Федеральное государственное бюджетное учреждение науки Институт
вычислительной математики им. Г.И. Марчука Российской академии наук
(ИВМ РАН)

На правах рукописи

Осинский Александр Игоревич

Существование и построение близких к оптимальным столбцовых и крестовых аппроксимаций матриц

Специальность 1.1.6 — «Вычислительная математика»

Диссертация на соискание ученой степени
доктора физико-математических наук

Москва – 2025

Оглавление

Введение

Актуальность

Рассмотрим простой пример. Пусть требуется найти решение линейной системы $y = Ax$, где матрица $A = Z + E \in \mathbb{C}^{N\times N}$, $\operatorname{rank} Z = r \ll N$, а $\|E\| \ll \|A\|$ в некоторой норме. Пусть размер матрицы N достаточно велик, так что применение сингулярного разложения слишком затратно с точки зрения объема вычислений. Заметим, что если бы была известна матрица Z в виде

$$Z = UV, \quad U \in \mathbb{C}^{N\times r}, \quad V \in \mathbb{C}^{r\times N},$$

то приближенное решение можно было бы найти, например, решив задачу наименьших квадратов

$$\|y - UV\tilde{x}\|_2^2 \to \min_x,$$

нормальное псевдорешение которой можно получить за $O\left(Nr^2\right)$ операций, используя псевдообратные к матрицам U и V:

$$\tilde{x} = V^+U^+y.$$

На практике вместо прямого псевдообращения часто используется регуляризация перед обращением.

Данный пример является одним из многих применений *малоранговой аппроксимации* матриц. Он также показывает, что такие аппроксимации может потребоваться строить очень быстро: если мы не хотим существенно увеличить сложность алгоритма по сравнению с решением задачи наименьших квадратов, матрицу Z желательно находить за число операций, близкое к Nr^2. В случае $r^2 < N$ мы получаем довольно жесткое ограничение на алгоритм: мы не можем даже позволить ему «увидеть» все элементы матрицы. При этом необходимо максимально уменьшить размер погрешности E.

Сразу заметим, что в такой постановке (одновременное ограничение погрешности и вычислительной сложности) решения задачи не существует: какой бы алгоритм ни был, если ему на вход не передается вся матрица целиком, погрешность может быть сколь угодно велика, так как может находиться в элементах, которые не были рассмотрены.

Еще одним явным примером, когда особенно важно быстрое построение аппроксимации, является случай, когда матрица A представляет собой дискретизацию интегральных уравнений. В отличие от дифференциальных уравнений, котором соответствуют матрицы с небольшим числом ненулевых диагоналей, интегральные уравнения приводят к плотным матрицам, что значительно усложняет построения решения соответствующей линейной системы. Тем не менее, хотя матрица A в этом случае обычно полного ранга, она может быть представлена в иерархической блочной форме, где ранг каждого блока ограничен. Затем решение получается одним из итеративных методов: малоранговое представление блоков ускоряет выполнение умножения матрицы на вектор

в N/r раз, где r и N – ранг и размер соответствующего блока. Мозаично-скелетонный метод [1] строит аппроксимацию каждого блока за $O\left(Nr^2\right)$ операций, что сопоставимо с вычислительной сложностью $O\left(Nrk\right)$ использования данного блока в дальнейшем, где k – требуемое число итераций (умножений матрицы на вектор). Более медленные методы (требующие не менее N^2 операций за счет просмотра всего блока) привели бы к доминированию сложности построения аппроксимации в общей сложности.

Классическим примером алгоритма быстрого построения аппроксимации является неполное разложение Гаусса. Если алгоритм Гаусса остановить на шаге r, то полученное неполное LU разложение, где $L \in \mathbb{C}^{N\times r}$, $U \in \mathbb{C}^{r\times N}$ можно использовать в качестве малоранговой аппроксимации. При этом такое приближение, очевидно, строится всего за $O\left(Nr^2\right)$ операций.

Основной проблемой такого алгоритма является его численная неустойчивость. Устойчивость разложения Гаусса можно гарантировать только если на каждом шаге переставлять строки и столбцы так, чтобы новый опорный элемент был близок к максимальному во всей матрице. Это, в частности, гарантирует не слишком быстрый рост максимального элемента после каждого шага [2], но стоимость каждого шага растет до $O\left(N^2\right)$. На практике, однако, гораздо чаще встречается LUP разложение, где P – матрица перестановки. Таким образом, переставляются только столбцы, что может приводить к экспоненциальному росту ошибки [3]. Тем не менее широкое применение разложения Гаусса с частичным поиском опорного элемента, в том числе для построения малоранговых аппроксимаций [4, 5], говорит о том, что обычно не требуется рассматривать всю матрицу для достижения высокой точности аппроксимации. Естественно, есть и другие методы поиска ведущего элемента, как, например, метод хода ладьи (rook pivoting), где опорный элемент выбирается максимальным в своей строке и столбце. Это позволяет избежать экспоненциального роста элементов погрешности [6].

О чем все это говорит? Что часто для построения аппроксимации матрицы не требуется рассматривать все элементы для поиска оптимальных перестановок строк и столбцов. Это одновременно основное достоинство и недостаток подобных методов: с одной стороны, они имеют сложность меньше, чем размер входных данных. С другой стороны, незнание большей части входных данных не позволяет гарантировать точность методов. Вместо этого приходится использовать различные допущения о свойствах выбранных строк и столбцов, и уже с этими допущениями возможно получить оценки на точность аппроксимации. Тем не менее широкое использование LUP разложения и неполного LU разложения [7] говорит о высокой эффективности данного подхода, а также о перспективности его развития в терминах улучшения свойств аппроксимаций, основанных на небольшом числе строк и столбцов.

Все методы неполного LU разложения при этом можно рассматривать как частный случай псевдоскелетной CGR аппроксимации [8], где $C \in \mathbb{C}^{N\times r}$ и $R \in \mathbb{C}^{r\times N}$ – строки и столбцы матрицы A, а $G \in \mathbb{C}^{r\times r}$ – матрица генератора. В частности, если $G = \hat{A}^{-1}$ является обратной к матрице

на пересечении строк R и столбцов C, то такое разложение называется скелетным. Таким образом, задачу быстрого построения аппроксимации можно свести к поиску подходящих строк R и столбцов C матрицы A. Свойства аппроксимации при этом оказываются тесно связаны со свойствами подматрицы $\hat{A}$ на их пересечении. В частности, высокая точность по норме Чебышева достигается для подматриц, обладающих максимальным объемом (модулем определителя) [9, 10].

Изучение свойств скелетных и псевдоскелетных аппроксимаций – активно развивающаяся область, где появляются все более точные оценки аппроксимации и алгоритмы, позволяющие их достичь [11, 12, 13, 14, 15]. Крестовые аппроксимации активно применяются при построении иерархических [1, 16] и H2-матриц [17], решении интегральных уравнений [18, 19], уравнений Смолуховского [20], при отборе/выделении признаков [21], в задачах предобуславливания [22], при построении неотрицательных аппроксимаций [23], постобработке [24] и сжатии данных [25], поиске глобального максимума/минимума функций [26], численном вычислении и работе с гладкими функциями [27], в рекомендательных системах [28].

Поиск подматриц большого объема требуется при поиске оптимальных точек для полиномиального базиса [29], в методе внутренней точки [30], многосеточных методах [31], дизайне экспериментов [32], при выборе наилучшей обучающей выборки в машинном обучении [33], в коммуникационных системах при выборе лучей [34] и передающих антенн [35].

Построение крестовой аппроксимации также применяется при аппроксимации тензоров на основе тензорных поездов [36, 37] и разложения Таккера [38].

Постановка проблемы

В общем случае задача крестовой аппроксимации заключается в поиске n столбцов $C \in \mathbb{C}^{M \times n}$ и m строк $R \in \mathbb{C}^{m \times N}$ матрицы $A \in \mathbb{C}^{M \times N}$ и матрицы-генератора $G \in \mathbb{C}^{m \times n}$ таких, что $\|A - CGR\|$ мала. Особенно интересны при этом оценки, близкие к оптимальным:

$$\|A - CGR\| \leqslant (1 + \varepsilon) \min_{Z,\ \mathrm{rank}\ Z \leqslant r} \|A - Z\| \tag{0.1}$$

или

$$\|A - CGR\| \leqslant C \min_{Z,\ \mathrm{rank}\ Z \leqslant r} \|A - Z\|, \tag{0.2}$$

где $\|\cdot\|$ – некоторая матричная норма, r – требуемый ранг аппроксимации (для которого наилучшая аппроксимация дает погрешность не выше требуемого уровня), а ε и C, вообще говоря, зависят от размеров матрицы A и генератора G. Естественно, особенно интересны оценки, где величины ε и C невелики или близки к минимально возможным для соответствующей нормы.

При этом крайне важно иметь возможность быстрого построения подобных аппроксимаций. Если ограничить число операций порядком $O\,(MNr)$, то наиболее выгодными оказываются неполное QR разложение [39] или приближенное SVD с помощью случайного проектирования

[40]. Если же и это слишком затратно, тогда применяется неполное LU разложение на основе адаптивного крестового метода (Cross 2D) [4] (в иностранной литературе этот метод известен как adaptive cross approximation, ACA [5]) или алгоритма maxvol [11], что позволяет строить крестовую аппроксимацию за $O\left((M+N)\,r^2\right)$ операций. Еще одним важным свойством крестовых аппроксимаций является использование небольшого числа элементов матрицы A, что позволяет не вычислять всю матрицу A целиком, если она дана сложной формулой, и её вычисление является крайне затратным. В связи с этим генератор G обычно выбирается на основе подматрицы $\hat{A} \in \mathbb{C}^{m\times n}$ на пересечении строк R и столбцов C. Таким образом, свойства аппроксимации напрямую зависят от свойств подматрицы $\hat{A}$.

