в эквивалентном виде

Утверждение 2.6. *Псевдообратной матрицей для матрицы A является матрица $A^+ \in \mathbb{C}^{n\times m}$, столбцы которой нормальные псевдорешения системы линейных уравнений вида*

$$Ax = e_i, \qquad i = 1, \ldots, m, \tag{2.12}$$

где e_i – столбцы единичной матрицы E_m.

Доказательство см., например, в [6, 5, 20]. Это свойство псевдообратной матрицы, также может быть принято в качестве ее определения.

Утверждение 2.7. *Пусть $u_*^{(1)}$ и $u_*^{(2)}$ – нормальные псевдорешения двух систем линейных уравнений $Au = f^{(1)}$ и $Au = f^{(2)}$. Тогда $\beta u_*^{(1)} + \gamma u_*^{(2)}$ является нормальным псевдорешением системы $Au = \beta f^{(1)} + \gamma f^{(2)}$.*

Доказательство. Из $A^*Au_*^{(1)} = A^*f^{(1)}$ и $A^*Au_*^{(2)} = A^*f^{(2)}$ следует, что $\beta u_*^{(1)} + \gamma u_*^{(2)}$ удовлетворяет нормальной системе

$$A^*A(\beta u_*^{(1)} + \gamma u_*^{(2)}) = A^*(\beta f^{(1)} + \gamma f^{(2)}).$$

Далее, существуют столбцы $z^{(1)}$ и $z^{(2)}$ такие, что $u_*^{(1)} = A^*z^{(1)}$ и $u_*^{(2)} = A^*z^{(2)}$. Поэтому $\beta u_*^{(1)} + \gamma u_*^{(2)} = A^*(\beta z^{(1)} + \gamma z^{(2)}))$. □

Естественно, утверждение 2.7 может быть распространено на линейные комбинации произвольного числа столбцов.

Утверждение 2.8. *Псевдорешение системы линейных уравнений (2.1) может быть записано в виде $u_* = A^+f$.*

Действительно, столбец свободных членов f представляет собой линейную комбинацию столбцов матрицы E_m:

$$f = \beta_1 e_1 + \ldots + \beta_m e_m.$$

По определению псевдообратной матрицы и согласно утверждению 2.7 псевдорешение u_* есть линейная комбинация столбцов a_i^+ псевдообратной матрицы с теми же коэффициентами

$$u_* = \beta_1 a_1^+ + \ldots + \beta_m a_m^+.$$

Это равносильно доказываемому утверждению.

Псевдообратная матрица обладает следующим экстремальным свойством.

Утверждение 2.9. *Для любой матрицы $X \in \mathbb{C}^{n\times m}$ выполнено соотношение*

$$\|AA^+ - E_m\|_E \le \|AX - E_m\|_E.$$

При этом, если для какой-нибудь матрицы X, отличной от A^+, здесь имеет место равенство, то $\|A^+\|_E < \|X\|_E$.

Доказательство. По определению при любом i столбец a_i^+ псевдообратной матрицы дает минимальную невязку при подстановке в (2.12). Поэтому для i-го столбца матрицы X

$$\|Aa_i^+ - e_i\| \le \|Ax_i - e_i\|.$$

Если же тут при $x_i \ne a_i^+$ достигается равенство, то $\|a_i^+\| < \|x_i\|$. Заметим, что квадрат евклидовой нормы матрицы равен сумме квадратов ее столбцов. Следовательно, возводя в квадрат и суммируя приведенные соотношения по всем $i = 1, \dots, m$, приходим к доказываемому утверждению. □

Утверждение 2.10. *Нормальное относительно u^0 псевдорешение системы (2.1) определяется формулой*

$$u_* = A^+ f + (E_n - A^+ A)u^0.$$

Доказательство. Согласно утверждению 2.8 столбец $A^+ f$ – нормальное псевдорешение и, следовательно, является частным решением системы (2.4). Остается доказать, что столбец $z = (E_n - A^+ A)u^0$ при произвольном u^0 – общее решение нормальной однородной системы $A^* Az = 0$. Докажем это.

Во-первых, для любого u^0

$$A^* A[(E_n - A^+)u^0] = A^* Au^0 - A^* AA^+ Au^0 = A^* Au^0 - A^* Au^0 = 0.$$

Это означает, что z – решение нормальной однородной системы.

Во-вторых, для любого решения z системы $A^* Az = 0$ найдется столбец u^0, при котором

$$z = (E_n - A^+ A)u^0.$$

В действительности можно просто положить $u^0 = z$, так как система $A^* Az = 0$ равносильна системе $Az = 0$, и потому

$$(E_n - A^* A)z = z - A^+ Az = z. \qquad \square$$

Замечание 2.3.1. Утверждения 2.8 и 2.10 имеют главным образом теоретическое значение, как и правило Крамера для невырожденных матриц. Нахождение псевдообратной матрицы не обязательно для вычисления нормального псевдорешения и требует больших вычислительных затрат.

2.3.4. Псевдообращение при помощи предельного перехода.

Теорема 2.3. *Имеют место соотношения*

$$\lim_{\lambda\to 0}(A^*A+\lambda^2E_n)^{-1}A^*=A^+ \tag{2.13}$$

и

$$\lim_{\lambda\to 0}A^*(AA^*+\lambda^2E_m)^{-1}=A^+. \tag{2.14}$$

Доказательство имеется в [1], [3].

Из теоремы 2.3 непосредственно получаем

Следствие 2.3.1. *Для матриц полного ранга* ($\operatorname{rank}A=\min(m,n)$) *имеют место соотношения*

$$A^+=\begin{cases}(A^*A)^{-1}A^*, & \text{если } \operatorname{rank}A=n;\\ A^*(AA^*)^{-1}, & \text{если } \operatorname{rank}A=m.\end{cases}$$

Рассмотрим систему линейных уравнений

$$(A^*A+\alpha E_n)u=A^*f,\qquad \alpha>0. \tag{2.15}$$

Обозначим ее решение через u_α. Тогда

$$u_\alpha=(A^*A+\alpha E_n)^{-1}A^*f \tag{2.16}$$

и следствие 2.3.1 показывают, что справедливо

Утверждение 2.11. *При $\alpha\to 0$ решение u_α системы (2.15) стремиться к нормальному псевдорешению системы $Au=f$.*

Это утверждение имеет существенное теоретическое и прикладное значение. Дело в том, что нормальное псевдорешение системы линейных уравнений не является непрерывной функцией от матрицы системы. Утверждение 2.13 показывает, что система может быть включена в семейство систем с параметром α таким образом, что решение системы непрерывно зависит от параметра. Этот результат получен с более общей точки зрения в теории регуляризирующих функцтоналов для некорректно поставленных задач (см. Тихонов и Арсенин [18]). Упомянутая теория в основном относится к уравнениям в бесконечномерных пространствах (например, интегральным и диффренциальным уравнениям в частных производных).

В конечномерном случае прямой необходимости во введении регуляризирующих функционалов нет. Однако отметим, что для СЛАУ роль регуляризирующего функционала может играть функция

$$F_\alpha(u, f, A) = \|f - Au\|_2^2 + \alpha^2\|u\|_2^2$$

на арифметическом пространстве $\mathbb{C}^n$. Найдем значение u, при котором она достигает минимума. Иначе F_α можно записать так:

$$F_\alpha(u, f, A) = (f - Au)^*(f - Au) + \alpha u^* u.$$

Дифференцирую это выражение по u, находим

$$dF_\alpha(u, f, A) = -2du^* A^* f + 2u^* A^* Au + 2\alpha du^* u.$$

Дифференциал обращается в нуль для векторов u, удовлетворяющих системе уравнений, в точности совпадающей с (2.15). Как мы видели выше, детерминант матрицы системы отличен от нуля, и система имеет единственное решение (2.16) при любых f, A и $\alpha \neq 0$.

Обозначим это решение через u_λ, и пусть $F_\alpha(u, f, A) = \xi$. Если $\|u\| > \sqrt{\xi/\alpha}$, то $F_\alpha(u, f, A) > \xi$. Поэтому на сфере радиуса $\sqrt{\xi/\alpha}+1$ и вне ее F_α принимает значения, большие чем ξ. Если u_α не попало внутрь сферы, увеличим ее радиус до $\|u_\alpha\| + 1$. Так мы получим сферу, содержащую u_α и такую, что на ней и вне ее $F_\alpha(u, f, A) > \xi$. $F_\alpha(u, f, A) > \xi$. Функция непрерывна и внутри сферы имеет единственную стационарную точку. Поэтому эта точка является точкой минимума. Приведенные рассуждения показывают, что это будет абсолютный минимум.

Утверждение 2.15, по существу, утверждает, что при $\alpha \to 0$ точка, где регуляризирующий функционал достигает минимума, стремиться к псевдорешению системы $Au = f$.

2.3.5. Свойства псевдообратной матрицы. Отметим следующие основные свойства псевдообратной матрицы:

1. $(A^*)^+ = (A^+)^*$;
2. $(A^+)^+ = A$;
3. $(AA^+)^2 = AA^+, \quad (A^+A)^2 = A^+A.$

Первое свойство означает, что операция перехода к сопряженной матрице и к псевдообратной матрице перестановочны между собой. Равенство 2 выражает собой взаимность понятия псевдообратной матрицы, так как согласно 2 псевдообратной матрицей для A^+ является исходная матрица A. Согласно равенствам 3 матрицы AA^+ и A^+A являются

эрмитовыми и *инволютивными* (квадрат каждой из этих матриц равен самой матрице).

Много других свойств псевдообратных матриц имеется в [1].

2.4. Вычисление псевдообратных матриц

В этом разделе приведены некоторые достаточно простые, но практически важные численные алгоритмы нахождения псевдообратных матриц. Рассмотрены три численных метода нахождения псевдообратных матриц: алгоритм Гревилля, итерационный метод Бен-Израэля и метод, использующий сингулярное разложение матрицы (SVD-разложение).

2.4.1. Метод Гревилля. Этот метод не требует вычисления детерминантов и может быть также использован для вычисления обратной матрицы A^{-1} (в случае, если $A \in \mathbb{C}^{n\times n}$ и $|A| \neq 0$). Метод Гревилля последовательного нахождения псевдообратной матрицы состоит в следующем. Пусть a_k – k-й столбец в матрице $A \in \mathbb{C}^{m\times n}$, $A_k = (a_1, \ldots, a_k) \in \mathbb{C}^{m\times k}$ – матрица, образованная первыми k столбцами матрицы A, b_k – последняя строка в матрице A_k^+ $(k = 1, \ldots, n,\ A_1 = a_1, A_n = A)$. Тогда

$$A_1^+ = a_1^+ = \begin{cases} (a_1^* a_1)^{-1} a_1^*, & \text{если } a_1 \neq 0, \\ 0, & \text{если } a_1 = 0, \end{cases}$$

и для $k > 1$ имеют место рекуррентные формулы

$$A_k^+ = \begin{pmatrix} B_k \\ b_k \end{pmatrix}, \quad B_k = A_{k-1}^+ - d_k b_k, \quad d_k = A_{k-1}^+ a_k,$$

$$b_k = c_k^+ = \begin{cases} (c_k^* c_k)^{-1} c_k^*, & \text{если } c_k \neq 0, \\ (1 + d_k^* d_k) d_k^* A_{k-1}^+, & \text{если } c_k = 0, \end{cases}$$

где $c_k = a_k - A_{k-1} d_k$, $A_{k-1} \in \mathbb{C}^{m\times(k-1)}$, $A_{k-1}^+ \in \mathbb{C}^{(k-1)\times m}$, $d_k \in \mathbb{C}^{k-1}$ и b_k – вектор-строка размерности m.

Матрица A_k^+, построенная по этим формулам, является псевдообратной к матрице A_k, $k = 1, 2, \ldots, n$. В частности, $A_n^+ = A^+$.

2.4.2. Метод Бен-Израэля. При вычислении псевдообратной матрицы A^+ к действительной матрице $A \in \mathbb{R}^{m\times n}$ можно пользоваться итерационной формулой Бен-Израэля

$$X^{(k+1)} = X^{(k)} \left[2E_m - AX^{(k)}\right], \quad X^{(0)} = \alpha A^\top, \quad k = 1, 2, \ldots . \qquad (2.17)$$

Если α – число, удовлетворяющее условию

$$0 < \alpha < \frac{2}{\lambda_{\max}},$$

где $\lambda_{\max} = \lambda_{\max}(A^\top A) = \lambda_{\max}(AA^\top)$, то

$$\lim_{k\to 0} \|X^{(k)} - A^+\| = 0.$$

Вместо $\lambda_{\max}$ можно брать какую-либо норму этих матриц, например, $\|A\|_1$ или $\|A\|_\infty$. При известном $\lambda_{\max}$ обычно на практике полагают $\alpha = 1.6/\lambda_{\max}$.

Для остановки итерационного алгоритма (2.17), можно воспользоваться критерием

$$\frac{\left|\|AX^{(k+1)} - E_m\|_E - \|AX^{(k)} - E_m\|_E\right|}{\|AX^{(k)} - E_m\|_E} \le \delta, \tag{2.18}$$

где δ – заданное (достаточно малое) число. В этом случае $X^{(k)}$, удовлетворяющее (2.18), принимается за приближенное значение A^+. Это правило остановки итерационного алгоритма основано на экстремальном свойстве псевдообратной матрицы, сформулированном в утверждении 2.9.

Иногда, возможно использовать, более простое по числу арифметических операций, правило остановки:

$$\frac{\|X^{(k+1)} - X^{(k)}\|}{\|X^{(k)}\|} \le \delta', \tag{2.19}$$

где $\|\cdot\|$ – какая-либо матричная норма, δ' – заданное малое число. В некоторых случаях, целесообразно требовать одновременного выполнение условий (2.18) и (2.19).

2.4.3. Метод основанный на сингулярном разложении матриц. Как показано в разделе 1.4 для любой матрицы $A \in \mathbb{C}^{m\times n}$ существует сингулярное разложение (SVD-разложение) вида (1.14). Тогда из утверждения 2.9 непосредственно следует, что псевдообратная матрица

$$A^+ = V\Sigma_r^+ U^*,$$

где $\Sigma_r^+ = \operatorname{diag}\left(\sigma_1^{-1}, \ldots, \sigma_r^{-1}\right)$.

Данный метод является самым надежным (по точности) способом определения псевдообратных матриц, но в тоже время он является самым трудоемким (с вычислительной точки зрения) методом.

2.5. Типовые примеры

Пример 2.5.1. Рассмотрим систему, состоящую из одного уравнения:

$$a_1u_1 + a_2u_2 = f \quad (u_1^2 + u_2^2 \neq 0,\ a_1, a_2, f \in \mathbb{R}). \tag{2.20}$$

Требуется найти нормальное решение этой системы.

Р е ш е н и е. Для нахождения нормального решения u_* системы (2.20) можно воспользоваться псевдообратной матрицей,

$$u_* = A^+ f,$$

где $A = (a_1, a_2) \in \mathbb{R}^{1\times 2}$, т.е. матрица A – полного строкового ранга.

Используя свойство псевдообратной матрицы (из следствия 2.3.1) имеем

$$A^+ = A^\top (AA^\top)^{-1} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \left[\begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \right]^{-1} = \frac{1}{a_1^2 + a_2^2} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}.$$

Отсюда нормальное решение системы (2.20) равно

$$u_* = \left(\frac{a_1 f}{a_1^2 + a_2^2}, \frac{a_2 f}{a_1^2 + a_2^2} \right)^\top.$$

Пример 2.5.2. Пусть

$$\left(A = \begin{matrix} 1 & -1 & 0 \\ -1 & 2 & 1 \\ 2 & -3 & -1 \\ 0 & 1 & 1 \end{matrix} \right).$$

Вычислить псевдообратную матрицу A^+ с помощью метода Гревилля.

Р е ш е н и е.

$$A_1^+ = (A_1^\top A_1)^{-1} = 1/6 \cdot A_1^\top = (1/6,\ -1/6,\ 1/3,\ 0),$$

$$d_2 = A_1^+ a_2 = -3/2, \quad c_2 = a_2 - A_1 d_2 = \begin{pmatrix} 1/2 \\ 1/2 \\ 0 \\ 1 \end{pmatrix},$$

$$b_2 = c_2^+ = (c_2^\top c_2)^{-1} c_2^\top = (1/3,\ 1/3,\ 0,\ 2/3),$$

$$B_2 = A_1^+ - d_2 b_2 = (2/3,\ 1/3,\ 1/3,\ 1).$$

Таким образом,

$$A_2^+ = \begin{pmatrix} 2/3 & 1/3 & 1/3 & 1 \\ 1/3 & 1/3 & 0 & 2/3 \end{pmatrix}.$$

Далее

$$d_3 = A_2^+ a_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{и} \quad c_3 = a_3 - A_2 d_3 = 0.$$

Поэтому

$$b_3 = (1 + d_3^\top d_3)^{-1} d_3^\top A_2^+ = (1/3,\ 1/3) A_2^+ = (1/3,\ 2/9,\ 1/9,\ 5/9)$$

и

$$B_3 = A_2^+ - d_3 b_3 = \begin{pmatrix} 2/3 & 1/3 & 1/3 & 1 \\ 1/3 & 1/3 & 0 & 2/3 \end{pmatrix} -$$

$$- \begin{pmatrix} 1/3 & 2/9 & 1/9 & 5/9 \\ 1/3 & 2/9 & 1/9 & 5/9 \end{pmatrix} = \begin{pmatrix} 1/3 & 1/9 & 2/9 & 4/9 \\ 0 & 1/9 & -1/9 & 1/9 \end{pmatrix},$$

$$A^+ = A_3^+ = \begin{pmatrix} 1/3 & 1/9 & 2/9 & 4/9 \\ 0 & 1/9 & -1/9 & 1/9 \\ 1/3 & 2/9 & 1/9 & 5/9 \end{pmatrix}.$$

Пример 2.5.3. Найти псевдообратную матрицу A^+ к матрице

$$A = \begin{pmatrix} 1 & i \\ -i & 1 \\ 1 & 1 \end{pmatrix}.$$

Р е ш е н и е. Ранг матрицы A равен числу ее столбцов. Поэтому применима формула из следствия 2.3.1. По ней получаем

$$A^+ = (A^* A)^{-1} A^* =$$

$$= \left(\begin{pmatrix} 1 & i & 1 \\ -i & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & i \\ -i & 1 \\ 1 & 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & i & 1 \\ -1 & 1 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 3 & 1+2i \\ 1-2i & 3 \end{pmatrix}^{-1} \begin{pmatrix} 1 & i & 1 \\ -i & 1 & 1 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 3 & -1-2i \\ -1+2i & 3 \end{pmatrix} \times$$

$$\times \begin{pmatrix} 1 & i & 1 \\ -i & 1 & 1 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 1+i & -1+i & 2-2i \\ -1-i & 1-i & 2+2i \end{pmatrix}.$$

Пример 2.5.4. Найти псевдообратную матрицу A^+ к матрице

$$A = \begin{pmatrix} 1 & -i & 1 \\ i & 1 & 1 \end{pmatrix}.$$

Р е ш е н и е. Ранг матрицы A равен числу ее строк. Поэтому применима формула из следствия 2.3.1. По ней получаем

$$A^+ = A^*(AA^*)^{-1} =$$

$$= \begin{pmatrix} 1 & i \\ -i & 1 \\ 1 & 1 \end{pmatrix} \left(\begin{pmatrix} 1 & i & 1 \\ -i & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & i \\ -i & 1 \\ 1 & 1 \end{pmatrix} \right)^{-1} =$$

$$= \begin{pmatrix} 1 & i \\ -i & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 3 & 1+2i \\ 1-2i & 3 \end{pmatrix}^{-1} = \frac{1}{4} \begin{pmatrix} 1 & i \\ -i & 1 \\ 1 & 1 \end{pmatrix} \times$$

$$\times \begin{pmatrix} 3 & -1-2i \\ -1+2i & 3 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 1+i & -1-i \\ -1+i & 1-i \\ 2-2i & 2+2i \end{pmatrix}.$$

Глава 3

Вычисление псевдорешений

3.1. Определение множества чисел с плавающей точкой

Множество $\mathbb{F}$ чисел с плавающей точкой характеризуется четырьмя параметрами: основанием системы счисления β, разрядностью p и интервалом показателей $[\nu^-, \nu^+]$. Каждое число $x \in \mathbb{F}$ представимо в виде

$$x = \pm \left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \ldots + \frac{d_p}{\beta^p} \right) \cdot \beta^\nu,$$

где целые числа β, ν, $d_1, \ldots, d_p$ удовлетворяют неравенствам

$$0 \le d_i \le \beta - 1, \quad i = 1, \ldots, p; \quad \nu^- \le \nu \le \nu^+.$$

Часто d_i называют *разрядами*, p – *длиной мантиссы*, ν – *порядком числа*. *Мантиссой* (дробной частью) x называется число в скобках.

Удобно считать, что *округление* – это некоторое отображение множества действительных чисел $\mathbb{R}$ в множество $\mathbb{F}$ чисел с плавающей точкой. Если y – такое действительное число, что результат отображения $fl(y) \in \mathbb{F}$, то имеет место аксиома

$$fl(y) = y(1 + \eta),$$

где в случае $fl(y) \neq 0$ $|\eta| \le \varepsilon_1$. Будем считать, что ε_1 есть точная верхняя грань для $|\eta|$. При традиционном способе округления чисел имеем $\varepsilon_1 = \frac{1}{2}\beta^{1-p}$, при округлении отбрасыванием разрядов $\varepsilon_1 = \beta^{1-p}$. Величину ε_1 часто называют *машинной точностью* или *машинным эпсилон* и обозначают *macheps* или ε_{mach}.

Величину ε_1 можно оценить непосредственно в ходе вычислительного процесса [2]. Для этого достаточно включить в программу фрагмент включающий следующий метод. Полагая $\varepsilon^{(0)} = 1$, следует вычислять последовательно $\varepsilon^{(1)} = 0.5\varepsilon^{(0)}$, $\varepsilon^{(2)} = 0.5\varepsilon^{(1)}, \ldots, \varepsilon^{(n)} = 0.5\varepsilon^{(n-1)}, \ldots$, проверяя каждый раз выполнение неравенства $1+\varepsilon^{(n)} > 1$. Как только при некотором n окажется, что $1 + \varepsilon^{(n)} = 1$, следует положить $\varepsilon_1 = \varepsilon^{(n-1)}$. Хотя полученное таким способом значение может отличаться от ε_1 в 2 раза, обычно оно используется так, что эта погрешность не имеет значения.

Оценим погрешность размещения вещественной матрицы в памяти ЭВМ. Пусть $A \in \mathbb{R}^{m\times n}$ – исходная матрица. Через $fl(A)$ обозначен результат размещения матрицы A в памяти ЭВМ. Для элементов a_{ij} и $fl(a_{ij})$ матриц A и $fl(A)$ соответственно выполнено неравенство [14]

$$|fl(a_{ij}) - a_{ij}| \le \max(\varepsilon_0, \varepsilon_1|a_{ij}|) \le \varepsilon_0 + \varepsilon_1|a_{ij}|, \quad \text{где} \quad \varepsilon_0 = \beta^{\nu^--1}.$$

Поэтому очевидно, что

$$\|fl(A) - A\|_E \le \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}(\varepsilon_0 + \varepsilon_1|a_{ij}|)^2} \le \varepsilon_0\sqrt{mn} +$$

$$+\varepsilon_1\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}a_{ij}^2} = \varepsilon_1\|A\|_E + \varepsilon_0\sqrt{mn}. \tag{3.1}$$

Оценка (3.1) по сравнению с реальной погрешностью может быть слегка завышена, но для анализа ошибок округления достаточно эффективна.

3.2. Обусловленность и числа обусловленности

В этом разделе книги систематически изучается одно из фундаментальных понятий современного численного анализа – обусловленность задачи. Обусловленность является характеристикой чувствительности к возмущениям исходных данных в математической задаче. Устойчивость определяет погрешность решения математической задачи на компьютере, вызванную конечной разрядностью машинной арифметики.

3.2.1. Обусловленность задачи

Любая математическая задача в дальнейшем будет рассматриваться как некоторая функция (отображение) $f : X \to Y$ из некоторого нормированного векторного пространства исходных данных X в нормированное векторное пространство Y решений. Эта функция f обычно нелинейная (даже в линейной алгебре), но в большинстве случаев является непрерывной.

Хорошо обусловленная задача, например, обладает свойством, что для всех достаточно малых возмущений в векторе исходных данных $x \in X$ изменения в $f(x)$ будут также достаточно малыми. *Плохо обусловленная* задача, когда достаточно малые возмущения в x будут приводить к большим изменениям $f(x)$.

3.2.2. Абсолютное число обусловленности

Пусть δx обозначает малое возмущение в x, тогда $\delta f = f(x+\delta x) - f(x)$. *Абсолютное число обусловленности* $\hat{\kappa} = \hat{\kappa}(x)$ задачи f на x определяется как

$$\hat{\kappa} = \lim_{\delta \to 0} \sup_{\|\delta x\| \leq \delta} \frac{\|\delta f\|}{\|\delta x\|}. \tag{3.2}$$

Для большинства задач, предел верней грани в этой формуле можно интерпретировать как верхнюю грань всех бесконечно малых возмущений δx, и в интересах читателей записать эту формулу просто как

$$\hat{\kappa} = \sup_{\delta x} \frac{\|\delta f\|}{\|\delta x\|}, \tag{3.3}$$

в предположении, что δx и δf бесконечномалые величины.

Если f дифференцируемая, то число обусловленности можно определить через соответствующие производные от f. Пусть $J(x)$ матрица (i,j) - элементами которой являются частные производные $\partial f_i / \partial x_j$. Матрица $J(x)$ известна как *Якобиан* функции f по x. Так как для производных первого порядка $\delta f \approx J(x)\delta x$, то в пределе при $\|\delta x\| \to 0$ получаем, что абсолютное число обусловленности равно

$$\hat{\kappa} = \|J(x)\|, \tag{3.4}$$

где $\|J(x)\|$ матричная норма для $J(x)$ индуцирована нормами в X и Y.

3.2.3. Относительное число обусловленности

Для характеристики относительных ошибок решений, вызванных относительными величинами возмущений исходных данных, вводится понятие *относительного числа обусловленности*, $\kappa = \kappa(x)$ определяемого как

$$\kappa = \lim_{\delta \to 0} \sup_{\|\delta x\| \le \delta} \left(\frac{\|\delta f\|}{\|f(x)\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right), \tag{3.5}$$

или, полагая δx и δf бесконечно малыми,

$$\kappa = \sup_{\delta x} \left(\frac{\|\delta f\|}{\|f(x)\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right). \tag{3.6}$$

Если f дифференцируемая, то относительное число обусловленности можно определить с использованием Якобиана:

$$\kappa = \frac{\|J(x)\|}{\|f(x)\|/\|x\|}. \tag{3.7}$$

Абсолютное и относительное числа обусловленности имеют различные применения, но наиболее важное значение они имеют в численном анализе. Это связано с тем фактом, что в компьютерных вычислениях в основном используется арифметика с плавающей точкой. Задача *хорошо обусловленная* если κ мало (например, 1, 10, 10^2), и *плохо обусловленная* если κ большое (например, 10^6, 10^{16}).

Примеры

Пример 3.2.1. Рассмотрим простейшую задачу вычисления скаляра $x/2$ из $x \in \mathbb{C}$. Якобиан J функции $f : x \mapsto x/2$ равен производной $f' = 1/2$. Тогда из (3.7) получаем

$$\kappa = \frac{\|J\|}{\|f(x)\|/\|x\|} = \frac{1/2}{(x/2)/x} = 1.$$

Эта задача хорошо обусловленная для всех $x \in \mathbb{C}$.

Пример 3.2.2. Рассмотрим задачу вычисления $\sqrt{x}$ для $x > 0$. Якобиан функции $f : x \mapsto \sqrt{x}$ есть производная $J = f' = 1/(2\sqrt{x})$. Тогда получаем

$$\kappa = \frac{\|J\|}{\|f(x)\|/\|x\|} = \frac{1/(2\sqrt{x})}{\sqrt{x}/x} = \frac{1}{2}.$$

Таким образом, снова имеем хорошо обусловленную задачу.

Пример 3.2.3. Рассмотрим задачу определения скаляра $f(x) = x_1 - x_2$ из вектора $x = (x_1, x_2)^* \in \mathbb{C}^2$. Для простоты воспользуемся кубической нормой (∞-нормой) в арифметическом пространстве $\mathbb{C}^2$. Якобиан функции f равен

$$J = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} & \dfrac{\partial f}{\partial x_2} \end{bmatrix} = [1 \quad -1],$$

с $\|J\|_\infty = 2$. Число обусловленности тогда равно

$$\kappa = \frac{\|J\|_\infty}{\|f(x)\|_\infty / \|x\|_\infty} = \frac{2}{|x_1 - x_2| / \max\{|x_1|, |x_2|\}}.$$

Следовательно, если $|x_1 - x_2| \approx 0$, то данная задача плохо обусловленная если $x_1 \approx x_2$. Интуитивно ясно, что вычитание близких чисел может привести к «катастрофически» большой ошибке.