Одним из свойств, гарантирующих высокую точность аппроксимации, является объем подматрицы $\mathcal{V}\left(\hat{A}\right)$, равный, в общем случае, произведению её сингулярных чисел. Потому алгоритмы построения крестовых аппроксимаций часто основаны на поиске подматрицы наибольшего объема.

Такой подход, однако, эффективен с точки зрения оценок вида (0.2) только если требуемый ранг r совпадает с одним из размеров подматрицы $\hat{A}$, причем величина C растет с ростом ранга аппроксимации. Как показано в данной работе, в том случае, если есть возможность выбора подматрицы большего размера (а такая возможность обычно есть), лучшие оценки можно получить, используя подматрицу большого r-проективного объема (равного произведению r наибольших сингулярных чисел подматрицы). Таким образом, возникает вопрос о точности аппроксимаций с числом строк и столбцов больших r, того, насколько выгодным является при этом использование подматриц большого проективного объема, а также можно ли такие подматрицы искать за разумное время.

Кроме того, стоит отметить, что многие методы и алгоритмы поиска сильно невырожденных подматриц, приведенные в литературе [7, 39, 41, 42, 43, 44], часто могут быть ускорены с использованием методов и оценок, аналогичных тем, что применяются для поиска подматриц локально максимального объема. Или для них аналогичным образом могут быть доказаны более точные оценки аппроксимации.

В качестве примеров задач, где особенно выгодно использование крестовых аппроксимаций, будут рассмотрены уравнения Смолуховского, построение неотрицательных приближений и восстановление матриц. В последних двух случаях необходимо построение крестовой аппроксимации, гарантирующей ε-точность по норме Фробениуса (0.1) с $\varepsilon \ll 1$, для чего r строк и столбцов оказывается недостаточно. Оказывается, что необходимой точности можно достичь, если строить аппроксимацию на основе подматрицы большого r-проективного объема с числом строк и столбцов, больших r. В связи с этим важно изучение свойств таких подматриц и соответствующих им крестовых аппроксимаций.

Цель и задачи исследования

Целью работы является построение быстрых алгоритмов крестовой аппроксимации матриц и обоснование высокой точности аппроксимаций соответствующего вида. Эту же цель можно сформулировать как построение общей теории крестовых и столбцовых аппроксимаций и рассмотрение соответствующих ей алгоритмов. Для её достижения необходимо было решить следующие задачи:

1. Доказать, что крестовые аппроксимации в целом могут достигать высокой точности в различных нормах.
2. Показать, что полученные оценки близки к оптимальным, сравнив их с оценками снизу.
3. Так как аппроксимации на основе подматриц локально максимального объема не дают высокой точности для всех матриц, необходимо рассмотреть вероятностную модель и показать, что принцип локально максимального объема позволяет строить аппроксимации высокой точности для большинства матриц.
4. Построить и оценить вычислительную сложность алгоритма поиска прямоугольных подматриц локально максимального объема, усовершенствовать существующие методы для квадратных подматриц и дальнейшего набора столбцов для них. Построить алгоритмы поиска подматриц большого проективного объема.
5. Проверить эффективность предложенных алгоритмов для построения крестовых аппроксимаций как на случайных матрицах, так и в различных задачах, где быстрая малоранговая аппроксимация играет важную роль.

Степень разработанности проблемы

Одной из первых работ, где была рассмотрена точность крестовых аппроксимаций на основе подматриц максимального объема является [8]. В ней впервые были предложены оценки точности на основе подматриц максимального объема по спектральной норме. Данные оценки, однако, малоинтересны из-за низкой эффективности крестовых аппроксимаций по спектральной норме в целом. Кроме того, для построения предложенных аппроксимаций требовалось знание сингулярного разложения, что сразу отметает практическую ценность данных оценок. Кроме того, поиск подматриц максимального объема является NP-сложной задачей [45], что делает данные оценки еще менее ценными. Тем не менее, предложенные в [8] оценки крестовой аппроксимации для аппроксимаций на основе r строк и столбцов оставались наилучшими известными оценками по спектральной норме.

Оценки крестовых аппроксимаций по норме Чебышева впервые были получены в [9, 10] для $r \times r$ подматриц и усовершенствованы соискателем для подматриц произвольного размера $m \times n$, $m, n \geqslant r$. Здесь также стоит упомянуть, что оценка из [9] была улучшена в [15], однако она все

равно уступает оценкам аппроксимаций, основанным на подматрицах локально максимального проективного объема.

Оценки столбцовых аппроксимаций по норме Фробениуса были подробно изучены в [43, 46]. Там же приведены нижние оценки, доказывающие асимптотическую оптимальность полученных результатов. Тем не менее, построение таких аппроксимаций требовало нескольких сингулярных разложений, что крайне неэффективно с вычислительной точки зрения. Алгоритм для r столбцов был усовершенствован в [47], однако все равно имел сложность, пропорциональную четвертой степени размера матрицы. До настоящей работы не было ясно, можно ли еще ускорить данный алгоритм или применить те же оценки к крестовым аппроксимациям. Существующие же оценки крестовых аппроксимаций по норме Фробениуса [13, 48, 49] были очень далеки от оптимальных, требовали существенных вычислительных ресурсов, а также слишком большого числа строк и столбцов.

Задача поиска существенно невырожденных подматриц рассматривалась во многих работах [50, 51], однако наилучшие известные алгоритмы (не требующие поиска локально максимального объема) были построены в [44]. Многие из них далее усовершенствованы в данной диссертации. Поиск подматриц локально максимального объема был впервые предложен в [39], в [52] была доказано оценка на число шагов алгоритма поиска, а в [53, 11] данный алгоритм был упрощен для поиска квадратных подматриц (хотя на практике использовался уже в [1]). В [42] был предложен алгоритм жадного набора столбцов. Именно на его основе соискателем в дальнейшем был построен алгоритм поиска прямоугольных подматриц локально максимального объема. В диссертации также получен асимптотически более быстрый вариант алгоритма из [42].

В данной диссертации рассматривается три возможных применения крестовых методов: ускорение решения температурно-зависимых уравнений Смолуховского, восстановление матриц и построение неотрицательных аппроксимаций. Методы малоранговой аппроксимации были впервые применены к уравнениям Смолуховского в [20]. В кандидатской диссертации соискателя было предложено использовать крестовые методы для решения температурно-зависимых уравнений Смолуховского [54], где ядро меняется со временем, а потому необходимо быстро обновлять аппроксимацию на каждом временном шаге. Метод восстановления матриц на основе переменных проекций с использованием сингулярного разложения был предложен в [55]. Взяв его за основу, можно существенно ускорить алгоритм, если использовать методы крестовой аппроксимации в качестве приближенного сингулярного разложения. Приближенное сингулярное разложение используется и может быть ускорено также и в других методах восстановления матриц, в частности, основанных на минимизации ядерной нормы [56]. Метод переменных проекций для неотрицательных аппроксимаций матриц и тензоров подробно разобран в [57]. Отметим, что крестовый метод также позволяет находить аппроксимации с неотрицательными факторами [58].

Описание методологии исследования

Точность столбцовых и крестовых аппроксимаций рассматривается в трех разных аспектах: с точки зрения верхних оценок, нижних оценок и вероятностных оценок.

Для построения верхних оценок точности часто предполагается, что некоторая аппроксимация Z ранга r заранее известна, и на её основе затем строится крестовая CGR аппроксимация. Зная точность аппроксимации Z затем можно оценить точность CGR аппроксимации на её основе. Такой подход не позволяет гарантировать существование быстрых алгоритмов поиска крестовых аппроксимаций, однако позволяет гарантировать достижимость аппроксимаций с определенной точностью.

Нижние оценки, наоборот, позволяют определить минимальную достижимую погрешность, также ничего не говоря о том, как её достичь. Они строятся путем поиска конкретных примеров матриц, для которых выбор любых столбцов приводит к высокой величине ошибки. Естественно, такие примеры должны быть достаточно универсальными, чтобы распространятся на произвольные размеры M и N приближаемой матрицы, произвольное число столбцов n и произвольный ранг r. Стоит отметить, что так как крестовые CGR аппроксимации являются частным случаем столбцовых CW аппроксимаций для $W = GR$, нижние оценки достаточно построить для столбцовых аппроксимаций. Как будет показано, особую роль при построении нижних оценок будут играть подматрицы унитарных матриц. Впервые данная связь была обнаружена в [8], и, используя аналогичные рассуждения, можно построить нижние оценки для столбцовых аппроксимаций.

Вероятностные оценки оказываются необходимы в связи с тем, что выбор подматрицы локально максимального объема или проективного объема гарантирует высокую точность аппроксимации по норме Чебышева, но не по норме Фробениуса. С другой стороны, на практике алгоритмы построения крестовых аппроксимаций на основе таких подматриц почти всегда приводят к оценкам вида (0.1) с $\varepsilon = \frac{r}{n-r+1}$ для подматриц размера $n \times n$. Чтобы обосновать наблюдаемую эффективность, предлагается использовать вероятностную модель, где сама матрица A является случайной. Для этого подойдет так называемый RANDSVD ансамбль, где сингулярные числа матрицы A являются произвольными и фиксированными, а сингулярные векторы – случайные, и распределены равномерно (согласно мере Хаара). Усреднение по всем возможным матрицам левых и правых сингулярных векторов тогда приводит к тому, что для большинства матриц погрешность крестовой аппроксимации оказывается мала.

Для построения крестовых аппроксимаций на практике предлагается использовать и усовершенствовать существующие алгоритмы поиска подматриц большого объема [11, 39, 42]. Если же дополнительно также набирать первые r строк/столбцов, чтобы максимизировать объем, можно воспользоваться оценками на объем такой стартовой подматрицы [29], чтобы ограничить число шагов при дальнейшем поиске локально максимального объема.

Основные результаты, выносимые на защиту

Основные результаты данной работы включают в себя новые нижние и верхние оценки точности крестовых и столбцовых аппроксимаций, эффективные алгоритмы построения крестовых и столбцовых аппроксимаций, оценки на число шагов построенных алгоритмов и на свойства подматриц, которые данные алгоритмы находят. На защиту выносятся следующие положения:

1. Использование подматриц локально максимального объема и проективного объема приводит к крестовым аппроксимациям, близким к оптимальным по норме Чебышева. При этом увеличение числа строк и столбцов всего в два раза уже ведет к улучшению асимптотической зависимости коэффициента погрешности от ранга аппроксимации.
2. Существуют скелетные аппроксимации, коэффициент относительной погрешности по норме Фробениуса которых не зависит от размеров приближаемой матрицы. Существуют столбцовые аппроксимации с коэффициентом погрешности, близким к его нижней границе для нормы Фробениуса.
3. Оценки точности столбцовой аппроксимации по спектральной норме совпадают со значением функции $t(r, n, N)$, определяемой свойствами подматриц унитарных матриц. Следствиями данного факта являются близкие друг к другу верхние и нижние оценки точности крестовых и столбцовых аппроксимаций по спектральной норме.
4. Поиск подматриц из принципа максимизации объема и проективного объема в матрицах из randsvd ансамблей (любая матрица принадлежит некоторому randsvd ансамблю) позволяет с большой вероятностью гарантировать, что точность соответствующих крестовых и столбцовых аппроксимаций по норме Фробениуса будет близка к точности аппроксимации на основе сингулярного разложения.
5. Подматрицы, обладающие ρ-локально максимальным объемом или проективным объемом могут быть найдены за полиномиальное время. Требуемое число замен строк и столбцов при этом зависит только от размеров подматрицы и коэффициента ρ.
6. Крестовые методы существенно ускоряют решение задач, требующих многократного построения аппроксимаций матриц. В частности, алгоритмы позволяют ускорить численное решение обобщенных уравнений Смолуховского и восстановление матриц.

Научная новизна

Научная новизна диссертации заключается в следующем.

- Полученные оценки существенно улучшают известные оценки крестовых и столбцовых аппроксимаций. Это, в частности, удалось сделать путем введения понятия проективного объема и его использования для построения соответствующих псевдоскелетных аппроксимаций.

- Вероятностные оценки точности являются новым способом оценки эффективности алгоритмов, для которых результат зависит от конкретных входных данных. В частности, это позволяет оценить эффективность в тех случаях, когда необходимо получить аппроксимацию, используя лишь малую часть входных данных. Что и было сделано для случаев поиска подматриц локально максимального объема и проективного объема.

- Предложенные алгоритмы позволяют достичь низких оценок на нормы $\left\|\hat{A}^{+}\right\|_2$ и $\left\|\hat{A}^{+}\right\|_F$ для найденной подматрицы $\hat{A}$ существенно быстрее большинства других аналогичных методов. По скорости они уступают лишь рандомизированным методам, требующим существенно большего числа строк и столбцов (что в итоге все равно приводит к большей вычислительной сложности построения аппроксимации) и гарантирующим оценки на $\left\|\hat{A}^{+}\right\|_2$ и $\left\|\hat{A}^{+}\right\|_F$ лишь с некоторой вероятностью.

- Предложенные алгоритмы позволяют строить приближение с коэффициентом погрешности по норме Фробениуса $1+\varepsilon$ за $O\left(Nr^2/\varepsilon + (r/\varepsilon)^3\right)$ операций, чего не могут достичь никакие другие известные алгоритмы. Среди алгоритмов, использующих порядка r/ε строк и столбцов, псевдоскелетная аппроксимация на основе подматриц локально максимального проективного объема содержит (в среднем) наименьший возможный коэффициент ошибки.

- Предложенные алгоритмы восстановления матриц и построения неотрицательных аппроксимаций асимптотически быстрее версий, основанных на других методах приближенного вычисления сингулярного разложения.

Теоретическое и практическое значение

Теоретическая значимость полученных результатов заключается в усовершенствовании существующих оценок столбцовых и крестовых аппроксимаций. При этом удалось максимально сблизить верхние и нижние оценки как по спектральной норме, так и по норме Фробениуса. Применение той же методологии к аппроксимации по норме Чебышева позволяет не только предсказывать эффективность методов крестовой аппроксимации, но и получить нижние оценки аппроксимации единичной матрицы. Построение вероятностных оценок является новым способом анализа эффективности методов малоранговой аппроксимации, для которых не существует гарантий на точность аппроксимации на произвольных матрицах. Ограничение числа шагов в алгоритмах поиска подматриц локально максимального также позволяет быть уверенным в их быстрой работе на практике.

Практическая значимость заключается в том, что построенные алгоритмы позволяют быстро находить крестовые аппроксимации высокой точности по норме Фробениуса, спектральной норме и норме Чебышева. Алгоритмы быстрой крестовой аппроксимации матриц играют важную

роль в аппроксимации тензоров с помощью тензорных поездов (TT-cross метод). Крестовые аппроксимации также позволяют существенно ускорить решения систем интегральных уравнений, в том числе рассматриваемых в диссертации уравнений Смолуховского. Быстрое построение аппроксимаций высокой точности применяется, например, в латентном семантическом анализе и распознавании лиц, а также позволяет строить быстрые алгоритмы восстановления матриц, которые используются, в частности, в рекомендательных системах, при поиске/восстановлении фазы и построении разреженной оценки канала в задачах обработки сигнала, сжатии и восстановлении поврежденных изображений. Кроме задачи восстановления матриц, в диссертации также рассматривается построение неотрицательных аппроксимаций, которые используются при анализе и распознавании изображений, а также при аппроксимации решения дифференциальных и интегральных уравнений, где появление отрицательных элементов может приводить к неустойчивости последующего решения.

Достоверность и апробация результатов

Полученные теоретические результаты основаны на строгом математическом выводе всех положений. Диссертация содержит строгий и подробный вывод всех используемых в алгоритмах формул, гарантирующих достижение требуемых выходных данных, а также вывод оценок на число шагов алгоритмов. Эффективность алгоритмов и точность доказанных оценок подтверждается большим количеством численных экспериментов на случайных и реальных данных, а также успешном применении построенных аппроксимаций для решения физических задач. Положения и выводы, сформулированные в диссертации, получили квалифицированную апробацию на международных научных конференциях. Достоверность также подтверждается публикацией результатов исследований в рецензируемых научных журналах.

Публикации. По теме диссертации опубликовано 15 работ, из них 15 в журналах, входящих в перечень рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание учёной степени доктора наук, в том числе 15 публикаций в изданиях, включенных в международные системы цитирования WoS и/или Scopus [A.1-A.15].

Личный вклад:

1. **Публикация A.1.** Полностью написана соискателем. Соискателем построен малоранговый Монте-Карло метод для уравнений Смолуховского.
2. **Публикация A.2.** Полностью написана соискателем. Соискателем был построен алгоритм поиска прямоугольных подматриц локально максимального объема, получены оценки на его число шагов и ограничения на нормы псевдообратной к полученной подматрице, показана его эффективность на практике.

3. **Публикация А.3.** Полностью написана соискателем. Соискателем получены нижние оценки точности столбцовых аппроксимаций по спектральной норме и норме Фробениуса. Показано, что верхние и нижние оценки по спектральной норме совпадают с соответствующими значениями t-функции.
4. **Публикация А.4.** Полностью написана соискателем. Соискателем был построен и описан метод поиска подматриц локально максимального объема с одновременными заменами строк и столбцов. Получены оценки на число шагов, доказано достижение сильного LU разложения и высокой точности по норме Чебышева, проведено сравнение с другими алгоритмами построения выявляющих ранг LU разложений.
5. **Публикация А.5.** Соискателем был построен и описан малоранговый Монте-Карло метод для температурно-зависимых уравнений, получены точные аналитические решения, проведены численные эксперименты.
6. **Публикация А.6.** Соискателем был построен и описан метод принятия-отклонения в применении к задачам с агрегацией и фрагментацией на мономеры, выполнены все связанные с ним численные эксперименты.
7. **Публикация А.7.** Соискателем проведены численные эксперименты с использованием малоранговых методов решения уравнений Смолуховского, выведены асимптотики решений.
8. **Публикация А.8.** Соискателем был получен первоначальный вариант доказательства возможности использования алгоритмов приближенного SVD для алгоритма singular value projection, предложен вариант использования крестовой аппроксимации в данном методе, написан сам алгоритм и проведены численные эксперименты для него.
9. **Публикация А.9.** Соискателем получен вывод оценок точности крестовых и столбцовых аппроксимаций в среднем, проведены численные эксперименты, обнаружена и численно подтверждена связь между точностью аппроксимации и свойствами подматриц ортогональных матриц.
10. **Публикация А.10.** Полностью написана соискателем. Соискателем построен малоранговый метод решения температурно-зависимых ОДУ типа Смолуховского, предложен быстрый способ использования адаптивного временного шага, а также аппроксимации хвоста распределения концентраций.
11. **Публикация А.11.** Соискателем была доказана сходимость алгоритма maxvol для частного случая аппроксимации ранга 1, оценена скорость сходимости, точность полученной аппроксимации, близость найденного элемента к максимальному, доказана возможность применения алгоритма при аппроксимации тензоров.
12. **Публикация А.12.** Полностью написана соискателем. Оценки крестовой аппроксимации по норме Чебышева были обобщены на случай тензоров.