Пример 3.2.4. Задача вычисления собственных значений несимметричной матрицы часто также оказывается плохо обусловленной. Это можно видеть на примере рассмотрения двух матриц

$$\begin{pmatrix} 1 & 1000 \\ 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 1000 \\ 0,001 & 1 \end{pmatrix},$$

с собственными значениями $\{1, 1\}$ и $\{0, 2\}$ соответственно. Наоборот, если матрица A симметричная (вообще, даже, если нормальная), тогда задача вычисления собственных значений хорошо обусловленная. Можно показать, если λ и $\lambda + \delta\lambda$ соответственно собственные значения матриц A и $A + \delta A$, тогда $|\delta\lambda| \leq \|\delta A\|_2$. Причем равенство имеет место, если δA скалярная матрица. Тогда абсолютное число обусловленности симметричной проблемы собственных значений $\hat{\kappa} = 1$, если возмущения измерять в квадратичной (l_2) норме, и относительное число обусловленности $\kappa = \|A\|_2 / |\lambda|$.

3.2.4. Обусловленность матрично-векторного умножения

Рассмотрим число обусловленности фундаментальной и важной для вычислительной линейной алгебры задачи.

Пусть задана матрица $A \in \mathbb{C}^{m \times n}$ и рассмотрим задачу вычисления Ax на векторе x, т.е. определим число обусловленности соответствующее возмущениям вектора x при фиксированной матрице A. Число

обусловленности κ определим для произвольной векторной нормы $\|\cdot\|$ и соответствующей ей индуцированной матричной нормы. Тогда

$$\kappa = \sup_{\delta x} \left(\frac{\|A(x+\delta x) - Ax\|}{\|Ax\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right) = \sup_{\delta x} \frac{\|A\delta x\|}{\|\delta x\|} \Big/ \frac{\|Ax\|}{\|x\|},$$

т.е.

$$\kappa = \|A\| \frac{\|x\|}{\|Ax\|} \tag{3.8}$$

(для случая (3.7)). Эта точная формула для κ зависит как от A, так и от x.

Предположим, что матрица A квадратная и невырожденная. Тогда используя тот факт, что $\|x\|/\|A\| \leq \|A^{-1}\|$, можно для (3.8) получить верхнюю границу независящую от x:

$$\kappa \leq \|A\|\|A^{-1}\|. \tag{3.9}$$

Или можно также записать в виде

$$\kappa = \alpha \|A\|\|A^{-1}\| \tag{3.10}$$

с

$$\alpha = \frac{\|x\|}{\|Ax\|} \Big/ \|A^{-1}\|. \tag{3.11}$$

Наверняка, при некоторых x можно получить $\alpha = 1$, и, следовательно, $\kappa = \|A\|\|A^{-1}\|$. Если $\|\cdot\| = \|\cdot\|_2$, тогда данное равенство имеет место если x правый сингулярный вектор соответствующий минимальному сингулярному числу матрицы A.

Пусть матрица A неквадратная. Если $A \in \mathbb{C}^{m\times n}$ с $m \geq n$ и имеет полный ранг, $\operatorname{rank} A = n$, то в выражениях (3.9) – (3.11) матрицу A^{-1} следует заменить на псевдообратную матрицу A^{+}.

Рассмотрим обратную задачу: дана матрица A, вычислить $A^{-1}b$ при входном векторе b. Математически эта задача идентична рассмотренной ранее задаче векторно-матричного умножения, когда вместо матрицы A используется матрица A^{-1}. Эти результаты можно объединить в следующую теорему.

Теорема 3.1. *Пусть $A \in \mathbb{C}^{m\times m}$ невырождена и рассмотрим уравнение $Ax = b$. Задача вычисления b, при данном x, имеет число обусловленности*

$$\kappa = \|A\| \frac{\|x\|}{\|b\|} \leq \|A\|\|A^{-1}\|, \tag{3.12}$$

соответствующее возмущениям в x. Задача вычисления x, при данном b, имеет число обусловленности

$$\kappa = \|A\|\frac{\|b\|}{\|x\|} \leq \|A\|\|A^{-1}\|, \tag{3.13}$$

соответствующее возмущениям в b. Если $\|\cdot\| = \|\cdot\|_2$, тогда равенство в (3.12) достигается на x пропорциональном правому сингулярному вектору матрицы A, соответствующем ее минимальному сингулярному значению σ_m, и равенство в (3.13) достигается на b пропорциональном левому сингулярному вектору матрицы A, соответствующему ее максимальному сингулярному значению σ_1.

3.2.5. Число обусловленности матрицы

Величина $\|A\|\|A^{-1}\|$ часто имеет другое название: ее называют *числом обусловленности матрицы* A (относительно нормы $\|\cdot\|$) и определяется:

$$\kappa(A) = \mathrm{cond}(A) = \|A\|\|A^{-1}\|. \tag{3.14}$$

Так в этом случае термин «число обусловленности» относится к матрице, но не к задаче. Если $\kappa(A)$ мало, то говорят, что матрица A *хорошо обусловенная*; если $\kappa(A)$ большое – A *плохо обусловленная*. Если A вырожденная, то считаем $\kappa(A) = \infty$.

Заметим, что если $\|\cdot\| = \|\cdot\|_2$, тогда $\|A\| = \sigma_1$ и $\|A^{-1}\| = 1/\sigma_m$. Тогда

$$\kappa(A) = \frac{\sigma_1}{\sigma_m} \tag{3.15}$$

в евклидовой векторной норме. Эта формула называется *спектральным числом обусловленности матриц*. Отношение σ_1/σ_m можно интерпретировать как эксцентриситет гиперэллипсоида, являющегося образом единичной сферы из $\mathbb{C}^n$ на $\mathrm{im}\, A$.

Для произвольных матриц $A \in \mathbb{C}^{m\times n}$ полного ранга ($\mathrm{rank}\, A = \min(m,n)$) число обусловленности определяется в терминах псевдоинверсии: $\kappa(A) = \|A\|\|A^+\|$. Так как A^+ связано с линейной задачей наименьших квадратов, то определение спектрального числа обусловленности имеет вид

$$\kappa_2(A) = \|A\|_2\|A^+\|_2 = \frac{\sigma_1}{\sigma_n}. \tag{3.16}$$

Для матриц неполного ранга ($\mathrm{rank}\, A < \min(m,n)$) принимается, что спектральное число обусловленности матрицы $\kappa_2(A) = \infty$.

С более педантичной точки зрения число обусловленности матрицы (3.14), это число обусловленности для задачи обращения матрицы. Задаче вычисления собственных значений матрицы A соответствует, например, другое число обусловленности [9].

3.3. Теория возмущений

3.3.1. Системы с квадратными невырожденными матрицами

Пусть имеем систему

$$Au = f, \quad A \in \mathbb{R}^{n\times n}, \quad \det A \neq 0 \tag{3.17}$$

и соответствующую ей возмущенную систему

$$(A + \delta A)\hat{u} = f + \delta f; \tag{3.18}$$

нашей целью является оценка нормы вектора $\delta u \equiv \hat{u} - u$. Вычитая из (3.18) равенство (3.17), получаем соотношение

$$\delta A u + (A + \delta A)\delta u = \delta f,$$

которое можно привести к виду

$$\delta u = A^{-1}(-\delta A\hat{u} + \delta f). \tag{3.19}$$

Беря здесь нормы и используя свойство согласованности матричной и векторной норм, а также неравенство треугольника для векторных норм, находим

$$\|\delta u\| \leq \|A^{-1}\|(\|\delta A\| \cdot \|\hat{u}\| + \|\delta f\|). \tag{3.20}$$

(Предполагается, что векторная и матричная нормы являются согласованными. Например, можно взять произвольную векторную норму и индуцированную ее матричную норму). Трансформируя это неравенство, получаем

$$\frac{\|\delta u\|}{\|\hat{u}\|} \leq \|A^{-1}\| \cdot \|A\| \cdot \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta f\|}{\|A\| \cdot \|\hat{u}\|}\right). \tag{3.21}$$

Как следует из 3.2.5, произведение $\kappa(A) = \|A^{-1}\| \cdot \|A\|$ в (3.21) является числом обусловленности матрицы A, поскольку оно оценивает относительное изменение $\|\delta u\|/\|\hat{u}\|$ как кратное относительного изменения

$\|\delta A\|/\|A\|$ во входных данных. (Строго говоря, мы должны были бы показать, что неравенство (3.21) превращается в равенство при некотором выборе возмущений δA и δf; в противном случае $\kappa(A)$ было бы лишь верхней оценкой для числа обусловленности.) Если δA и δf малы, то мала и величина, на которую умножается $\kappa(A)$, что приводит к малой верхней границе для относительной ошибки $\|\delta u\|/\|\hat{u}\|$.

Наша верхняя оценка зависит от δu (входящего в $\hat{u}$), что, по видимости, затрудняет ее интерпретацию. В действительности, однако, эта оценка весьма полезна с практической точки зрения, поскольку вычисленное решение $\hat{u}$ известно, следовательно, и оценка легко может быть вычислена. Далее покажем, как вывести теоретически более привлекательную оценку, не зависящую от δu.

Лемма 3.1. *Пусть выбранная матричная норма $\|\cdot\|$ обладает свойством мультипликативности $\|AB\| \le \|A\|\|B\|$. Тогда, если $\|X\| < 1$, то матрица $E - X$ обратима, $(E-X)^{-1} = \sum_{i=0}^{\infty} X^i$ и $\|(E-X)^{-1}\| \le \frac{1}{1-\|X\|}$.*

Доказательство. Матричная сумма $\sum_{i=1}^{\infty} X^i$ сходится тогда и только тогда, когда сходится каждый элемент этой матрицы. Используя тот факт (свойство эквивалентности норм в конечномерных пространствах), что для произвольной нормы существует константа c, такая, что $|x_{jk}| \le c\|X\|$. Тогда $|(X^i)_{jk}| \le c\|X^i\| \le c\|X\|^i$, т.е. каждый элемент матрицы $\sum X^i$ мажорируется сходящейся геометрической прогрессией $\sum c\|X\|^i = \frac{1}{1-\|X\|}$ и, следовательно, сам сходится. Поэтому последовательность $S_n \sum_{i=0}^{n} X^i$ сходится при $n \to \infty$ к некоторому S и $(E-X)S_n = (E-X)(E + X^2 + \ldots + X^n) = E - X^{n+1} \to E$ при $n \to \infty$, поскольку $\|X^i\| \le \|X\|^i \to 0$. Таким образом, $(E-X)S = E$ и $S = (E-X)^{-1}$. Заключительная оценка выводится так: $\|(E-X)^{-1}\| = \|\sum_{i=0}^{\infty} X^i\| \le \sum_{i=0}^{\infty} \|X^i\| \le \sum_{i=0}^{\infty} \|X\|^i = \frac{1}{1-\|X\|}$. □

Разрешая уравнение (3.18) относительно δu, получаем

$$\begin{aligned} \delta u &= (A + \delta A)^{-1}(-\delta A u + \delta f = \\ &= [A(E + A^{-1}\delta A)]^{-1}(-\delta A u + \delta f) = \\ &= (E + A^{-1}\delta A)^{-1} A^{-1}(-\delta A u + \delta f). \end{aligned}$$

Беря здесь нормы, деля обе части на $\|u\|$, используя свойство согласованности матричной нормы с векторной и неравенство треугольника, предполагая, наконец, что δA настолько мало, что

$$\|A^{-1}\delta A\| \le \|A^{-1}\|\|\delta A\| < 1,$$

приходим к желаемой оценке:

$$\begin{aligned}
\frac{\|\delta u\|}{\|u\|} &\le \|(E + A^{-1}\delta A)^{-1}\| \cdot \|A^{-1}\| \left(\|\delta A\| + \frac{\|\delta f\|}{\|u\|} \right) \le \\
&\le \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|\delta A\|} \left(\|\delta A\| + \frac{\|\delta f\|}{\|u\|} \right) = \quad \text{по лемме (3.1)} \\
&= \frac{\|A^{-1}\| \cdot \|A\|}{1 - \|A^{-1}\| \cdot \|A\| \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta f\|}{\|A\| \cdot \|u\|} \right) \le \\
&\le \frac{\kappa(A)}{1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta f\|}{\|f\|} \right), \qquad (3.22)
\end{aligned}$$

поскольку $\|f\| = \|Au\| \le \|A\| \cdot \|u\|$.

Эта оценка выражает относительную ошибку решения $\|\delta u\|/\|u\|$ как кратное относительных ошибок $\|\delta A\|/\|A\|$ и $\|\delta f\|/\|f\|$ во входных данных. Если величина $\|\delta A\|$ достаточно мала, то множитель

$$\kappa(A) / \left(1 - \kappa(A) \frac{\|\delta A\|}{\|A\|} \right)$$

близок к числу обусловленности $\kappa(A)$.

Следующая теорема полнее раскрывает смысл предположения

$$\|A^{-1}\| \cdot \|\delta A\| = \kappa(A) \frac{\|\delta A\|}{\|A\|} < 1 :$$

оно гарантирует, что матрица $A + \delta A$ невырожденна; это необходимо для существования δu. Теорема также дает геометрическую характеризацию числа обусловленности.

Теорема 3.2. *Пусть матрица A невырожденна. Тогда*

$$\min \left\{ \frac{\|\delta A\|_2}{\|A\|_2} : \det(A + \delta A) \neq 0 \right\} = \frac{1}{\|A^{-1}\|_2 \cdot \|A\|_2} = \frac{1}{\kappa(A)}.$$

Следовательно, расстояние до ближайшей вырожденной матрицы (некорректная задача)=1/(число обусловленности).

Доказательство. Достаточно показать, что

$$\min\{\|\delta A\|_2 : \det(A + \delta A) \neq 0\} = \frac{1}{\||A^{-1}\|_2}.$$

Чтобы убедиться, что указанный минимум не меньше, чем $\|A^{-1}\|_2^{-1}$, заметим: если $\|\delta A\|_2 < \|A^{-1}\|$, то $1 > \|\delta A\|_2\|A^{-1}\|_2 \geq \|A^{-1}\delta A\|_2$. Согласно лемме 3.1, матрица $E + A^{-1}\delta A$ обратима, а потому обратима матрица $A + \delta A$.

Чтобы показать, что минимум равен $\|A^{-1}\|_2^{-1}$, построим возмущение δA с нормой $\|A^{-1}\|_2^{-1}$, такое, что матрица $A+\delta A$ вырожденна. Поскольку

$$\|A^{-1}\|_2 = \max_{u\neq 0} \frac{\|A^{-1}u\|_2}{\|u\|_2},$$

найдется вектор u, такой, что $\|u\|_2 = 1$ и $\|A^{-1}\|_2 = \|A^{-1}u\| > 0$. Положим теперь $y = \frac{A^{-1}u}{\|A^{-1}u\|_2} = \frac{A^{-1}u}{\|A^{-1}\|}$; таким образом, $\|y\|_2 = 1$. Положим $\delta A = \frac{-uy^\top}{\|A^{-1}\|_2}$.