13. **Публикация А.13.** Соискателем доказано существование крестовых аппроксимаций высокой точности по норме Фробениуса.
14. **Публикация А.14.** Соискателем упрощены доказательства для точности столбцовой и крестовой аппроксимаций по спектральной норме, доказаны оценки точности крестовой аппроксимации по норме Чебышева, проведены численнные эксперименты, предложен и реализован алгоритм поиска подматриц локально максимального проективного объема.
15. **Публикация А.15.** Соискателем получены оценки точности крестовой аппроксимации по норме Чебышева, предложен сам вид крестовой аппроксимации и требования к выбору подматрицы, позволяющие гарантировать полученные оценки.

Результаты работы были представлены на ведущих российских и международных конференциях:

1. Osinsky A. I. A probability proof for the rank-one cross approximation method for matrices and tensors // The Sixth China-Russia Conference on Numerical Algebra with Applications (CRCNAA 2017). — August 2017.

2. Желтков Д. А., Осинский А. И. Исследование метода крестовой оптимизации в многомерных задачах ранга 1 // Ломоносовские чтения-2018, секция «Вычислительная математика и кибернетика», МГУ имени М.В. Ломоносова, Россия, 16-27 апреля 2018. Опубликовано в сборнике Ломоносовские чтения 2018 ф-т ВМК МГУ, место издания Макс-Пресс, тезисы, с. 58-59.

3. Осинский А. И., Лебедева О. С., Петров С. В. Приближенные алгоритмы малоранговой аппроксимации в задаче восстановления матрицы по элементам на случайном шаблоне // Ломоносовские чтения-2018, секция «Вычислительная математика и кибернетика», МГУ имени М.В. Ломоносова, Россия, 16-27 апреля 2018. Опубликовано в сборнике Ломоносовские чтения 2018 ф-т ВМК МГУ, место издания Макс-Пресс, тезисы, с. 76-77.

4. Osinsky A. I. Cross method accuracy estimates in consistent norms // SIAM Conference on Applied Linear Algebra (SIAM-ALA18). — May 2018.

5. Osinsky A. I., Zamarashkin N. L. Probabilistic estimates for matrix cross approximation // The 5th International Conference on Matrix Methods in Mathematics and applications (MMMA 2019). — August 2019.

6. Petrov S. V., Osinsky A. I. Low-Rank Approximation Algorithms for Matrix Completion with Random Sampling // The 5th International Conference on Matrix Methods in Mathematics and applications (MMMA 2019). — August 2019.

7. Kalinov A., Matveev S., Osinsky A. Direct simulation Monte Carlo and oscillations in aggregation-fragmentation kinetics // 15th International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing. — July 2022.

8. Осинский А. И. Быстрый поиск неотрицательных аппроксимаций матриц с помощью крестового разложения // Ломоносовские чтения-2023, секция «Вычислительные технологии и моделирование», МГУ имени М.В. Ломоносова, Россия, 4-14 апреля 2023. Опубликовано в сборнике Ломоносовские чтения 2023 ф-т ВМК МГУ, место издания Макс-Пресс, тезисы, с. 125-127.

9. Osinsky A. Close to optimal column approximation from SVD // The 6th international conference on matrix methods and machine learning in mathematics and applications (MMMA 2023). — August 2023.

10. Осинский А. И. Поиск ρ-локально максимального объёма за полиномиальное время // Матричные методы и интегральные уравнения, Сириус. — Август 2023.

11. Осинский А. И. Достижимое за полиномиальное время приближение в задаче поиска подматриц максимального объема // Ломоносовские чтения-2024, секция «Вычислительные технологии и моделирование», МГУ имени М.В. Ломоносова, Россия, 20 марта - 3 апреля 2024. Опубликовано в сборнике Ломоносовские чтения 2024 ф-т ВМК МГУ, место издания Макс-Пресс, тезисы, с. 161-163.

12. Осинский А. И. Крестовые аппроксимации на основе подматриц большого проективного объема // IV Конференция математических центров России, секция «Прикладная математика и математическое моделирование», Санкт-Петербургский международный математический институт имени Леонарда Эйлера. — Август 2024.

13. Осинский А. И. Компромисс между скоростью и точностью в адаптивном крестовом методе // Матричные методы и интегральные уравнения, Сириус. — Август 2024.

Приглашенные доклады:

Osinsky A. I. Low-rank Monte-Carlo for temperature dependent Smoluchowski equations // Stochastic processes and Pattern Formation, Skoltech. — 13 September 2019.

Результаты работы обсуждались на семинарах:

- Объединенный научный семинар ИМ СО РАН, ИВМиМГ СО РАН, МЦА, кафедра ММГФ ММФ НГУ «Прикладные обратные задачи и искусственный интеллект», 2024 (онлайн).
- Семинар международной лаборатории стохастических алгоритмов и анализа многомерных данных факультета компьютерных наук НИУ ВШЭ, 2024.
- Семинар лаборатории «Многомерная аппроксимация и приложения» механико-математического факультета МГУ, 2024.
- Семинар «Теория функций» кафедры общих проблем управления механико-математического факультета МГУ, 2024.
- Семинар центра прикладного ИИ, цикл «Математика ИИ», Сколтех, 2024.

- Семинар «вычислительная математика и приложения» ИВМ РАН, 2024.
- Общеинститутский семинар ИПМ РАН, 2024.

Структура и объем диссертации

Работа состоит из введения, 7 глав, заключения и приложения. Полный объём работы составляет 252 страницы. Работа включает 17 рисунков и 17 таблиц. Список литературы содержит 154 наименования.

Благодарности

Автор выражает признательность за ценные обсуждения результатов диссертации и их презентации с Николаем Леонидовичем Замарашкиным. Автор также выражает благодарность Евгению Евгеньевичу Тыртышникову за поддержку и постоянно высокую оценку результатов.

Глава 1. Общие сведения о столбцовых и крестовых аппроксимациях

Для построения малоранговых аппроксимаций матриц выгодно использовать крестовые CGR (или CUR) аппроксимации, основанные на небольшом числе столбцов $C \in \mathbb{C}^{M\times n}$ и строк $R \in \mathbb{C}^{m\times N}$ приближаемой матрицы $A \in \mathbb{C}^{M\times N}$.

Определение 1.1. *Крестовой* или *псевдоскелетной* аппроксимацией матрицы $A \in \mathbb{C}^{M\times N}$ называется произведение CGR, где $C \in \mathbb{C}^{M\times n}$ – некоторые n столбцов матрицы A, а $R \in \mathbb{C}^{m\times N}$ – m её строк. Матрица $G \in \mathbb{C}^{n\times m}$ называется генератором аппроксимации.

В случае, если матрица G квадратная $m = n$ и совпадает с обратной к матрице $\hat{A} \in \mathbb{C}^{m\times n}$ на пересечении строк R и столбцов C, $G = \hat{A}^{-1}$, то такая аппроксимация называется *скелетной*.

Крестовое приближение является частным случаем *столбцового* приближения матриц.

Определение 1.2. Столбцовым приближением матрицы $A \in \mathbb{C}^{M\times N}$ называется произведение CW, где $C \in \mathbb{C}^{M\times n}$ – некоторые n столбцов матрицы A, а $W \in \mathbb{C}^{n\times N}$ называется матрицей весов.

Везде далее мы будем использовать букву C для обозначения столбцов, а букву R для обозначения строк.

При построении крестовых и столбцовых аппроксимаций ранга r будем предполагать, что $\operatorname{rank} G = r$ и $\operatorname{rank} W = r$ соответственно.

Столбцовые аппроксимации полезны не только как обобщение крестовых аппроксимаций (что позволяет использовать нижние оценки для них в качестве нижних оценок крестовых аппроксимаций), но они также позволяют строить эффективные крестовые аппроксимации, используя идеи для выбора столбцов также и для выбора строк. Такой подход можно встретить во многих работах, например в [7, 8, 13].

1.1. Связь с неполными QR и LU разложениями

Скелетное $C\hat{A}^{-1}R$ приближение имеет важную интерпретацию: оно задается неполным LU разложением матрицы A с переставленными соответствующим образом строками и столбцами. Другими словами, оно задается первыми r шагами исключения Гаусса. При этом позиции ведущих элементов соответствуют диагонали подматрицы $\hat{A}$.

Таким образом, задачу поиска $\hat{A}$ можно рассматривать в терминах построения соответствующего ей неполного LU разложения. Как уже было сказано, неполное LU разложение может служить критерием малоранговости матрицы, если строки и столбцы подобрать соответствующим образом. Такие разложения называют rank revealing (выявляющими ранг) LU разложениями

или, сокращенно, RRLU. Аналогично, для выявляющих ранг QR разложений используют аббревиатуру RRQR.

Неполное QR разложение задает столбцовую аппроксимацию специального вида, а именно $CC^{+}A = Q\,(Q^{*}A)$, которую мы часто будем встречать в дальнейшем.

Рассмотрим некоторые известные оценки и свойства RRQR и RRLU разложений. В данных определениях под словом «разложение» мы также будем подразумевать и некоторый алгоритм, позволяющий его построить. В зависимости от того, какие оценки алгоритм позволяет гарантировать, соответствующее разложение может относиться к нескольким типам.