Тогда

$$\|\delta A\|_2 = \max_{z\neq 0} \frac{\|uy^\top z\|_2}{\|A^{-1}\|_2\|z\|_2} = \max_{z\neq 0} \frac{|y^\top z|}{\|z\|_2}\frac{\|u\|_2}{\|A^{-1}\|_2} = \frac{1}{\|A^{-1}\|_2},$$

где максимум достигается, когда z есть произвольное ненулевое кратное вектора y. Матрица $A + \delta A$ вырожденна, так как

$$(A + \delta A)y = Ay - \frac{uy^\top y}{\|A^{-1}\|_2} = \frac{u}{\|A^{-1}\|_2} - \frac{u}{\|A^{-1}\|_2} = 0.$$

□

Из теоремы 3.2 видно, что расстояние до ближайшей некорректной задачи и число обусловленности взаимно обратны для задачи решения системы линейных уравнений.

Имеется несколько иной способ построения теории возмущений для задачи $Au = f$. Пусть $\hat{u}$ – произвольный вектор, тогда разность $\delta u \equiv \hat{u} - u = \hat{u} - A^{-1}f$ может быть оценена следующим образом. Как отмечалось, вектор $r = A\hat{u} - f$ называется невязкой; невязка $r = 0$, если $\hat{u} = A^{-1}f$. Это позволяет нам написать $\delta u = A^{-1}r$ и получить оценку

$$\|\delta u\| = \|A^{-1}r\| \leq \|A^{-1}\| \cdot \|r\|. \tag{3.23}$$

Эта простая оценка привлекательна с практической точки зрения, поскольку r легко вычисляется, если дано приближенное решение $\hat{u}$. Кроме того, нет видимой нужды в оценивании δA и δf. В действительности, оба описанных подхода весьма тесно связаны, что выявляется следующей теоремой.

Теорема 3.3. *Пусть $r = A\hat{u} - f$. Тогда найдется возмущение δA, такое, что $\|\delta A\| = \frac{\|r\|}{\|\hat{u}\|}$ и $(A + \delta A)\hat{u} = f$. Не существует никакого возмущения δA меньшей нормы, удовлетворяющего уравнению $(A + \delta A)\hat{u} = f$. Таким образом, δA есть наименьшая возможная обратная ошибка (измеряемая посредством нормы). Это утверждение справедливо для любой векторной нормы и индуцированной ее матричной нормы (или выборе $\|\cdot\|_2$ для векторов и $\|\cdot\|_E$ для матриц).*

Доказательство. Равенство $(A + \delta A)\hat{u} = f$ эквивалентно соотношениям $\delta A\hat{u} = f - A\hat{u} = -r$, поэтому $\|r\| = \|\delta A \cdot \hat{u}\| \le \|\delta A\| \cdot \|\hat{u}\|$, откуда $\delta A\| \le \frac{\|r\|}{\|\hat{u}\|}$. Остальную часть доказательства проведем только для 2-нормы и индуцированной ею матричной нормы. Положим $\delta A = \frac{-r\hat{u}^\top}{\|\hat{u}\|_2^2}$. Легко проверить, что $\delta A\hat{u} = -r$ и $\|\delta A\|_2 = \frac{\|r\|_2}{\|\hat{u}\|_2}$. □

Таким образом, наименьшее значение $\|\delta A\|$, позволяющее получить вектор $\hat{u}$, удовлетворяющий условиям $(A + \delta A)\hat{u} = f$ и $r = A\hat{u} - b$, указывается теоремой 3.3. Привлекая оценку (3.20) (с $\delta f = 0$), получаем

$$\|\delta u\| \le \|A^{-1}\| \left(\frac{r}{\hat{u}} \cdot \|\hat{u}\|\right) = \|A^{-1}\| \cdot \|r\|,$$

т.е. приходим к оценке (3.23).

Все наши оценки зависят от возможности оценить число обусловленности $\|A\| \cdot \|A^{-1}\|$. Оценки чисел обусловленности могут быть вычислены программами библиотеки LAPACK, например программой `sgesvx`.

3.3.2. Теория относительных возмущений

В предыдущем разделе было показано, как оценить норму ошибки $\delta u = \hat{u} - u$ приближенного решения $\hat{u}$ системы $Au = f$. Оценка для $\|\delta u\|$ была пропорциональна $\kappa(A) = \|A\| \cdot \|A^+\|$ и нормам $\|\delta A\|$ и $\|\delta f\|$ в предположении, что $\hat{u}$ удовлетворяет уравнению (3.18).

Во многих случаях эта оценка вполне удовлетворительна, но все же не всегда. Цель данного раздела – указать ситуации, когда эта оценка слишком пессимистична, и развить альтернативную теорию возмущений, дающую лучшие оценки.

Вот пример ситуации, где оценка ошибки из предыдущего раздела чересчур пессимистична.

Пример 3.3.1. Пусть $A = \mathrm{diag}(\gamma, 1)$ (диагональная матрица с диагональными элементами $a_{11} = \gamma$ и $a_{22} = 1$) и $f = (\gamma, 1)^\top$, где $\gamma > 1$. Тогда $u_* = A^{-1}f = (1, 1)^\top$. Всякий разумный прямой метод вычислит

очень точное приближение к $\hat{u}$ к решению u_* (посредством двух делений f_i/a_{ii}); в то же время число обусловленности $\kappa(A) = \gamma$ может быть как угодно велико. Следовательно, как угодно велика может быть и оценка ошибки (3.21).

Причина, почему число обусловленности $\kappa(A)$ заставляет переоценивать ошибку, состоит в следующем: оценка (3.20), содержащая это число, основана на предположении, что возмущение δA ограничено по норме, но *в остальном произвольно*; на последнем свойстве построено доказательство достижимости оценки (3.20) (см. [9]). В противоположность этому, возмущения δA, отвечающие реальным ошибкам округления, не являются произвольными, а имеют специальную структуру, не отражаемую нормой самой по себе. Наименьшее δA, соответствующее в нашей задаче вектору $\hat{u}$, можно определить так: простой анализ ошибок округления показывает, что $\hat{u}_i = (f_i/a_{ii})/(1+\delta_i)$, где $|\delta_i| \le \varepsilon$. Тем самым $(a_{ii} + \delta_i a_{ii})\hat{u}_i = f_i$. Это можно переписать как $(A + \delta A)\hat{u} = f$, где $\delta A = \mathrm{diag}(\delta_1 a_{11}, \delta_2 a_{22})$. Тогда $\|\delta A\|$ может достигать величины $\max_i |\varepsilon a_{ii}| = \varepsilon\gamma$. Применяя оценку ошибки (3.21) с $\delta f = 0$, находим

$$\frac{\|\delta u\|_\infty}{\|\hat{u}\|_\infty} \le \gamma \left(\frac{\varepsilon\gamma}{\gamma}\right) = \varepsilon\gamma.$$

В отличие от этого, реальная ошибка удовлетворяет соотношениям

$$\begin{aligned}
\|\delta u\| &= \|\hat{u} - u\|_\infty = \\
&= \left\| \begin{bmatrix} (f_1/a_{11})/(1+\delta_1) - (f_1/a_{11}) \\ (f_2/a_{22})/(1+\delta_2) - (f_2/a_{22}) \end{bmatrix} \right\|_\infty = \\
&= \left\| \begin{bmatrix} -\delta_1/(1+\delta_1) \\ -\delta_2/(1+\delta_2) \end{bmatrix} \right\|_\infty \le \\
&\le \frac{\varepsilon}{1-\varepsilon}
\end{aligned}$$

или

$$\frac{\|\delta u\|_\infty}{\|\hat{u}\|_\infty} \le \varepsilon/(1-\varepsilon)^2,$$

что примерно в γ раз меньше предыдущей оценки. $\diamond$

(Символ $\diamond$ – обозначает завершение примера.)

Для данного примера структуру реального возмущения δA можно описать следующим образом: $|\delta a_{ij} \le \varepsilon |a_{ij}|$, где ε – некоторое очень малое число. В более сжатой форме это можно описать как

$$|\delta A| \le \varepsilon |A|. \tag{3.24}$$

(Здесь для $m \times n$-матрицы A под $|A|$ понимается $m \times n$-матрица, составленная из абсолютны величин (или соответственно модулей для комплексных матриц) элементов A: $(|A|)_{ij} = |a_{ij}|$. Неравенства типа $|A| \le |B|$ следует понимать как системы покомпонентных неравенств: $|a_{ij}| \le |b_{ij}|$ для всех i и j. Аналогичные обозначения в дальнейшем будут использованы для векторов: $(|x|)_i = |x_i|$.) В таких случаях говорят, что δA есть *малое относительное покомпонентное возмущение* матрицы A. Поскольку на практике часто можно добиться того, чтобы δA удовлетворяла оценке (3.24), а δf – оценке $|\delta f| \le \varepsilon |f|$, то возможно построить теорию возмущений, основываясь на этих оценках для δA и δf.

Начнем с уравнения (3.19):

$$\delta u = A^{-1}(-\delta A \hat{u} + \delta f).$$

Переходя к абсолютным величинам и применяя неравенство треугольника, получаем

$$\begin{aligned} |\delta u| &= |A^{-1}(-\delta A \hat{u} + \delta f)| \le \\ &\le |A^{-1}|(|\delta A| \cdot |\hat{u}| + |\delta f|) \le \\ &\le |A^{-1}|(\varepsilon |A| \cdot |\hat{u}| + \varepsilon |f| = \\ &= \varepsilon (|A^{-1}|)(|A| \cdot |\hat{u}| + |f|)). \end{aligned}$$

Используя любую векторную норму, для которой $\||z|\| = \|z\|$ (таковы 1-норма, 2-норма и ∞-норма) приходим к оценке

$$\|\delta u\| \le \varepsilon \||A^{-1}|(|A| \cdot |\hat{u}| + |f|)\|. \tag{3.25}$$

Предположим, пусть $\delta u = 0$. Тогда (3.25) можно ослабить до оценки

$$\|\delta u\| \varepsilon \||A^{-1}| \cdot |A|\| \cdot \|\hat{u}\|$$

или

$$\frac{\|\delta u\|}{\|\hat{u}\|} \le \||A^{-1}| \cdot |A|\|. \tag{3.26}$$

Это неравенство служит обоснованием для того, чтобы назвать величину $\kappa_{CR}(A) \equiv \||A^{-1}| \cdot |A|\|$ *относительным покомпонентным числом обусловленности* матрицы A или, для краткости, ее *относительным числом обусловленности*. Иногда эту величину называют также числом обусловленности Бауэра или числом обусловленности Скила. По

поводу доказательства того, что оценки (3.25) и (3.26) достижимы, см. [9].

Пример 3.3.2. Вернемся к предыдущему примеру с $A = \mathrm{diag}(\gamma, 1)$ и $f = (\gamma, 1)^\top$. Легко проверить, что $\kappa_{CR}(A) = 1$, поскольку $|A^{-1}||A| = E$. В действительности, $\kappa_{CR}(A) = 1$ для любой диагональной матрицы A, что соответствует нашему интуитивному ощущению, согласно которому системы уравнений с диагональными матрицами должны решаться с высокой точностью. ◇

Рассмотрим более общий случай, где D – произвольная невырожденная диагональная матрица, а B – произвольная невырожденная матрица. Тогда

$$\begin{aligned}\kappa_{CR}(DB) &= \||(DB)^{-1}| \cdot |(DB)|\| = \\ &= \||B^{-1}D^{-1}| \cdot |DB|\| = \\ &= \||B^{-1}| \cdot |B|\| = \\ &= \kappa_{CR}(B).\end{aligned}$$

Это означает, что если матрица DB *плохо масштабирована*, т.е. B – хорошо обусловленная матрица, а DB обусловлена плохо (из-за того, что диагональные элементы в D сильно различаются по величине), то можно надеяться на возможность решения системы $(DB)u = f$ с высокой точностью, несмотря на плохую обусловленность матрицы DB.

В заключение приведем, как и в предыдущем разделе, оценку ошибки использующую только невязку $r = A\hat{u} - f$:

$$\|\delta u\| = \|A^{-1}r\| \le \||A^{-1}| \cdot |r|\|. \tag{3.27}$$

Здесь было применено неравенство треугольника. Эта оценка иногда может быть много меньше аналогичной оценки (3.23) [9]; так будет в частности, если A плохо масштабирована. Справедливо, кроме того, утверждение, аналогичное теореме 3.3.

Теорема 3.4. *Наименьшее число $\varepsilon > 0$, для которого существуют возмущения δA и δf, удовлетворяющие оценкам $|\delta A| \le \varepsilon|A|$ и $\delta f \le \varepsilon|f|$ и уравнению $(A + \delta A)\hat{u} = f + \delta f$, называется покомпонентной относительной обратной ошибкой. Оно выражается через невязку $r = A\hat{u} - f$ формулой*

$$\varepsilon = \max_i \frac{|r_i|}{(|A| \cdot |\hat{u}| + |f|)_i}.$$

Относительно доказательства теоремы см. [9].

Программы библиотеки LAPACK, такие, как `sgevx`, вычисляют покомпонентную относительную обратную ошибку ε (для соответствующей переменной в LAPACK'е принято имя `BERR`).

3.3.3. Теория возмущений для задачи наименьших квадратов

Для неквадратной матрицы полного ранга A спектральное число обусловленности, как показано в разделе 3.2.5, определяется как $\kappa_2(A) = \sigma_{\max}(A)/\sigma_{\min}(A)$. Это определение в случае квадратной матрицы A сводится к обычному. Следующая теорема обосновывает это новое определение.

Теорема 3.5. *Пусть* $A \in \mathbb{C}^{m\times n}$, $\operatorname{rank} A = n$, *где* $m \geq n$. *Предположим, что вектор* u_* *минимизирует* $\|Au - f\|_2$ *и* $r_* = Au_* - f$ *есть соответствующая невязка. Пусть вектор* $\tilde{u}$ *минимизирует* $\|(A+\delta A)u - (f+\delta f)\|_2$. *Предположим, что* $\varepsilon \equiv \max\left(\frac{\|\delta A\|_2}{\|A\|_2}, \frac{\|\delta f\|_2}{\|f\|_2}\right) < \frac{1}{\kappa_2(A)} = \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)}$. *Тогда*

$$\frac{\|\tilde{u}-u\|_2}{\|u\|_2} \leq \varepsilon \cdot \left\{ \frac{2\cdot\kappa_2(A)}{\cos\theta} + \tan\theta\cdot\kappa_2^2(A) \right\} + O(\varepsilon^2) \equiv \varepsilon\cdot\kappa_{LS} + O(\varepsilon_{mach}^2),$$

где $\sin\theta = \frac{\|r_*\|_2}{\|f\|_2}$. *Другими словами,* θ *есть угол между векторами* f *и* Au_*; *он измеряет велика или мала норма невязки* $\|r_*\|_2$ *(т.е., соответственно,* $\|r_*\|_2$ *близка к* $\|f\|$ *или 0). Символ* κ_{LS} *означает число обусловленности для задачи наименьших квадратов.*

Здесь символ "O" является стандартным в математике для указания точности: выражение $\varphi(t) = O(\psi(t))$ обозначает, что существуют некоторые константы $C > 0$ и t_0 такие, что при всех $t < t_0$ (или при $t > t_0$, т.е. при $t \to 0$ или при $t \to \infty$ соответственно) выполняется неравенство $|\varphi(t)| \leq C\psi(t)$.