Определение 1.3 ([39])**.** Неполное QR разложение с выбором ведущих столбцов

$$A\Pi = Q \begin{bmatrix} R & B \\ & C \end{bmatrix} \in \mathbb{C}^{M\times N},$$

где матрица $Q \in \mathbb{C}^{M\times M}$ унитарная, а матрица $R \in \mathbb{C}^{r\times r}$ верхняя треугольная называется *выявляющим ранг* (rank revealing), если

$$\sigma_r(R) \geqslant \sigma_r(A)/p_1(r,N)$$

и

$$\sigma_1(C) \leqslant \sigma_{r+1}(A)p_2(r,N),$$

где $p_1(r,N)$ и $p_2(r,N)$ – функции, ограниченные полиномами от r и N.

Оно называется *выявляющим спектр* (spectrum revealing), если дополнительно

$$\sigma_i(R) \geqslant \sigma_i(A)/p_1(r,N), \quad i = \overline{1,r-1}.$$

Оно называется *сильным* (strong), если оно выявляет спектр и дополнительно

$$\sigma_i(C) \leqslant \sigma_{r+i}(A)p_2(r,N), \quad i = \overline{1,\min(M-r,N-r)}$$

и

$$\|R^{-1}B\|_C \leqslant f = const.$$

Аналогичные определения существуют для неполного LU разложения.

Определение 1.4. Неполное LU разложение с выбором ведущих элементов

$$\Pi_1 A \Pi_2 = \begin{bmatrix} L_{11} & \\ L_{21} & I_{M-r} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ & U_{22} \end{bmatrix} \in \mathbb{C}^{M\times N}, \tag{1.1}$$

где $L_{11} \in \mathbb{C}^{r\times r}$ – нижняя унитреугольная матрица, а $U_{11} \in \mathbb{C}^{r\times r}$ – верхняя треугольная матрица, называется *выявляющим ранг* (rank revealing), если

$$\sigma_r(L_{11}U_{11}) \geqslant \sigma_r(A)/p_1(r,M,N)$$

и

$$\sigma_1(U_{22}) \leqslant \sigma_{r+1}(A)p_2(r, M, N),$$

где $p_1(r, M, N)$ и $p_2(r, M, N)$ – функции, ограниченные полиномами от r, M и N.

Оно называется *выявляющим спектр* (spectrum revealing) [14], если дополнительно

$$\sigma_i(L_{11}U_{11}) \geqslant \sigma_i(A)/p_1(r, M, N), \quad i = \overline{1, r-1}.$$

Оно называется *сильным* (strong) [59], если оно выявляет спектр и дополнительно

$$\sigma_i(U_{22}) \leqslant \sigma_{r+i}(A)p_2(r, M, N), \quad i = \overline{1, \min(M-r, N-r)},$$

$$\|L_{21}L_{11}^{-1}\|_C \leqslant f = const$$

и

$$\|U_{11}^{-1}U_{12}\|_C \leqslant f = const.$$

Построение неполных QR или LU разложений можно считать эквивалентным выбору строк и столбцов, так как после этого само построение занимает $O\left(\mathrm{nnz}\,(\mathrm{A})\,r + Mr^2\right)$ операций в случае неполного QR разложения (через $\mathrm{nnz}\,(\mathrm{A})$ мы обозначим число ненулевых элементов в матрице A) и $O\left((M+N)\,r^2\right)$ операций в случае неполного LU разложения, что асимптотически меньше времени работы всех известных RRQR и RRLU алгоритмов. Таким образом, например, задачу построения сильного RRQR разложения можно свести к поиску «хорошей» подматрицы (как мы увидим далее, такая подматрица должна обладать локально максимальным объемом) и доказательству для столбцовой аппроксимации на её основе свойств из определения (1.3).

Отметим, что применение сокращенного сингулярного разложения не портит свойства выявления спектра (так как не меняет первых r сингулярных чисел, а лишь зануляет остальные), что может помочь дополнительно сократить размер разложения, не увеличив существенно его погрешность [14]. Его можно применить к CGR аппроксимации общего вида следующим образом. Сначала производится QR разложение строк и столбцов:

$$C = Q_1R_1, \quad R^* = Q_2R_2, \quad G' = R_1GR_2^*.$$

Для матрицы G' строится сингулярное разложение

$$G' = U\Sigma V, \quad U \in \mathbb{C}^{m\times\min(m,n)}, \Sigma \in \mathbb{C}^{\min(m,n)\times\min(m,n)}, V \in \mathbb{C}^{\min(m,n)\times n}.$$

Далее данное разложение сокращается до ранга r: для этого выбираются подматрицы $\hat{U}$, $\hat{\Sigma}$ и $\hat{V}$, соответствующие наибольшим сингулярным числам. Такое сокращенное сингулярное разложение будем обозначать нижним индексом r:

$$G'_r = \hat{U}\hat{\Sigma}\hat{V}, \quad U \in \mathbb{C}^{m\times r}, \Sigma \in \mathbb{C}^{r\times r}, V \in \mathbb{C}^{r\times n}.$$

В итоге получаем аппроксимацию $\tilde{A} = (CGR)_r$, факторы которой имеют размеры $M \times r$ и $r \times N$ соответственно:

$$\tilde{A} = \left(Q_1 \hat{U}\right) \left(\hat{\Sigma} \hat{V} Q_2^*\right),$$

что займет в общей сложности $O\left(Mnr + Nmr + mn \min\left(m, n\right)\right)$ операций. Так как само построение аппроксимации занимает не меньше времени, это оказывается крайне выгодно, когда сингулярные числа матрицы A быстро убывают. А именно, справедлива следующая оценка [14, 60], где мы ввели проектор на первые r сингулярных векторов матрицы A, $P_r A = A_r$:

$$\begin{aligned}
\|A - (CGR)_r\|_F &\leqslant \|A - CGR\|_F + \|CGR - (CGR)_r\|_F \\
&\leqslant \|A - CGR\|_F + \|CGR - P_r CGR\|_F \\
&\leqslant \|A - CGR\|_F + \|(A - CGR) - P_r(A - CGR)\|_F + \|A - P_r A\|_F \\
&\leqslant 2\,\|A - CGR\|_F + \|A - A_r\|_F .
\end{aligned}$$

Если сингулярные числа A быстро убывают, то $\|A - CGR\|_F \sim \|A - A_n\|_F \ll \|A - A_r\|_F$, а потому такой прием может существенно уменьшить относительную погрешность разложения по сравнению с построением аппроксимации на r строках и столбцах.

1.2. Связь с задачей одновременной аппроксимации

Задача построения оптимальной столбцовой аппроксимации произвольной матрицы $A \in \mathbb{C}^{M \times N}$ по норме Фробениуса

$$\left\|A - CC^{+}A\right\|_F \to \min_{C \in \mathbb{C}^{M \times n}},$$

часто называемая задачей выбора подмножества столбцов (column subset selection problem, CSS, CSSP) [61], может быть также переформулирована в виде задачи одновременной аппроксимации [62], также известной как задача разреженного представления нескольких векторов измерений (sparse representation of multiple measurement vectors, MMV) [63]. В последнем случае в основном рассматривается вопрос возможности точного восстановления, нежели возможность малоранговой аппроксимации. При этом последняя задача является более общей, поскольку в ней рассматривается случай, когда приближаемая матрица (из векторов измерений) не обязана совпадать с матрицей, из которой выбираются столбцы (называемой словарем). А именно, в общем случае можно решать задачу

$$\|A - C_D W\|_F \to \min_{C_D \in \mathbb{C}^{M \times n}}, \tag{1.2}$$

где столбцы C_D выбираются из некоторого словаря D (не обязательно конечного), а $W \in \mathbb{C}^{n \times N}$ – произвольная матрица весов.

Интересно, что популярный жадный алгоритм из [64] оказывается полностью эквивалентен и записывается точно так же, как и ортогональный жадный алгоритм для одновременной аппроксимации, также известный как MMP (modified marching persuit) или M-OMP (modified orthogonal

matching persuit) [65]. Использовать его, однако, следует аккуратно, поскольку он использует одноранговый пересчет матрицы E^*E, $E = A - C_D W$, что приводит к потери половины точности. Эту проблему можно решить, вычислив соответствующие элементы E^*E напрямую, когда они упадут достаточно сильно (в машинную точность раз), повторив идею, используемую при выборе ведущих столбцов в QR разложении [66]. Другие жадные алгоритмы решения данной задачи рассмотрены в [62]. В частности, чистый жадный алгоритм вместо оптимального $W = C_D^+ A$ выбирает веса итеративно, так что веса всех предыдущих столбцов фиксированы при добавлении нового.

Стоит, однако, отметить, что в теоретических результатах, посвященных задаче (1.2), чаще всего рассматривается аппроксимация относительно нормы матрицы A или подразумевается, что приближаемые векторы (столбцы A) лежат в выпуклой оболочке векторов словаря, которые, в свою очередь, нормированы на 1. Например, в [67] для ортогонального жадного алгоритма получена оценка

$$\left\|A - CC^+A\right\|_F^2 \leqslant \min_{\tilde{C}\in\mathbb{C}^{M\times n}} \left\|A - \tilde{C}\tilde{C}^+A\right\|_F^2 + \frac{16r}{n} \|A\|_F^2 / \sigma_n\left(C'\right),$$

где столбцы C' нормированы на 1, $C'_{:,i} = C_{:,i} / \left\|C_{:,i}\right\|_2$.