Набросок доказательства. Разложим вектор $\tilde{u} = ((A+\delta A)^\top(A+\delta A))^{-1}(A+\delta A)^\top(f+\delta f)$ по степеням величин δA и δf, а затем отбросим все, кроме членов, линейных по δA и δf. □

Условие $\varepsilon_{mach}\kappa_2(A) < 1$ играет ту же роль, что и при выводе оценки (3.22) для возмущенного решения системы линейных уравнений $Au = f$ с квадратной матрицей: оно гарантирует, что $\operatorname{rank}(A+\delta A) = n$, поэтому вектор $\tilde{u}$ определен однозначно.

Полученную оценку можно интерпретировать следующим образом: если угол θ очень мал или равен нулю, то мала и невязка. В этом случае эффективное число обусловленности равно примерно $2\kappa_2(A)$, т.е. примерно тому же, что и при решении обычной системы линейных уравнений. Если θ не мал, но и не близок к $\pi/2$, то невязка умеренно велика, и эффективное число обусловленности может быть много больше, чем в первом случае, а именно равно $\kappa_2^2(A)$. Если θ близок к $\pi/2$, так что точное решение почти вырождается, то эффективное число обусловленности становится неограниченным даже при малом $\kappa_2(A)$.

Оценке из теоремы 3.5 можно придать другую форму, в которой отсутствует член с $O(\varepsilon_{mach}^2)$ [22]:

$$\frac{\|\tilde{u} - u_*\|}{\|u_*\|} \le \frac{\varepsilon\kappa_2(A)}{1 - \varepsilon\kappa_2(A)} \left(2 + (\kappa_2(A) + 1)\frac{\|r_*\|_2}{\|A\|_2\|u_*\|_2}\right),$$

$$\frac{\|\tilde{r} - r_*\|_2}{\|r_*\|_2} \le 1 + 2\varepsilon\kappa_2(A).$$

Здесь $\tilde{r}$ – невязка возмущенной задачи: $\tilde{r} = (A + \delta A)\tilde{u} - (f + \delta f)$.

Приведенные оценки точности позволяют провести сравнительный анализ для основных способов решения линейных задач наименьших квадратов полного ранга. Для данной задачи эти способы (они подробно описаны в [9] и [8]):

1. нормальные уравнения,
2. QR-разложение,
3. SVD-метод,
4. преобразование в расширенную линейную систему (2.5).

Здесь QR-метод – метод, основанный на ортогональном разложении матрицы $A = QR$, а SVD-метод – метод, основанный на сингулярном разложении матриц. При правильной реализации оба метода – и QR, и SVD, – *численно устойчивы*, т.е. дают приближенное решение $\tilde{u}$, которое (точно) минимизирует некоторую функцию вида $\|(A + \delta A)\tilde{u} - (f + \delta f)\|_2$, причем

$$\max\left(\frac{\|\delta A\|}{\|A\|}, \frac{\|\delta f\|}{\|f\|}\right) = O(\varepsilon_{mach}).$$

Комбинируя этот результат с приведенными выше оценками возмущений, можно получить оценки для ошибки в решении задачи наименьших квадратов подобно тому, как были получены оценки для ошибки в решении системы линейных уравнений.

Метод нормальных уравнений не столь точен. Поскольку решается система $(A^*A)u = A^*f$, точность определяется числом обусловленности $\kappa_2(A^*A) = \kappa_2^2(A)$. Таким образом, ошибка всегда ограничена величиной типа $\kappa_2^2(A)\varepsilon_{mach}$, а не просто $\kappa_2(A)\varepsilon_{mach}$. Следует ожидать поэтому, что при решении нормальных уравнений может быть потеряно вдвое больше верных разрядов, чем в методах, основанных на QR-разложении и SVD.

Кроме того, метод нормальных уравнений *не* всегда устойчив, т.е. вычисленное решение $\tilde{u}$, вообще говоря, не является точкой минимума функции вида $\|(A + \delta A)\tilde{u} - (f + \delta f)\|_2$ с малыми δA и δf. Все же, если число обусловленности мало, метод, скорей всего, даст приближенное решение примерно того же качества, что и QR-разложение и SVD. Поскольку задача наименьших квадратов быстрей всего (по числу арифметических операций) решается именно посредством нормальных уравнений, этот метод является предпочтительным в случае хорошо обусловленной матрицы A.

Последний метод (основанный на преобразовании в расширенную систему) позволяет проводить итерационное уточнение в случае плохо обусловленной задачи полного ранга. Все методы, кроме третьего (SVD-метода), могут быть адаптированы для эффективной обработки разреженных матриц.

3.4. Вычисление нормальных псевдорешений линейных систем

На ЭВМ можно вычислить только хорошо устойчивые объекты. Первый вывод, который вытекает из теории возмущений, состоит в том, что нормальное псевдорешение и псевдообратная матрица непригодны для вычислений на ЭВМ. Легко убедиться, рассматривая только диагональные матрицы и диагональные возмущения, что нормальное псевдорешение неустойчиво к изменению ранга матрицы.

В данном разделе рассматривается задача вычисления $u_* = A^+f$, где $A \in \mathbb{C}^{m\times n}$, $\tau = \operatorname{rank} A < \min(m, n)$. Нормальное псевдорешение $u_* = A^+f$ можно устойчиво вычислить, если исходные данные (A, f)

заданы точно и выполняется условие:

$$\frac{\sigma_\tau(A)}{\sigma_1(A)} > \varepsilon_{mach} \quad (\sigma_1(A) > 0). \tag{3.28}$$

Для отыскания псевлорешений требуется находить псевдообратную матрицу A^+. Однако, как будет показано в этом разделе, можно использовать рекуррентную процедуру (предложенную Р.Ш. Липцером и А.Н. Ширяемым [13]) нахождения псевдорешений, не требующую псевдообращения матрицы A.

В начале рассмотрим совместные системы с $m \le n$.

Введем некоторые обозначения. Пусть $k = 1, 2, \ldots, m$ – номера строк матрицы A, a_k – строки матрицы A, $A_k = \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix} \in \mathbb{C}^{k\times n}$, $f^k = (f_1, \ldots, f_k)^\top \in \mathbb{C}^k$.

Рассмотрим для каждого k совместные СЛАУ

$$A_k u = f^k.$$

Положим также

$$u_k = A_k^* f^k, \qquad \gamma_k = E_n - A_k^+ A_k.$$

Теорема 3.6. *Пусть система $Au = f$ удовлетворяет условиям $m \le n$ и $\operatorname{rank}(A\vdots f) = \operatorname{rank} A = m$. Тогда векторы u_k и матрицы γ_k $(k = 1, 2, \ldots, m)$ удовлетворяют системе рекуррентных соотношений*

$$\begin{aligned} u_{k+1} &= u_k + \gamma_k a_{k+1}^\top (a_{k+1}\gamma_k a_{k+1}^\top)^+ (f_{k+1} - a_{k+1}^\top u_k), & u_0 = 0, \\ \gamma_{k+1} &= \gamma_k - (a_{k+1}\gamma_k a_{k+1}^\top)^+ \gamma_k a_{k+1}^\top a_{k+1}\gamma_k, & \gamma_0 = E_n, \end{aligned}$$

где

$$(a_{k+1}\gamma_k a_{k+1}^\top)^+ = \begin{cases} (a_{k+1}\gamma_k a_{k+1}^\top)^{-1}, & \text{если } a_{k+1}\gamma_k a_{k+1}^\top > 0; \\ 0, & \text{если } a_{k+1}\gamma_k a_{k+1}^\top > 0. \end{cases}$$

$k = 1, 2, \ldots, m-1$ и вектор $u_m = u_* = A^+ f$.

Доказательство теоремы имеется в [13].

Если $\operatorname{rank} A = m$, то $(a_{k+1}\gamma_k a_{k+1}^\top)^+ = (a_{k+1}\gamma_k a_{k+1}^\top)^{-1}$ при всех $k = 1, 2, \ldots, m-1$.

Пусть исходная система $Au = f$ является произвольной (возможно несовместной и неполного ранга). Тогда для вычисления нормального псевдорешения воспользуемся расширенной системой (2.5):

$$\begin{pmatrix} E_m & A \\ A^* & 0 \end{pmatrix} \begin{pmatrix} r \\ u \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix} \quad \Leftrightarrow \quad Gz = b. \tag{3.29}$$

Из (3.29) непосредственно следует, что

$$z_* = G^+ b = \begin{pmatrix} f - Au_* \\ u_* \end{pmatrix},$$

где $u_* = A^+ f$.

Основное преимущество системы (3.29) состоит в том, что в отличие от исходной системы $Au = f$, она всегда совместна. Таким образом, для вычисления z_* можно воспользоваться рекуррентными формулами из теоремы 3.6, что дает возможность определить одновременно нормальное псевдорешение u_* и минимальные (в смысле наименьших квадратов) невязки r_* исходной системы.

Глава 4

Вычисление решений приближенных систем

4.1. Корректность вычислительной задачи

Постановка вычислительной задачи включает в себя *множество допустимых входных данных* X и *множество возможных решений* Y. Таким образом, как и в разделе 3.2, любая математическая задача в дальнейшем будет рассматриваться как некоторая функция (отображение) $g : X \to Y$ из некотрого нормированного векторного пространства X в нормированное векторное пространство Y решений.

Анализ важнейших требований, предъявляемых к различным прикладным задачам, приводит к понятию корректности математической задачи, которое было впервые сформулировано Ж.Адамаром. Вычислительная задача называется *корректной* (по Адамару), если выполнены следующие три требования: 1) ее решение $y \in Y$ существует при любых входных данных $x \in X$; 2) это решение единственно; 3) решение устойчиво по отношению к малым возмущениям входных данных. В том случае, когда хотя бы одно из этих требований не выполнено, задача называется *некорректной*. Рассмотрим подробнее эти требования.

Существование. Существование решения задачи – естественное требование к ней. Отсутствие решения может свидетельствовать, например, о непригодности математической модели либо о неправильной постановке задачи. Иногда отсутствие решения является следствием неправильного выбора множества допустимых входных данных X или множества возможных решений Y. Так как математическая модель не является абсолютно точным отражением действительности, то даже в

случае, когда исходная проблема заведомо имеет решение, соответствующая вычислительная задача может и не оказаться разрешимой. Конечно, такая ситуация говорит о серьезном дефекте в постановке задачи. В некоторых случаях отсутствие решения математической задачи приводит к пониманию того, что первоначально сформулированная проблема неразрешима и нуждается в серьезной корректировке.

Единственность. Для некоторых вычислительных задач единственность является естественным свойством; для других же решение может не быть единственным. Например, квадратное уравнение имеет два корня. Как правило, если задача имеет реальное содержание, то неединственность может быть ликвидирована введением дополнительных ограничений на решение (т.е. сужением множества Y). В некоторых случаях проблема снимается тем, что признается целесообразным найти набор всех решений, отвечающих входным данным x, и тогда за решение y принимается этот набор.

Неединственность решения вычислительной задачи – весьма неприятное свойство. Оно может быть проявлением неправильной постановки исходной прикладной проблемы, неоднозначности ее решения или сигналом о неудачном выборе математической модели.

Устойчивость решения. Решение y вычислительной задачи называется *устойчивым по входным данным* x, если оно зависит от входных данных непрерывным образом. Это означает, что для любого $\varepsilon > 0$ существует $\delta = \delta(\varepsilon) > 0$ такое, что всякому исходному данному x', удовлетворяющему условию $\|x' - x\| < \delta$, отвечает приближенное решение y', для которого $\|y' - y\| < \varepsilon$. Таким образом, для устойчивой вычислительной задачи ее решение теоретически можно найти со сколь угодно высокой точностью ε, если обеспечена достаточно высокая точность δ входных данных.

Для задач решения произвольных СЛАУ, как было показано в 2.1, первые два условия всегда устраняются за счет введения *нормального псевдорешения*. Однако, в силу того, что нормальное псевдорешение соответствует случаю когда матрица A системы имеет неполный ранг ($\operatorname{rank} A < \min(m, n)$), – оно неустойчиво к возмущениям в матрице A. Это непосредственно следует из того факта, что ранг матрицы (в случае ее неполного ранга) не будет непрерывной функцией от возмущений в элементах матрицы.

В случае, когда матрица матрица системы $Au = f$ является почти вырожденной, т.е. $\frac{\sigma_{\min}(A)}{\sigma_{\max}(A)} \leq macheps$, систему будем называть системой с *неполным машинным рангом*. Для таких систем может оказаться

предпочтительнее сразу же искать нормальное псевдорешение.

4.2. Постановка задачи

Методам решения некорректных систем линейных алгебраических уравнений посвящено огромное число работ. Среди этих работ особого внимания заслуживает В.А. Морозова [15], в которой дается глубокий и наиболее полный анализ основных подходов к приближенному решению данного класса некорректных задач. Основные подходы к приближенному решению некорректных задач связаны с тем или иным возмущением исходной задачи, переходом к некоторой "близкой", но уже корректной (*условно корректной* или *корректной по Тихонову*) задаче. На этом пути получаются различные регуляризирующие алгоритмы. При этом важнейшее значение имеет проблема выбора параметра регуляризации, его согласование с погрешностями в исходных данных. Следует особо отметить тот общий момент, что выбор параметра регуляризации приводит к существенному увеличению вычислительной работы и носит в той или иной степени итерационный характер.

Исключение составляет класс некорректных СЛАУ, когда точная исходная система является *совместной*. Для этого класса задач В.А. Морозовым и С.Ф. Гилязовым получена важная теорема (см. теорема 3 в [7]), которая снимает проблему выбора параметра регуляризации вообще.

Бьорком [21] для итерационного уточнения решений в линейных задачах наименьших квадратов (только для переопределенных СЛАУ полного ранга) был предложен метод эквивалентных расширенных систем. Этот метод позволяет произвольную несовместную СЛАУ преобразовать к эквивалентной расширенной *совместной* системе. На основе преобразования произвольной исходной СЛАУ к эквивалентной расширенной (но уже *совместной*) системе результаты полученные В.А. Морозовым и С.Ф. Гилязовым распостраняются на класс некорректных СЛАУ без ограничений на меру совместности исходных систем.

Матрица коэффициентов расширенной СЛАУ является симметричной матрицей. Этот факт дает возможность использовать для вычисления регуляризованных решений метод мнимого сдвига спектра (В.Н. Фаддеевой), что позволяет существенно снизить число обусловленности вычислительной задачи.

Постановка задачи о решении приближенной СЛАУ является фун-

даментальной задачей вычислительной математики [17].

Пусть точная СЛАУ

$$Au = f, \tag{4.1}$$

задается с помощью (априори неизвестных) исходных данных $d = \{A, f\}$, где $A = (a_{ij}) \in \mathbb{R}^{m\times n}$, $f = (f_1, \ldots, f_m)^\top \in \mathbb{R}^m$, $u = (u_1, \ldots, u_n)^\top \in \mathbb{R}^n$. Если $m \neq n$ или $m = n$ и $\det A = 0$, то условия разрешимости формулируются в виде точных равенств (теорема Кронекера – Капелли).

В общем случае под "решением"точной СЛАУ (4.1) будем понимать нормальное псевдорешение

$$u_* = A^+ f, \tag{4.2}$$

где A^+ – *псевдообратная* матрица или обобщенная обратная матрица Мура – Пенроуза.

Тогда *мера несовместности* точной СЛАУ (4.1) определяется величиной

$$\mu = \inf_{u\in\mathbb{R}^n} \|Au - f\| = \|Au_* - f\| \geq 0.$$

Везде в дальнейшем векторные нормы в $\mathbb{R}^m$ и $\mathbb{R}^n$ полагаются евклидовыми (квадратичными) нормами, т.е. $\|r\| = \|r\|_2$, где r – вектор невязки, $r = f - Au$.

Информация о системе (4.1) задается приближенными данными $\tilde{d} = \{\tilde{A}, \tilde{f}\}$ (индивидуальной приближенной СЛАУ $\tilde{A}u = \tilde{f}$), такими, что

$$\|\tilde{A} - A\| \leq h, \qquad \|\tilde{f} - f\| \leq \delta,$$

где $h \geq 0$ и $\delta \geq 0$ характеризуют погрешности задания приближенных данных $\tilde{d}$; $\|A\|$ – спектральная матричная норма, т.е.

$$\|\tilde{A} - A\| = \|\tilde{A} - A\|_2 = \sup_{\|u\|=1} \|\tilde{A}u - Au\|.$$

Как отмечалось в [15], решение (в смысле (4.2)) системы (4.1) неполного ранга ($\operatorname{rank} A < \min(m, n)$) по приближенным данным $\tilde{d}$ при $h > 0$ является некорректно поставленной по Адамару задачей, так как приближенное нормальное псевдорешение

$$\tilde{u}_* = \tilde{A}^+ \tilde{f}$$

неустойчиво к бесконечно малым возмущениям исходных данных.

Для нахождения устойчивых решений системы (4.1) по приближенным данным $\tilde{d}$ применяются различные методы регуляризации [15]. Одним из наиболее универсальных методов является метод регуляризации А.Н. Тихонова [17]. Как хорошо известно, регуляризованное решение $\tilde{u}_\alpha$ в этом методе определяется как (единственное) решение *уравнения Эйлера*

$$(\tilde{A}^\top \tilde{A} + \alpha E_n)u = \tilde{A}^\top \tilde{f}, \quad \alpha > 0, \tag{4.3}$$

где E_n – единичная матрица порядка n, α – параметр регуляризации.

Когда точная (априори неизвестная) система (4.1) совместна ($\mu = 0$), то, как показали В.А. Морозов и С.Ф. Гилязов (см. теорема 3 в [7]), если положить $\alpha = h$ в (4.3), тогда

$$\|\tilde{u}_\alpha - u_*\| = O(h + \delta), \tag{4.4}$$

где $\tilde{u}_\alpha$ – решение уравнения (4.3).

Для несовместных систем ($\mu > 0$) В.А. Морозов и С.Ф. Гилязов предложили двухэтапный алгоритм. Однако в этом алгоритме требуется вначале находить проекцию вектора $\tilde{f}$ на образ матрицы $\tilde{A}$. Эта процедура приводит к увеличению ошибки уклонения регуляризованного решения по сравнению с (4.4).

Здесь рассматривается другой подход для систем с $\mu > 0$. В этом подходе исходная система преобразуется к эквивалентной расширенной *совместной* системе [10] и затем для регуляризации последней используются результаты теоремы В.А. Морозова и С.Ф Гилязова [7]. Этот подход обеспечивает для регуляризованных решений такую же ошибку уклонения как в (4.4). Кроме того, предлагаемый подход позволяет получить эффективные численные алгоритмы для решения поставленной задачи.

4.3. Регуляризация на основе расширенных систем

Известно, что нормальное псевдорешение $\tilde{u}_* = \tilde{A}^+ \tilde{f}$ является нормальным решением нормальной системы уравнений

$$\tilde{A}^\top \tilde{A}u = \tilde{A}^\top \tilde{f} \tag{4.5}$$

или $\tilde{A}^\top \tilde{r} = 0$, где $\tilde{r} = \tilde{f} - \tilde{A}u$.

Таким образом, нормальная система уравнений (4.5) эквивалентна расширенной системе уравнений

$$\begin{array}{rcl} \tilde{r} + \tilde{A}u & = & \tilde{f}, \\ \tilde{A}^\top \tilde{r} & = & 0 \end{array} \iff \begin{pmatrix} E_m & \tilde{A} \\ \tilde{A}^\top & 0 \end{pmatrix} \begin{pmatrix} \tilde{r} \\ u \end{pmatrix} = \begin{pmatrix} \tilde{f} \\ 0 \end{pmatrix} \iff \tilde{G}z = \tilde{b}. \quad (4.6)$$

где $z = (\tilde{r}^\top, u^\top)^\top \in \mathbb{R}^{m+n}$.

Расширенная система вида (4.6) была впервые использована Бьорком [21] при $m \geq n$ для итерационного уточнения решений в линейных задачах наименьших квадратов.

Так как нормальная система уравнений (4.5) всегда совместна, то из этого непосредственно следуют

Утверждение 4.1. *Расширенная система (4.6) при любых исходных данных $\tilde{d} = \{\tilde{A}, \tilde{f}\}$ совместна.*

Утверждение 4.2. *Нормальное решение расширенной системы (4.6) определяется как $\tilde{z}_* = \tilde{G}^+\tilde{b} = (\tilde{r}_*^\top, \tilde{u}_*^\top)^\top$, где $\tilde{u}_* = \tilde{A}^+\tilde{f}$ и $\tilde{r}_* = \tilde{f} - \tilde{A}\tilde{u}_*$.*

Для нахождения устойчивых решений системы (4.6) воспользуемся методом регуляризации А.Н. Тихонова.

Так как $\tilde{G} = \tilde{G}^\top$, то $\tilde{G}^\top\tilde{G} = \tilde{G}^2$. Следовательно, регуляризованное решение $\tilde{z}_\alpha$ системы (4.6) определяется как (*единственное*) решение *уравнения Эйлера*

$$(\tilde{G}^2 + \alpha E_{m+n})z = \tilde{G}\tilde{b}. \quad (4.7)$$

Возмущения в приближенной системе $\{\tilde{G}, \tilde{b}\}$ удовлетворяют неравенствам

$$\|\tilde{G} - G\| = \left\| \begin{pmatrix} 0 & \tilde{A} - A \\ \tilde{A}^\top - A^\top & 0 \end{pmatrix} \right\| = \|\tilde{A} - A\| \leq h,$$

$$\|\tilde{b} - b\| = \|\tilde{f} - f\| \leq \delta.$$

Из утверждений 4.1 и 4.2 непосредственно следует, что точная расширенная система

$$\begin{pmatrix} E_m & A \\ A^\top & 0 \end{pmatrix} \begin{pmatrix} r \\ u \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix} \iff Gz = b$$

всегда совместна,

$$\bar{\mu} = \inf_{z \in \mathbb{R}^{m+n}} \|Gz - b\| = \|Gz_* - b\| = 0,$$

где $z_* = G^+ f = \begin{pmatrix} f - Au_* \\ u_* \end{pmatrix}$.

Тогда из теоремы В.А. Морозова и С.Ф. Гилязова (теорема 3 в [7]) непосредственно следует

Теорема 4.1. *Если в уравнении (4.7) положить $\alpha = h$, тогда ошибка уклонения*

$$\|\tilde{z}_\alpha - z_*\| = O(h + \delta),$$

где $\tilde{z}_\alpha$ – решение (4.7).

Таким образом, основные результаты полученные В.А. Морозовым и С.Ф. Гилязовым [7] для совместных систем ($\mu = 0$), с использованием эквивалентных расширенных систем непосредственно обобщаются на класс несовместных систем ($\mu > 0$).

Для численного решения (как прямыми, так и итерационными методами) уравнения Эйлера (4.7) для расширенных систем важно исследовать числа обусловленности этой системы.

4.4. Обусловленность вычислительной задачи

Для исследования чисел обусловленности вычислительной задачи исследуем спектр матрицы $G = \begin{pmatrix} E_m & A \\ A^\top & 0 \end{pmatrix}$.

Теорема 4.2. *Пусть $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_\tau > \sigma_{\tau+1} = \ldots = \sigma_s = 0$ сингулярные числа матрицы A, т.е. $\sigma_k = \sigma_k(A) = \sqrt{\lambda_k}$, где λ_k – собственные числа матрицы $A^\top A$, $k = 1, 2, \ldots, s$; $\tau = \operatorname{rank}(A) > 0$, $s = \min(m, n)$. Тогда расширенная матрица G имеет собственные числа $\omega_1 \geq \ldots \geq \omega_{m+n}$: $\frac{1}{2} + \sqrt{\frac{1}{4} + \sigma_1^2}$, $\ldots$, $\frac{1}{2} + \sqrt{\frac{1}{4} + \sigma_\tau^2}$, $\underbrace{0, \ldots, 0}_{n-\tau}$, $\underbrace{1, \ldots, 1}_{m-\tau}$, $\frac{1}{2} - \sqrt{\frac{1}{4} + \sigma_\tau^2}$, $\ldots$, $\frac{1}{2} - \sqrt{\frac{1}{4} + \sigma_1^2}$ соответственно.*

Доказательство. Если $Gz = \omega z$, $z = (v^\top, u^\top)^\top \neq 0$, тогда

$$v + Au = \omega v, \quad A^\top v = \omega u.$$

Исключая v, получаем

$$\omega u + A^\top A u = \omega^2 u,$$

или

$$A^\top Au = (\omega^2 - \omega)u.$$

Следовательно, если $u \neq 0$, тогда u собственный вектор и $(\omega^2 - \omega)$ соответствующее ему собственное число матрицы $A^\top A$, т.е.

$$\omega^2 - \omega - \lambda_k = 0. \tag{4.8}$$

Если $u = 0$, то

$$A^\top v = 0, \quad v = \omega v, \quad v \neq 0. \tag{4.9}$$

Из (4.8) и (4.9) непосредственно следует утверждение теоремы. □

Исследуем обусловленность регуляризованной системы (уравнения Эйлера) (4.7). Для этого оценим число обусловленности матрицы $\tilde{G}$.

Из теоремы 4.2 следует, что

$$\begin{aligned}\lambda_{\max}(\tilde{G}^2 + \alpha E) &= \lambda_{\max}(\tilde{G}^2) + \alpha = \omega_1^2 + \alpha = \\ &= \left(\frac{1}{2} + \sqrt{\frac{1}{4} + \tilde{\sigma}_1^2}\right)^2 + \alpha = \frac{1}{4}\left(1 + \sqrt{1 + 4\tilde{\sigma}_1^2}\right)^2 + \alpha\end{aligned}$$

и

$$\lambda_{\min}(\tilde{G}^2 + \alpha E) = \lambda_{\min}(\tilde{G}^2) + \alpha.$$

Тогда

$$\operatorname{cond}_2(\tilde{G}^2 + \alpha E) = \frac{\lambda_{\max}(\tilde{G}^2 + \alpha E)}{\lambda_{\min}(\tilde{G}^2 + \alpha E)} = \frac{\left(1 + \sqrt{1 + 4\tilde{\sigma}_1^2}\right)^2 + \alpha}{4(\lambda_{\min}(\tilde{G}^2) + \alpha)}. \tag{4.10}$$

Для вычисления (4.10) необходимо вычислять все сингулярные числа матрицы $\tilde{A}$. Поэтому можно воспользоваться оценкой

$$\operatorname{cond}_2(\tilde{G}^2 + \alpha E) \leq \frac{(1 + \sqrt{1 + 4\tilde{\sigma}_1^2})^2 + \alpha}{4\alpha} = 1 + \frac{1}{4\alpha}\left(1 + \sqrt{1 + 4\tilde{\sigma}_1^2}\right)^2, \tag{4.11}$$

где $\tilde{\sigma}_1$ – максимальное сингулярное число матрицы $\tilde{A}$.

Равенство в (4.11) достигается когда $\operatorname{rank}(A) < n$, так как в этом случае, на основании теоремы 4.2, $\lambda_{\min}(\tilde{G}^2) = \min\limits_{1 \leq k \leq m+n} |\tilde{\omega}_k| = 0$, где $\tilde{\omega}_k$ – собственные числа матрицы $\tilde{G}$.

На практике в (4.11) вместо $\tilde{\sigma}_1$ можно взять одну из легко вычисляемых норм $\|\tilde{G}\|_1 = 1 + \|\tilde{A}\|_1$ или $\|\tilde{G}\|_\infty = 1 + \|\tilde{A}\|_\infty$.

4.5. Метод мнимого сдвига спектра

Используя свойство симметричности матрицы $\tilde{G}$ можно понизить число обусловленности регуляризованной вычислительной задачи. Для этого воспользуемся методом мнимого сдвига спектра (В.Н. Фаддеевой).

В соответствии с этим методом вместо уравнения (4.7) рассмотрим уравнение

$$(\tilde{G} + i\sqrt{\alpha} \cdot E)z = \tilde{b}, \tag{4.12}$$

где $i = \sqrt{-1}$ – мнимая единица.