Нас же здесь будут интересовать оценки вида

$$\left\|A - CC^+A\right\|_F \leqslant f\left(r, n, M, N\right) \|A - A_r\|_F, \tag{1.3}$$

где коэффициент f зависит только от размеров матрицы, числа столбцов и ранга аппроксимации, а приближение сравнивается с сокращенным сингулярным разложением ранга $r \leqslant n$ (что соответствует выбору n столбцов из бесконечного словаря), поскольку на практике часто встречается именно случай, когда данные с высокой точностью являются малоранговыми. Таким образом, нас далее будет интересовать вопрос: насколько наилучшая одновременная аппроксимация с помощью конечного словаря близка к аппроксимации с помощью бесконечного словаря? Оценки вида (1.3) будут рассмотрены в разделе 2.3.1.

1.3. Вид наилучших аппроксимаций

Напомним одну из основных теорем линейной алгебры.

Теорема 1.1 (Эккарт-Янг-Мирский)**.** *Для произвольной матрицы $A \in \mathbb{C}^{M\times N}$, решением задачи минимизации*

$$\left\|A - \tilde{A}\right\|_{2,F} \to \min_{\operatorname{rank}\tilde{A}\leqslant r}$$

по спектральной норме или норме Фробениуса (или любой другой унитарно инвариантной норме) является матрица $\tilde{A} = A_r$, полученная из сокращенного сингулярного разложения матрицы A.

Она говорит о том, что наилучшая аппроксимация ранга r по спектральной норме и норме Фробениуса достигается с помощью сокращенного сингулярного разложения.

Пусть теперь, например, мы построили некое крестовое приближение $\tilde{A} = CGR$ ранга выше r. Будет ли приближение на основе сокращенного сингулярного разложения $(CGR)_r$ наилучшим крестовым приближением ранга r при тех же строках и столбцах? Оказывается, что если матрица G (или, в случае столбцовых аппроксимаций, матрица W) выбрана оптимальным образом, и речь идет о норме Фробениуса, то такое утверждение оказывается справедливым. Покажем это.

Во-первых, заметим, что оптимальным выбором матрицы W для минимизации нормы Фробениуса ошибки является $W = C^+A$. Для этого рассмотрим задачу наименьших квадратов для произвольного i-го столбца $W_{:,i}$:

$$\left\|A_{:,i} - CW_{:,i}\right\|_2 \to \min .$$

Решением данной задачи будет $W_{:,i} = C^+A_{:,i}$. Объединяя все столбцы i вместе, получаем $W = C^+A$. Таким образом, для построения наилучшей аппроксимации по норме Фробениуса достаточно оптимизировать выбор столбцов:

$$\min_{C,W} \|A - CW\|_F = \min_{C} \left\|A - CC^+A\right\|_F .$$

Аналогично можно определить вид наилучшей аппроксимации ранга r на основе n столбцов $C \in \mathbb{C}^{n\times N}$ [68]. Для этого нам потребуется следующее утверждение. Запишем его сразу и для спектральной нормы.

Утверждение 1.1 (Матричная теорема Пифагора)**.** *Пусть матрицы $A \in \mathbb{C}^{M\times N}$ и $B \in \mathbb{C}^{M\times N}$ ортогональны: $A^*B = 0_{N\times N}$. Тогда*

$$\|A + B\|_F^2 = \|A\|_F^2 + \|B\|_F^2$$

и

$$\|A + B\|_2^2 \leqslant \|A\|_2^2 + \|B\|_2^2$$

Доказательство. Для нормы Фробениуса

$$\|A + B\|_F^2 = \operatorname{tr}\left((A + B)^* (A + B)\right) = \operatorname{tr}(A^*A) + \operatorname{tr}(B^*B) + 2\operatorname{Re}\operatorname{tr}(A^*B) = \|A\|_F^2 + \|B\|_F^2 + 0.$$

Для спектральной нормы выберем единичный вектор-столбец $u \in \mathbb{C}^N$, $\|u\|_2 = 1$, для которого

$$(A + B)\, u = \|A + B\|_2 .$$

Тогда для него же получим

$$\|A + B\|_2^2 = u^* (A + B)^* (A + B)\, u = u^*A^*Au + u^*B^*Bu \leqslant \|A^*A\|_2 + \|B^*B\|_2 = \|A\|_2^2 + \|B\|_2^2 .$$

□

Данное утверждение чаще всего применяется в случае, если A и B ненулевые в разных блоках (например, $A = [C_1\ 0]$, $B = [0\ C_2]$, $A + B = [C_1\ C_2]$) или если они расположены в ортогональных подпространствах (например, $A = QQ^*C$, $B = (I - QQ^*)\,C$, $A + B = C$, $Q^*Q = I$).

Рассмотрим QR разложение $C = QR$, $Q \in \mathbb{C}^{M\times n}$, $R \in \mathbb{C}^{n\times n}$. С его помощью задача поиска оптимальных C и W сводится к задаче поиска оптимальных Q и $W' = RW$. Нас теперь будет интересовать, что если дополнительно наложить ограничение $\operatorname{rank} W' = r$, чтобы ранг столбцовой аппроксимации был не выше r. Погрешность такой аппроксимации можно выразить как

$$\begin{aligned}\|A - QW'\|_F^2 &= \|(A - QQ^*A) + (QQ^*A - QW')\|_F^2 \\ &= \|A - QQ^*A\|_F^2 + \|QQ^*A - QW'\|_F^2 \\ &= \|A - QQ^*A\|_F^2 + \|Q^*A - W'\|_F^2\,,\end{aligned}$$

где мы воспользовались матричной теоремой Пифагора и инвариантностью нормы Фробениуса относительно унитарных преобразований.

Оптимальное значение W' полностью определяется вторым слагаемым. Так как погрешность по норме Фробениуса минимизируется сингулярным разложением, то $W' = (Q^*A)_r$, откуда получаем, что

$$CW = QW' = Q\,(Q^*A)_r = (QQ^*A)_r = \left(CC^+A\right)_r\,,$$

то есть наилучшая столбцовая аппроксимация ранга r достигается путем применения сокращенного сингулярного разложения к CC^+A аппроксимации.

Наилучшее значение генератора $G \in \mathbb{C}^{m\times n}$ при фиксированных строках R и столбцах C находится точно так же. Рассмотрим сразу общий случай, когда требуется $\operatorname{rank} G \leqslant r$. Он является более общим, так как случай отсутствия ограничений соответствует $r = \min(m, n)$. Рассмотрим QR разложения $C = Q_1R_1$ и $R^* = Q_2R_2$. Тогда задача сводится к поиску наилучшей матрицы $G' = R_1GR_2^*$, минимизирующей погрешность

$$\begin{aligned}\left\|A - Q_1G'Q_2^*\right\|_F^2 &= \left\|A - Q_1Q_1^*A + Q_1Q_1^*A - Q_1G'Q_2^*\right\|_F^2 \\ &= \left\|A - Q_1Q_1^*A\right\|_F^2 + \left\|Q_1Q_1^*A - Q_1G'Q_2^*\right\|_F^2 \\ &= \left\|A - Q_1Q_1^*A\right\|_F^2 + \left\|Q_1^*A - G'Q_2^*\right\|_F^2\,,\end{aligned}$$

где мы снова воспользовались матричной теоремой Пифагора и унитарной инвариантностью. Теперь воспользуемся теми же свойствами в пространстве строк.

$$\begin{aligned}\left\|Q_1^*A - G'Q_2^*\right\|_F^2 &= \left\|Q_1^*A - Q_1^*AQ_2Q_2^* + Q_1^*AQ_2Q_2^* - G'Q_2^*\right\|_F^2 \\ &= \left\|Q_1^*A - Q_1^*AQ_2Q_2^*\right\|_F^2 + \left\|Q_1^*AQ_2Q_2^* - G'Q_2^*\right\|_F^2 \\ &= \left\|Q_1^*A - Q_1^*AQ_2Q_2^*\right\|_F^2 + \left\|Q_1^*AQ_2 - G'\right\|_F^2\,.\end{aligned}$$

Теперь матрица G' осталась лишь в последнем слагаемом, размер которого минимален, когда $G' = \left(Q_1^*AQ_2\right)_r$ получена из сокращенного сингулярного разложения. В итоге оптимальная

крестовая аппроксимация имеет вид

$$CGR = Q_1 \left(Q_1^* A Q_2\right)_r Q_2^* = \left(Q_1 Q_1^* A Q_2 Q_2^*\right)_r = \left(CC^+ AR^+ R\right)_r,$$

а в случае отсутствия ограничения на ранг $G = C^+ AR^+$.

Заметим, однако, что на практике построение оптимальной крестовой аппроксимации требует как минимум $O\left(MN \min\left(m, n\right)\right)$ операций, что существенно более затратно, чем использование скелетного разложения вида $C\hat{A}^{-1}R$. Это, в частности, означает, что применение сокращенного сингулярного разложения не всегда является оптимальным способом понизить ранг крестовой аппроксимации, см., например, численные эксперименты в [69]. Поэтому вместо $CC^+ AR^+ R$ сокращенное сингулярное разложение имеет смысл применять, в общем случае, к $C\hat{A}_k^+ R$, $r \leqslant k \leqslant n$, что позволяет одновременно уменьшить сложность и (возможно) увеличить точность аппроксимации. В [69] авторами было предложено использовать разложения вида $C_r C_r^+ A$ и $C_r C_r^+ AR_r^+ R_r$, однако стоимость их построения мало отличается от наилучших (по норме Фробениуса) разложений $\left(CC^+ A\right)_r$ и $\left(CC^+ AR^+ R\right)_r$, а оценки точности (в том числе и по спектральной норме) оказываются хуже.