Пусть $z = x + iy$. Тогда (4.12) можно записать в виде расширенной вещественной СЛАУ

$$\begin{pmatrix} \tilde{G} & -\sqrt{\alpha}E \\ \sqrt{\alpha}E & \tilde{G} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \tilde{b} \\ 0 \end{pmatrix}. \tag{4.13}$$

Из (4.13) непосредственно следует, что $x = \operatorname{Re} z$ является решением уравнения

$$(\tilde{G}^2 + \alpha E)x = \tilde{G}\tilde{b}.$$

Это означает, что если z_* – решение (4.12), а $\tilde{z}_\alpha$ – решение (4.7), то $\tilde{z}_\alpha = \operatorname{Re} z_*$. Таким образом, задачу решения СЛАУ (4.7) можно заменить на задачу решения СЛАУ (4.12) с мнимым сдвигом спектра.

Вычислительная трудоемкость решения СЛАУ (4.12) не намного выше трудоемкости решения СЛАУ (4.7). Однако как будет показано, число обусловленности СЛАУ (4.12) существенно меньше числа обусловленности СЛАУ (4.7).

Исследуем число обусловленности матрицы $W_\alpha = \tilde{G} + i\sqrt{\alpha}E$. Собственные значения $\lambda_k(W_\alpha)$ матрицы W_α равны

$$\lambda_k(W_\alpha) = \tilde{\omega}_k + i\sqrt{\alpha}, \quad k = 1, 2, \ldots, m + n.$$

Тогда сингулярные числа $\sigma_k(W_\alpha)$ матрицы W_α определяются как

$$\sigma_k(W_\alpha) = \sqrt{\lambda_k(W_\alpha^* W_\alpha)} = \sqrt{\lambda_k(\tilde{G}^2 + \alpha E)} = \sqrt{\tilde{\omega}_k^2 + \alpha},$$

где $W_\alpha^* = \overline{W_\alpha}^\top = \tilde{G} - i\sqrt{\alpha}E$.

Следовательно

$$\operatorname{cond}_2(\tilde{G}^2 + \alpha E) = (\operatorname{cond}_2(\tilde{G} + i\sqrt{\alpha}E))^2$$

и

$$\operatorname{cond}_2(\tilde{G}+i\sqrt{\alpha}E)\le\sqrt{1+\frac{1}{4\alpha}\left(1+\sqrt{1+4\tilde{\sigma}_1^2}\right)^2}. \tag{4.14}$$

Равенство в (4.14) имеет место когда $\operatorname{rank}(\tilde{A})<n$. Если $\operatorname{cond}_2(\tilde{A})\gg 1$, то приближенно

$$\operatorname{cond}_2(\tilde{G}+i\sqrt{\alpha}E)\le\frac{1+\sqrt{1+4\tilde{\sigma}_1^2}}{2\sqrt{\alpha}}. \tag{4.15}$$

Таким образом, из (4.14) и (4.15) видно, что для плохо обусловленных и неполного столбцового ранга матриц $\tilde{A}$ вычисление регуляризованных решений $\tilde{z}_\alpha$ на основе СЛАУ (4.12) с мнимым сдвигом спектра существенно эффективнее, чем на основе СЛАУ (4.7).

4.6. Численные примеры

В качестве иллюстрации возможностей предлагаемого метода регуляризации неустойчивых конечномерных линейных задач рассмотрим вырожденную несовместную (точную) систему уравнений

$$\begin{cases} x_1+\sqrt{2}x_2=1,\\ 2x_1+2\sqrt{2}x_2=\sqrt{2}, \end{cases}\qquad A=\begin{pmatrix}1 & \sqrt{2}\\ 2 & 2\sqrt{2}\end{pmatrix},\quad f=\begin{pmatrix}1\\ \sqrt{2}\end{pmatrix}. \tag{4.16}$$

Система уравнений (4.16) не имеет решений. Найдем ее нормальное псевдорешение $u_*=A^+f$. Для этого вычислим псевдообратную матрицу

$$A^+=\lim_{\alpha\to 0}(A^\top A+\alpha E)^{-1}A'=\frac{1}{15}\begin{pmatrix}1 & 2\\ \sqrt{2} & 2\sqrt{2}\end{pmatrix}.$$

Тогда нормальное псевдорешение (точной) СЛАУ (4.16)

$$u_*=\frac{1}{15}\begin{pmatrix}1 & 2\\ \sqrt{2} & 2\sqrt{2}\end{pmatrix}\begin{pmatrix}1\\ \sqrt{2}\end{pmatrix}=\begin{pmatrix}\frac{1+2\sqrt{2}}{15}\\ \frac{4+\sqrt{2}}{15}\end{pmatrix}\approx\begin{pmatrix}0,255\\ 0,360\end{pmatrix}.$$

Теперь будем решать систему (4.16) любым классическим (машинным) методом. Положим (используя округление до двух десятичных знаков) $\sqrt{2}\approx 1,41$ и $2\sqrt{2}\approx 2,83$. Тогда приближенная система уравнений будет

$$\begin{cases} x_1+1,41x_2=1,\\ 2x_1+2,83x_2=1,41, \end{cases}\qquad \tilde{A}=\begin{pmatrix}1 & 1,41\\ 2 & 2,83\end{pmatrix},\quad \tilde{f}=\begin{pmatrix}1\\ 1,41\end{pmatrix}. \tag{4.17}$$

Система (4.17) формально невырожденная: $\det \tilde{A} = 0,01 \neq 0$. Любым классическим (машинным) методом найдем ее решение с точностью 10^{-3}:

$$\tilde{u}_* = \begin{pmatrix} 84,140 \\ -59,000 \end{pmatrix}.$$

Очевидно, погрешности приближенных данных системы (4.17) определяются величинами:

$$\|\tilde{A} - A\|_2 \leq \|\tilde{A} - A\|_E \leq h = 10^{-2} \quad \text{и} \quad \|\tilde{f} - f\|_2 \leq \delta = 10^{-2},$$

где $\|A\|_E$ – сферическая (евклидова) матричная норма,

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

Регуляризованные решения $\tilde{u}_\alpha$ системы (4.17) определялись из СЛАУ (4.13). На основании теоремы 1.1 параметр регуляризации выбирается $\alpha = h = 0,01$. Тогда

$$\tilde{G}_\omega = \begin{pmatrix} 1 & 0 & 1 & 1,41 \\ 0 & 1 & 2 & 2,83 \\ 1 & 2 & 0 & 0 \\ 1,41 & 2,83 & 0 & 0 \end{pmatrix} \quad \text{и} \quad \tilde{b} = \begin{pmatrix} 1 \\ 1,41 \\ 0 \\ 0 \end{pmatrix}.$$

Таким образом, в данном примере СЛАУ (4.13) имеет вид:

$$\begin{pmatrix} 1 & 0 & 1 & 1,41 & -0,1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 2,83 & 0 & -0,1 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 & -0,1 & 0 \\ 1,41 & 2,83 & 0 & 0 & 0 & 0 & 0 & -0,1 \\ 0,1 & 0 & 0 & 0 & 1 & 0 & 1 & 1,41 \\ 0 & 0,1 & 0 & 0 & 0 & 1 & 2 & 2,83 \\ 0 & 0 & 0,1 & 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0,1 & 1,41 & 2,83 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1,41 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

и $\tilde{u}_\alpha = (\tilde{u}_{1\alpha}, \tilde{u}_{2\alpha})^\top = (x_3, x_4)^\top$.

Решая эту систему любым классическим методом получаем регуляризованное решение с точностью 10^{-3}:

$$\tilde{u}_\alpha = \begin{pmatrix} \tilde{u}_{1\alpha} \\ \tilde{u}_{2\alpha} \end{pmatrix} \approx \begin{pmatrix} 0,255 \\ 0,359 \end{pmatrix}.$$

Сравнивая регуляризованное решение $\tilde{u}_\alpha$ с точным нормальным псевдорешением u_*, получаем

$$\|\tilde{u}_\alpha - u_*\| < 10^{-2},$$

что соответствует приведенным теоретическим результатам.

Рассмотрим другую приближенную к (4.16) систему. Пусть коэффициенты матрицы системы (4.16) получены округлением до пяти десятичных знаков ($\sqrt{2} \approx 1,41421$, $2\sqrt{2} \approx 2,82843$), а элементы вектора правой части округлением до двух десятичных знаков, т.е. $\sqrt{2} \approx 1,41$. Таким образом, имеем приближенную систему

$$\tilde{A} = \begin{pmatrix} 1 & 1,41421 \\ 2 & 2,82843 \end{pmatrix}, \quad \tilde{f} = \begin{pmatrix} 1 \\ 1,41 \end{pmatrix}. \tag{4.18}$$

Приближенная система (4.18) также невырожденная и ее решение может быть получено любым классическим методом:

$$\tilde{u}_* = \begin{pmatrix} -4,14 \cdot 10^4 \\ 2,93 \cdot 10^4 \end{pmatrix}.$$

Для системы (4.18) погрешности приближенных данных определяются соответственно величинами:

$$\|\tilde{A} - A\|_2 \le \|\tilde{A} - A\|_E \le h = 10^{-5} \quad \text{и} \quad \|\tilde{f} - f\|_2 \le \delta = 10^{-2}.$$

Тогда параметр регуляризации $\alpha = h = 10^{-5}$ и принимая $\sqrt{\alpha} \approx 0,03$, аналогично предыдущему примеру получаем регуляризованное решение для приближенной СЛАУ:

$$\tilde{u}_\alpha = \begin{pmatrix} \tilde{u}_{1\alpha} \\ \tilde{u}_{2\alpha} \end{pmatrix} \approx \begin{pmatrix} 0,255 \\ 0,360 \end{pmatrix}.$$

В этом примере также достигается точность приближенного решения соответствующая, приведенным здесь теоретическим результатам.

Глава 5

Идентификация нестационарных AR-моделей

В этой главе рассмотрен пример использования методов регуляризации к прикладной задаче идентификации параметров моделей процессов роста, описываемых нестационарными моделями авторегрессии (AR-моделями). Эта задача относится к классу неустойчивых вычислительных задач.

5.1. Стохастические непрерывные модели

При построении детерминированных математических моделей процессов роста экономических, и ряда других, показателей исходят из следующих двух предположений: скорость изменения показателя $x(t)$ пропорциональна его значению; скорость изменения пропорциональна также функции $g(x)$, монотонно убывающей со временем. Эта функция вводится для учета возможных ограничений ресурсов и влияния конкурентов. Дифференциальное уравнение имеет вид:

$$\frac{dx}{dt} = xg(x). \tag{5.1}$$

Функция $g(x)$ выбирается так, чтобы значение эндогенной переменной не стремилась к бесконечности. Стабилизация численности популяции при $t \to \infty$ есть свойство модели, которое обеспечивается специальным выбором функции $g(x)$. Определенный выбор функции $g(x)$ задает конкретный вид модели процесса роста. В исследованиях процессов

роста экономических показателей, аналогично процессам роста биологических популяций, наиболее распространены следующие два способа выбора функций $g(x)$.

Первому способу соответствует выбор функции $g(x)$ в виде

$$g(x) = c_2(1 - x/c_1), \quad c_1, c_2 > 0.$$

При таком выборе функции $g(x)$, дифференциальное уравнение (5.1) будет соответствовать так называемой модели Верхулста. Для модели Верхулста аналитическое решение дифференциального уравнения (5.1) равно

$$x(t) = \frac{c_1}{1 + (c_1/x(0) - 1)\exp(-c_2 t)}. \tag{5.2}$$

Функция (5.2) представляет собой классическую логистическую функцию. Начиная от значения $x(0)$, где $x(0) < c_1$, значение эндогенной переменной постоянно возрастает во времени t и стремиться к c_1 при $t \to \infty$.

Для оценивания параметров c_1 и c_2 модели Верхулста по экспериментальным данным используются методы нелинейного регрессионного анализа, в предположении, что значения эндогенной переменной измеряется с аддитивным шумом, т.е. наблюдаются

$$y(t) = x(t) + e(t), \tag{5.3}$$

где $e(t)$ – аддитивные помехи типа белого шума.

Существенным недостатком модели (5.2) является предположение об аддитивном характере (5.3) возмущений $e(t)$, так как это предположение теоретически не обеспечивает выполнение условия: $y(t) > 0$ с вероятностью 1, что противоречит физической сущности процесса роста.

Второй способ выбора функции $g(x)$:

$$g(x) = c_1 - c_2 \ln(x), \quad c_1 \geq 0, c_2 > 0.$$

Такой способ выбора функции $g(x)$ соответствует так называемой модели Гомпертца. Для модели Гомпертца аналитическое решение дифференциального уравнения (5.1) имеет вид:

$$\ln x(t) = (c_1/c_2) + (\ln x(0) - c_1/c_2)\exp(-c_2 t). \tag{5.4}$$

Функция $\ln x(t)$ монотонно убывает при $t \to \infty$, а $x(t)$ стремиться к положительному числу $\exp(c_1/c_2)$.

Для оценивания параметров c_1 и c_2 модели Гомертца (5.4) по экспериментальным данным также используются методы нелинейного регрессионного анализа, в предположении, что эндогенная переменная измеряется с мультипликативными возмущениями, т.е. вместо $x(t)$ наблюдаются

$$y(t) = x(t) \cdot \xi(t), \tag{5.5}$$

где $\xi(t)$ – неотрицательная случайная величина с $\mathbf{M}[\xi(t)] = 1$ ($\mathbf{M}$ – оператор математического ожидания) и $\ln \xi(t)$ является случайным процессом типа белого шума.

Дальнейшим усовершенствованием модели (5.1), с целью повышения прогнозирующих свойств, является ее естественное обобщение на стохастический случай, которое получается добавлением на входе случайного процесса, отражающего совокупное действие влияющих на эндогенную переменную экзогенных факторов, таких как ресурсов и т.д. Интенсивность случайного входного сигнала в любой момент времени также пропорциональна значению эндогенной переменной в тот же момент. Таким образом, стохастический вариант уравнения (5.1) имеет вид

$$\frac{dx}{dt} = xg(x) + x\eta(t),$$

где $\eta(t)$ – некоторый случайный процесс отражающий влияние выше перечисленных (неучтенных в модели) экзогенных факторов. Простейшее предположение относительно $\eta(t)$ – считать $\eta(t)$ белым шумом с $\mathbf{M}[\eta(t)] = 0$. Однако такое предположение не совсем реалистично, так как трудно представить, что совокупное действие всех важных, влияющих на эндогенную переменную, факторов было бы некоррелировано во времени. Здесь будет рассмотрен стохастический вариант модели Гомпертца в предположении, что входное воздействие не является белым шумом, а подчиняется стохастическому дифференциальному уравнению.