Что касается вида оптимальных W и G для спектральной нормы, то он в общем случае неизвестен. Стоит, однако, отметить, что в случае столбцовых аппроксимаций полного ранга он совпадает с видом оптимальной аппроксимации по норме Фробениуса: $W = C^+ A$. Это легко увидеть из следующего неравенства:

$$\left\|A - CC^+ A\right\|_2 = \left\|\left(I - CC^+\right) A\right\|_2 = \left\|\left(I - CC^+\right)\left(A - CW\right)\right\|_2 \leqslant \left\|A - CW\right\|_2 \tag{1.4}$$

для произвольной матрицы W, поскольку $\left(I - CC^+\right) CW = 0$.

1.4. Свойства объема и проективного объема

Как упоминалось выше, многих оценок точности можно достичь, используя подматрицы локально максимального объема.

Определение 1.5. Объемом $\mathcal{V}(A)$ произвольной матрицы $A \in \mathbb{C}^{m \times n}$ называется произведение её сингулярных чисел

$$\mathcal{V}(A) = \prod_{i=1}^{\min(m,n)} \sigma_i(A),$$

В частности, если $m \geqslant n$, то

$$\mathcal{V}(A) = \sqrt{\det\left(AA^*\right)},$$

если $m \geqslant n$, то

$$\mathcal{V}(A) = \sqrt{\det\left(A^* A\right)},$$

и если $m = n$, то

$$\mathcal{V}(A) = |\det A|\,.$$

Проективным или r-проективным объемом матрицы A называется величина

$$\mathcal{V}_r(A) = \prod_{i=1}^{r} \sigma_i(A).$$

Название происходит из геометрической интерпретации: объем матрицы есть отношение объема образа к объему прообраза при данном линейном преобразовании.

Определение 1.6. Говорят, что подматрица полного ранга $\hat{A} \in \mathbb{C}^{m\times n}$ матрицы $A \in \mathbb{C}^{M\times N}$ обладает *ρ-локально максимальным объемом* (во всей матрице), $\rho \geqslant 1$, если перестановка её произвольной i-й строки и/или j-го столбца с любой другой строкой l и/или столбцом k матрицы A увеличивает её объем не более, чем в ρ раз.

Говорят, что матрица $\hat{A}$ обладает ρ-локально максимальным объемом в своих строках и столбцах, если перестановка её произвольной i-й строки или j-го столбца с любой другой строкой l или столбцом k матрицы A увеличивает её объем не более, чем в ρ раз (одновременные замены строки и столбца не допускаются).

Говорят, что подматрица обладает локально максимальным объемом (во всей матрице), если $\rho = 1$.

Аналогично определяется локально максимальный и ρ-локально максимальный проективный объем.

Подматрицу локально максимального объема в своих строках и столбцах также зовут *доминантной* [11].

В [39] и [59] было показано, что подматрицы ρ-локально максимального объема позволяют строить сильные RRQR и RRLU разложения. Мы вернемся к вопросу построения сильного RRLU разложения в разделе 4.4.2.

Вернемся к определению объема. Легко видеть, что он является обобщением модуля определителя матрицы. Поскольку определитель обладает многими важными свойствами, упрощающими работу с ним, полезно иметь схожие свойства для объема и проективного объема матриц. Многие из них следуют из следующего классического результата.

Теорема 1.2 (Бине-Коши)**.** *Пусть $A, B \in \mathbb{C}^{r\times N}$, $r \leqslant N$. Тогда*

$$\det AB^* = \sum_{\mathcal{I},|\mathcal{I}|=r} \det A_{:,\mathcal{I}} B^*_{:,\mathcal{I}}.$$

Каллиграфическими нижними индексами мы здесь и далее будем обозначать наборы индексов строк и столбцов. Через индекс с двоеточием : будет обозначаться набор, содержащий все строки (или столбцы) матрицы.

При $A = B$ получаем определение объема прямоугольной матрицы через объемы её квадратных подматриц.

Следствие 1.1.

$$\mathcal{V}^2(A) = \sum_{\mathcal{I},|\mathcal{I}|=r} \mathcal{V}^2\left(A_{:,\mathcal{I}}\right).$$

Одно из важных свойств объема проще всего доказать, используя так называемые *минорные* матрицы.

Определение 1.7. Минорная матрица $\mathcal{A} \in \mathbb{C}^{C_M^r \times C_N^r}$ матрицы $A \in \mathbb{C}^{M\times N}$ – это матрица с элементами, равными всем возможным минорам матрицы A размера $r \times r$.

Пусть $\mathcal{I}, |\mathcal{I}| = r$ – i-й (в лексикографическом порядке) набор из r строк матрицы A, а $\mathcal{J}, |\mathcal{J}| = r$ – j-й набор из r столбцов. Тогда

$$\mathcal{A}_{ij} = \det A_{\mathcal{I},\mathcal{J}}.$$

Следствие 1.2. Минорная матрица диагональной матрицы диагональна.

Следствие 1.3. Минорная матрица $\mathcal{I}$ единичной матрицы I является единичной матрицей.

Из теоремы Бине-Коши получаем, что произведение минорных матриц есть минорная матрица произведения.

Утверждение 1.2. *Пусть $C = AB$. Тогда*

$$\mathcal{C} = \mathcal{A}\mathcal{B}$$

Доказательство.

$$\sum_k \mathcal{A}_{ik}\mathcal{B}_{kj} = \sum_{\mathcal{K},|\mathcal{K}|=r} \det A_{\mathcal{I},\mathcal{K}} \det B_{\mathcal{K},\mathcal{J}} = \det A_{\mathcal{I},:}B_{:,\mathcal{J}} = \det C_{\mathcal{I},\mathcal{J}} = C_{ij},$$

где теорема Бине-Коши была применена для $A_{\mathcal{I},:} \in \mathbb{C}^{r\times N}$ и $B^*_{:,\mathcal{J}} \in \mathbb{C}^{r\times N}$. □

Следствие 1.4. Минорная матрица матрицы U с ортонормированными строками и/или столбцами ($UU^* = I$ и/или $U^*U = I$) также является матрицей с ортонормированными строками и/или столбцами.

Следствие 1.5. Если $A = USV$ – сингулярное разложение матрицы A, то $\mathcal{A} = \mathcal{U}\mathcal{S}\mathcal{V}$ – сингулярное разложение её минорной матрицы.

Эти свойства позволяют получить следующий важный результат.

Лемма 1.1 ([43])**.**

$$\mathcal{V}_r^2(A) \leqslant \sum_{\mathcal{I},|\mathcal{I}|=r} \sum_{\mathcal{J},|\mathcal{I}|=r} \mathcal{V}^2\left(A_{\mathcal{I},\mathcal{J}}\right) = \sum_{\mathcal{J},|\mathcal{I}|=r} \mathcal{V}^2\left(A_{:,\mathcal{J}}\right) = \sum_{i_1<\ldots<i_r} \sigma_{i_1}^2(A)\cdot\ldots\cdot\sigma_{i_r}^2(A), \qquad (1.5)$$

где в правой части стоит сумма произведений всех возможных наборов из r сингулярных чисел матрицы A с различными индексами.

Доказательство. Неравенство

$$\mathcal{V}_r^2(A) \leqslant \sum_{i_1<\ldots<i_r} \sigma_{i_1}^2(A) \cdot \ldots \cdot \sigma_{i_r}^2(A)$$

следует из определения проективного объема, поэтому достаточно доказать оставшиеся равенства.

Первое равенство в (1.5) следует из применения следствия 1.1 к каждому набору из r столбцов в $A^*_{:,\mathcal{J}} \in \mathbb{C}^{r\times N}$.

Таким образом, осталось только доказать

$$\sum_{\mathcal{I},|\mathcal{I}|=r} \sum_{\mathcal{J},|\mathcal{I}|=r} \mathcal{V}^2\left(A_{\mathcal{I},\mathcal{J}}\right) = \sum_{i_1<\ldots<i_r} \sigma_{i_1}^2(A) \cdot \ldots \cdot \sigma_{i_r}^2(A). \tag{1.6}$$

Для этого рассмотрим сингулярное разложение минорной матрицы

$$\mathcal{A} = \mathcal{U}\mathcal{S}\mathcal{V}$$

и возьмем квадрат нормы Фробениуса левой и правой части

$$\|\mathcal{A}\|_F^2 = \|\mathcal{S}\|_F^2. \tag{1.7}$$

В левой части (1.7) получим сумму квадратов модулей определителей всех $r \times r$ подматриц, что есть не что иное, как сумма квадратов объемов всех $r \times r$ подматриц, то есть левая часть (1.6).

В правой части (1.7) получим диагональную матрицу, каждый элемент которой есть произведение r различных сингулярных чисел из S (которые также есть сингулярные числа матрицы A), то есть в точности правую часть (1.6). Это доказывает последнее равенство из (1.5). □

Использование сингулярного разложения минорных матриц также позволяет легко вывести неравенство для проективного объема произведения матриц.

Утверждение 1.3. *Пусть $C = AB$. Тогда*

$$\mathcal{V}_r(C) \leqslant \mathcal{V}_r(A)\,\mathcal{V}_r(B) \tag{1.8}$$

Доказательство. Рассмотрим минорные матрицы $\mathcal{C}$, $\mathcal{A}$ и $\mathcal{B}$. Их максимальные сингулярные числа есть произведения r первых сингулярных чисел исходных матриц. Следовательно, неравенство (1.8) эквивалентно неравенству

$$\|C\|_2 \leqslant \|\mathcal{A}\|_2 \|\mathcal{B}\|_2.$$

□

Лемма 1.1 играет важную роль в построении оценок крестовых и столбцовых аппроксимаций по норме Фробениуса и норме Чебышева. Из нее также можно вывести, что подматрицы локально максимального объема образуют систему векторов-столбцов, разложение по которой других столбцов возможно с использованием небольших коэффициентов. Для этого нам сначала понадобится рассмотреть, насколько растет объем некоторой подматрицы $\hat{A} \in \mathbb{C}^{r\times n}$, $n \geqslant r$, матрицы $A \in \mathbb{C}^{r\times N}$ при добавлении одного столбца.