Стохастический вариант модели Гомпертца можно записать в виде:

$$\frac{dx}{dt} = x(c_1 - c_2 \ln x) + x\eta(t). \tag{5.6}$$

Разделив (5.6) на x, получаем

$$\frac{d \ln x}{dt} = c_1 - c_2 \ln x + \eta(t). \tag{5.7}$$

Предположение, что случайное входное воздействие подчиняется стохастическому дифференциальному уравнению первого порядка

$$\frac{d\eta}{dt} = a\eta + d(t) + \varpi(t), \tag{5.8}$$

где $a < 0$, $\varpi(t)$ – белый шум с непрерывным временем и нулевым средним, $d(t)$ – сумма постоянного сигнала и функций тренда. Конкретный вид функции тренда будет рассмотрен ниже.

Исключая $\eta(t)$ из уравнений (5.7) и (5.8) и дифференцируя (5.7) по t, получаем

$$\frac{d^2 \ln x}{dt^2} = -c_2 \frac{d \ln x}{dt} + \frac{d\eta(t)}{dt}. \tag{5.9}$$

Подставив в уравнение (5.8) выражение для η через x, получаемое из (5.7), а затем вместо производной $d\eta/dt$ подставив ее выражение через x, получаемое из (5.9), будем иметь:

$$\frac{d^2 \ln x}{dt^2} + c_2 \frac{d \ln x}{dt} = a \frac{d \ln x}{dt} - c_1 \ln x + +d(t) + \varpi(t)$$

или

$$\frac{d^2 \ln x}{dt^2} + \alpha_1 \frac{d \ln x}{dt} + \alpha_2 \ln x = d_1(t) + \varpi(t), \tag{5.10}$$

где $\alpha_1 = c_2 - a$, $\alpha_2 = -\alpha c_2$ и $d_1(t) = d(t) - ac_1$.

5.2. Стохастические дискретные модели

Рассмотрим дискретный вариант стохастической непрерывной модели (5.10). Для этого произведем дискретизацию по времени стохастического дифференциального уравнения.

Пусть значения эндогенной переменной $x(t)$ известны в дискретные моменты времени $t_0, t_0 + \Delta t, t_0 + 2 \cdot \Delta t, \ldots, t_0 + k \cdot \Delta t, \ldots$ с равномерным по времени шагом Δt. Обычно в экономических исследованиях Δt – сутки, неделя, квартал, год и т.д. Обозначим $x_k = x(t_0 + k \cdot \Delta t)$, где $k = 0, 1, 2, \ldots$ – дискретное время. Тогда на основании стохастического дифференциального уравнения (5.10) получаем стохастическое разностное уравнение:

$$\ln x_k = \theta_1 \ln x_{k-1} + \theta_2 \ln x_{k-2} + d_1(k) + e_k, \tag{5.11}$$

где k – дискретное время, $k = 0, 1, 2, \ldots$, а e_k – дискретный белый шум, удовлетворяющий условиям:

$$\begin{aligned} \mathbf{M}(e_{k+1}|\mathfrak{F}_k) &= 0 \quad \text{п.н.}, \\ \mathbf{M}(e_{k+1}^2|\mathfrak{F}_k) &= \sigma^2 < \infty \quad \text{п.н.}, \end{aligned}$$

$\mathfrak{F}_k$ – σ-алгебра, $\mathfrak{F}_k = \sigma\{e_1, \ldots, e_k\}$, $\mathfrak{F}_0 = \emptyset$, п.н. – почти наверное.

Уравнение (5.11) является нелинейным стохастическим разностным уравнением относительно переменной x_k и линейным стохастическим разностным уравнением (AR-уравнением) относительно $\tilde{x}_k = \ln x_k$. Уравнение (5.11) перепишем в виде:

$$\tilde{x}_k = \theta_1 \tilde{x}_{k-1} + \theta_2 \tilde{x}_{k-2} + d_1(k) + e_k. \tag{5.12}$$

Слагаемое $d_1(k)$ в уравнении (5.12) представляет собой функцию тренда и позволяет учитывать нестационарность среднего значения процесса роста. В исследованиях динамики роста экономических процессов чаще всего используются следующие способы задания функции тренда:

$$\begin{aligned} d_1(k) &= \theta_0, \\ d_1(k) &= \theta_0 + \theta_3 t, \\ d_1(k) &= \theta_0 + \sum_{j=1}^{p} \theta_{j+2} k^j, \\ d_1(k) &= \theta_0 + \theta_3 \cos \omega t + \theta_4 \sin \omega t. \end{aligned}$$

Таким образом, в экономических исследованиях используются четыре основных типа AR-моделей с нестационарными трендами для описания процессов роста, выводимых из непрерывных стохастических дифференциальных моделей Гомпертца:

$$\tilde{x}_k = \theta_0 + \theta_1 \tilde{x}_{k-1} + \theta_2 \tilde{x}_{k-2} + e_k, \tag{5.13}$$

$$\tilde{x}_k = \theta_0 + \theta_1 \tilde{x}_{k-1} + \theta_2 \tilde{x}_{k-2} + \theta_3 k + e_k, \tag{5.14}$$

$$\tilde{x}_k = \theta_0 + \theta_1 \tilde{x}_{k-1} + \theta_2 \tilde{x}_{k-2} + \sum_{j=1}^{q} k^j + e_k, \tag{5.15}$$

$$\tilde{x}_k = \theta_0 + \theta_1 \tilde{x}_{k-1} + \theta_2 \tilde{x}_{k-2} + \theta_3 \cos(\omega k) + \theta_4 \sin(\omega k) + e_k. \tag{5.16}$$

Модели (5.13) – (5.16) являются AR-моделями с постоянным, линейным, полиномиальным и периодическим трендами соответственно. AR-модели с периодическим трендом (5.16) широко используются в экономике для описания процессов роста с сезонными колебаниями.

5.3. Постановка задачи идентификации

Для решения задачи идентификации неизвестных параметров моделей (5.13) – (5.16) по экспериментальным данным $\{x(k)\}$ можно воспользоваться наиболее часто используемым в прикладном регрессионном анализе, методом наименьших квадратов (МНК), наиболее часто используемым в прикладном регрессионном анализе.

Для определенности в начале рассмотрим модель (5.13). В соответствии с МНК оценки $\hat{u}$ неизвестных параметров $u = (\theta_0, \theta_1, \theta_2)^\top$ модели (5.13) будут определяться как псевдорешение системы линейных алгебраических уравнений $Au = f$, т.е.

$$\hat{u} = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \| Au - f \|_2^2, \tag{5.17}$$

где $n = 3$, m – объем экспериментальных данных, $m > n$,

$$A = \begin{pmatrix} 1 & \tilde{x}_1 & \tilde{x}_0 \\ \vdots & \vdots & \vdots \\ 1 & \tilde{x}_m & \tilde{x}_{m-1} \end{pmatrix} \in \mathbb{R}^{m \times 3}, \qquad f = \begin{pmatrix} \tilde{x}_2 \\ \vdots \\ \tilde{x}_{m+1} \end{pmatrix} \in \mathbb{R}^m.$$

Аналогично, для модели (5.14): $n = 4$,

$$A = \begin{pmatrix} 1 & \tilde{x}_1 & \tilde{x}_0 & 2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \tilde{x}_m & \tilde{x}_{m-1} & m+1 \end{pmatrix} \in \mathbb{R}^{m \times 4}.$$

Для модели (5.15): $n = 5$,

$$A = \begin{pmatrix} 1 & \tilde{x}_1 & \tilde{x}_0 & \cos(2 \cdot \omega) & \sin(2 \cdot \omega) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \tilde{x}_m & \tilde{x}_{m-1} & \cos((m+1) \cdot \omega) & \sin((m+1) \cdot \omega) \end{pmatrix} \in \mathbb{R}^{m \times 5}.$$

И для модели (5.16): $n = 3 + q$,

$$A = \begin{pmatrix} 1 & \tilde{x}_1 & \tilde{x}_0 & 2 & 2^2 & \dots & 2^q \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & \tilde{x}_m & \tilde{x}_{m-1} & m+1 & (m+1)^2 & \dots & (m+1)^q \end{pmatrix} \in \mathbb{R}^{m \times (3+q)}.$$

Матрица регрессоров A в реальных прикладных задачах часто оказывается плохо обусловленной. К тому же часть элементов матрицы A

не может быть абсолютно точно представлена в компьютере. Это связано с тем фактом, что наиболее распространенная форма представления действительных чисел в компьютерах – это *числа с плавающей точкой.* Таким образом, все числа $\tilde{x}_k = \ln x_k$ в компьютере могут быть заданы лишь приближенно и вместо точных матрицы регрессоров A и вектора f в компьютере имеются лишь их приближенные значения $\tilde{A}$ и $\tilde{f}$. Это приводит к тому, что решение задачи (5.17) по приближенным данным является неустойчивой вычислительной задачей. В итоге, это не только может ухудшить эффективность самого статистического метода, основанного на МНК и используемого для решения поставленной задачи идентификации параметров моделей (5.13) – (5.16), но и, вообще, поставить под сомнение полученные результаты. В сложившихся обстоятельствах встает проблема применения численно устойчивых алгоритмов решения указанных задач идентификации параметров AR-моделей вида (5.13) – (5.16). Эффективные методы регуляризации указанных неустойчивых вычислительных задач идентификации будут рассмотрены далее в главе 4.

Для применения рассмотренных в главе 4 методов необходимо оценить погрешности представления исходных данных (A, f) в компьютерной арифметике с плавающей точкой, т.е. оценить величины h и δ такие, что

$$\|\tilde{A} - A\|_E \le h, \qquad \|\tilde{f} - f\|_2 \le \delta,$$

где $\|A\|_E$ – евклидова (сферическая) матричная норма, а $\|f\|_2$ – евклидова векторная норма.

Пусть $\tilde{A} = fl(A)$ и $\tilde{f} = fl(f)$. Тогда из результатов [14] непосредственно следует, что

$$\delta = \varepsilon_1 \sqrt{\sum_{k=2}^{m+1} \tilde{x}_k^2} + \varepsilon_0 \sqrt{m},$$

где $\varepsilon_0 = \beta^{\nu^- - 1}$.

Параметр h в соответствии с [14] оценивается величиной

$$h = \varepsilon_1 \Delta_1 + \varepsilon_0 \Delta_2,$$

где для моделей (5.13), (5.14), (5.16):

$$\Delta_1 = \sqrt{2 \sum_{k=1}^{m} \tilde{x}_k^2 + \tilde{x}_0^2 + \tilde{x}_{m+1}^2} \qquad \text{и} \qquad \Delta_2 = \sqrt{2m},$$

а для модели (5.15):

$$\Delta_1 = \sqrt{2\sum_{k=1}^{m}\tilde{x}_k^2 + \tilde{x}_0^2 + \tilde{x}_{m+1}^2 + \sum_{k=2}^{m+1}\left(\cos^2(k\omega) + \sin^2(k\omega)\right)}$$

и

$$\Delta_2 = 2\sqrt{m}.$$

Для практических расчетов достаточно для всех моделей (5.13) – (5.16) принять приближенно

$$h \approx \varepsilon_1\Delta_1 \quad \text{и} \quad \delta \approx \varepsilon_1\sqrt{\sum_{k=2}^{m+1}\tilde{x}_k^2}.$$

Применение методов регуляризации на основе расширенных моделей дает возможность существенно повысить точность вычисленных оценкок нестационарных AR-моделей, по сравнению с обычным методом наименьцццццццццццццццц

Литература

[1] *Алберт А.* Регрессия, псевдоинверсия и рекуррентное оценивание: Пер. с англ. – М.: Наука, 1977. – 224 с.

[2] *Амосов А.А., Дубинский Ю.А., Копченова Н.В.* Вычислительные методы для инженеров. – М.: Высш. шк., 1994. – 544 с.

[3] *Беклемишев Д.В.* Дополнительные главы линейной алгебры: Учебное пособие. – М.: Наука, 1983. – 336 с.

[4] *Беллман Р.* Введение в теорию матриц: Пер. с англ. – М.: Наука, 1976. – 368 с.

[5] *Воеводин В.В.* Вычислительные основы линейной алгебры. – М.: Наука, 1977.

[6] *Гантмахер Ф.Р.* Теория матриц. – М. Наука, 1967. – 576 с.

[7] *Гилязов С.Ф., Морозов В.А.* О регуляризации некорректно поставленных задач с нормально разрешимыми операторами // Ж. вычисл. матем. и матем. физ. 1997. Т. 37. № 2. С. 139–144.

[8] *Голуб Дж., Ван Лоун Ч.* Матричные вычисления: Пер. с англ. – М.: Мир, 1999. – 548 с.

[9] *Деммель Дж.* Вычислительная линейная алгебра. Теория и приложения: Пер. с англ. – М.: Мир, 2001. – 430 с.

[10] *Жданов А.И.* Регуляризация неустойчивых конечномерных линейных задач на основе расширенных систем//Ж. вычисл. матем. и матем. физ. 2005. Т. 45. № 11. С.1918–1926.

[11] *Ильин В.А., Ким Г.Д.* Линейная алгебра и аналитическая геометрия: Учебник. – 2-е изд. – М.: Изд-во МГУ, 2002. – 320 с.

[12] *Ильин В.А., Садовничий В.А., Сендов Бл.Х.* Математический анализ: Учебник. – М.: Наука, 1979. – 720 с.

[13] *Липцер Р.Ш., Ширяев А.Н.* Статистика случайных процессов – М.: Наука, 1974. – 696 с.

[14] *Малышев А.Н.* Введение в вычислительную линейную алгебру. – Новосибирск: Наука. Сиб. отд-ние. 1991. – 229 с.

[15] *Морозов В.А.* Алгоритмические основы методов решения некорректных задач//Вычисл. методы и программирование. 2003. Т. 45. С. 130–141.

[16] *Рао С.Р.* Линейные статистические методы и их применения: Пер. с англ. – М.: Наука, 1968. – 548 с.

[17] *Тихонов А.Н.* О приближенных системах линейных алгебраических уравнений//Ж. вычисл. матем. и матем. физ. 1980. Т. 20. № 6. С. 1373–1383.

[18] *Тихонов А.Н., Арсенин В.Я.* Методы решения некорректных задач: Учебное пособие. – М.: Наука, 1986. – 288 с.

[19] *Хорн Р., Джонсон Ч.* Матричный анализ: Пер. с англ. – М.: Мир, 1989. – 655 с.

[20] *Шевцов Г.С.* Линейная алгебра: Учебное пособие. – М.: Гардарики, 1999. – 360 с.

[21] *Björck Å.* Numerical stability of methods for solving augmented systems // Contemporary Math. 1997. V. 204. P.51–61.

[22] *Higham N.J.* Accuracy and Stability of Numerical Algorithms. – SIAM, Philadelphia, PA, 1996.

[23] *Penrose R.* A generalized inverse for matrices // Proc. Cambridge Philos. Soc. 1955. V. 51. P. 406–413.

[24] *Trefethen L.N., Bau D.* Numerical Linear Algebra. – SIAM, Philadelphia, 1997. – 373 p.