Рассмотрим разложение $\hat{A} = XQ$, $QQ^* = I$, $X \in \mathbb{C}^{r\times r}$, $Q \in \mathbb{C}^{r\times n}$, которое, например, может быть получено из LQ-разложения. Рассмотрим разбиение матрицы $A = \begin{bmatrix}\hat{A} & a & \tilde{A}\end{bmatrix}$, где $a \in \mathbb{C}^{r\times 1}$ – новый столбец, который мы собираемся добавить, а $\tilde{A} \in \mathbb{C}^{r\times(N-n-1)}$ – оставшаяся часть матрицы A. Запишем вид матрицы $\hat{A}^+A$:

$$\hat{A}^+A = \hat{A}^+ \begin{bmatrix}\hat{A} & a & \tilde{A}\end{bmatrix} = (XQ)^+ \begin{bmatrix}XQ & a & \tilde{A}\end{bmatrix} = Q^* \begin{bmatrix}Q & X^{-1}a & X^{-1}\tilde{A}\end{bmatrix}.$$

Обозначим

$$C = \begin{bmatrix}Q & c & \tilde{A}\end{bmatrix} = \begin{bmatrix}Q & X^{-1}a & X^{-1}\tilde{A}\end{bmatrix}. \tag{1.9}$$

Заметим, что

$$\frac{\mathcal{V}\left(\begin{bmatrix}\hat{A} & a\end{bmatrix}\right)}{\mathcal{V}\left(\hat{A}\right)} = \frac{\mathcal{V}\left(X^{-1}\begin{bmatrix}\hat{A} & a\end{bmatrix}\right)}{\mathcal{V}\left(X^{-1}\hat{A}\right)} = \mathcal{V}\left(\begin{bmatrix}Q & c\end{bmatrix}\right).$$

Посчитаем объем этой расширенной матрицы:

$$\begin{aligned}\mathcal{V}^2\left(\begin{bmatrix}Q & c\end{bmatrix}\right) &= \det\left(\begin{bmatrix}Q & c\end{bmatrix}\begin{bmatrix}Q & c\end{bmatrix}^*\right)\\ &= \det\left(QQ^* + cc^*\right)\\ &= \det\left(I + cc^*\right)\\ &= \left(1 + \|c\|_2^2\right)\\ &= \left(1 + \left\|X^{-1}a\right\|_2^2\right)\\ &= \left(1 + \left\|\hat{A}^+a\right\|_2^2\right).\end{aligned}$$

Мы только что доказали следующую лемму, версия которой для матрицы $\hat{A}^+A \in \mathbb{C}^{n\times N}$ была доказана в [42]. Её также можно было бы получить напрямую из леммы об определителе матрицы $\hat{A}\hat{A}^*$ (его изменении при одноранговом обновлении).

Лемма 1.2. *Для объема расширения невырожденной матрицы $\hat{A} \in \mathbb{C}^{r\times n}$ с помощью столбца $a \in \mathbb{C}^{r\times 1}$ справедливо равенство*

$$\mathcal{V}^2\left(\begin{bmatrix}\hat{A} & a\end{bmatrix}\right) = \mathcal{V}^2\left(\hat{A}\right)\left(1 + \left\|\hat{A}^+a\right\|_2^2\right) = \mathcal{V}^2\left(\hat{A}\right)\left(1 + \|c\|_2^2\right).$$

Теперь докажем оценку на нормы столбцов $\left\|\hat{A}^+a\right\|_2$, если подматрица $\hat{A}$ обладает локально-максимальным объемом.

Лемма 1.3. *Пусть подматрица $\hat{A} \in \mathbb{C}^{r\times n}$ обладает локально максимальным объемом в матрице $A \in \mathbb{C}^{r\times N}$. Тогда для любого столбца $a \in \mathbb{C}^{r\times 1}$ вне подматрицы $\hat{A}$ выполнено неравенство*

$$\left\|\hat{A}^{+}a\right\|_2 \leqslant \sqrt{\frac{r}{n-r+1}}. \tag{1.10}$$

Суммируя квадраты норм по всем столбцам,

$$\left\|\hat{A}^{+}A\right\|_F \leqslant \sqrt{r + \frac{r}{n-r+1}\,(N-n)} = \sqrt{r\frac{N-r+1}{n-r+1}}. \tag{1.11}$$

Следствие 1.6. При поиске подматрицы $\hat{U} \in \mathbb{C}^{r\times n}$ в ортонормированных строках $U \in \mathbb{C}^{r\times N}$, $UU^{*} = I$, получаем

$$\left\|\hat{U}^{+}\right\|_F = \left\|\hat{U}^{+}U\right\|_F \leqslant \sqrt{r + \frac{r}{n-r+1}\,(N-n)}$$

и

$$\left\|\hat{U}^{+}\right\|_2 \leqslant \sqrt{\left\|\hat{U}^{+}\right\|_F^2 - (r-1)} \leqslant \sqrt{1 + \frac{r}{n-r+1}\,(N-n)}.$$

Обобщение данной леммы на подматрицы ρ-локально максимального объема мы докажем в разделе 4.3, теорема 4.1.

Доказательство. Рассмотрим расширение $\tilde{A} = \left[\hat{A}\ a\right] \in \mathbb{C}^{r\times(n+1)}$ матрицы $\hat{A}$ столбцом a, на котором достигается максимум в (1.10).

Согласно следствию 1.1,

$$\mathcal{V}^2\left(\tilde{A}\right) = \sum_{\substack{X \in \mathbb{C}^{r\times r} \\ X \subset \tilde{A}}} \mathcal{V}^2\left(X\right).$$

Аналогично, для произвольной подматрицы $W \in \mathbb{C}^{r\times n} \subset \tilde{A}$ верно

$$\mathcal{V}^2\left(W\right) = \sum_{\substack{X \in \mathbb{C}^{r\times r} \\ X \subset W}} \mathcal{V}^2\left(X\right). \tag{1.12}$$

Суммируя (1.12) по всем $n+1$ подматрицам W, каждая подматрица X матрицы $\tilde{V}$ будет учтена $n-r+1$ раз (ровно столько подматриц размера n включают в себя все r столбцов матрицы X). Таким образом,

$$\sum_{\substack{W \in \mathbb{C}^{r\times n} \\ W \subset \tilde{A}}} \mathcal{V}^2\left(W\right) = (n-r+1) \sum_{\substack{X \in \mathbb{C}^{r\times r} \\ X \subset \tilde{A}}} \mathcal{V}^2\left(X\right) = (n-r+1)\,\mathcal{V}^2\left(\tilde{A}\right). \tag{1.13}$$

Так как $\hat{A}$ – подматрица локально максимального объема, то объем произвольной матрицы W будет не больше, а значит

$$\sum_{\substack{W \in \mathbb{C}^{r\times n} \\ W \subset \tilde{A}}} \mathcal{V}^2(A) \leqslant (n+1)\mathcal{V}^2\left(\hat{A}\right). \tag{1.14}$$

Кроме того, согласно лемме 1.2,

$$\mathcal{V}^2\left(\tilde{A}\right) = \mathcal{V}^2\left(\hat{A}\right)\left(1 + \left\|\hat{A}^+ a\right\|_2^2\right). \tag{1.15}$$

Объединяя вместе (1.13), (1.14) и (1.15), получаем

$$\mathcal{V}^2\left(\hat{A}\right)\left(1 + \left\|\hat{A}^+ a\right\|_2^2\right) = \mathcal{V}^2\left(\tilde{A}\right) = \frac{1}{n-r+1} \sum_{\substack{W \in \mathbb{C}^{r\times n} \\ W \subset \tilde{A}}} \mathcal{V}^2(W) \leqslant \frac{n+1}{n-r+1}\mathcal{V}^2\left(\hat{A}\right).$$

Разделив на $\mathcal{V}^2\left(\hat{A}\right)$ левую и правую часть, получим в итоге

$$\left\|\hat{A}^+ a\right\|_2^2 \leqslant \frac{r}{n-r+1}. \tag{1.16}$$

Суммируя оценку (1.16) по $N-n$ столбцам вне $\hat{A}$, а также по столбцам $\hat{A}$: $\left\|\hat{A}^+\hat{A}\right\|_F = \sqrt{r}$, получаем (1.11). □

Можно записать и доказательство напрямую, без использования предыдущих результатов, что также позволит оценить достижимость (1.10).

Альтернативное доказательство. Рассмотрим LQ разложение $\tilde{A} = \left[\hat{A}\ a\right] = LQ = L\left[\hat{Q}\ q\right]$. Тогда $\hat{A}^+ a = \hat{Q}^+ q$. Заметим, что

$$\left\|\hat{Q}^+ q\right\|_2^2 = \left\|\hat{Q}^+ Q\right\|_F^2 - \left\|\hat{Q}^+\hat{Q}\right\|_F^2 = \left\|\hat{Q}^+\right\|_F^2 - r. \tag{1.17}$$

При этом $\hat{Q}$ отличается от Q изменением ранга 1. Раз все сингулярные числа Q равны, то среди сингулярных чисел $\hat{Q}$ будет лишь одно не равное 1. Таким образом, максимизация объема $\hat{Q}$ (а потому и $\hat{A} = L\hat{Q}$) соответствует минимуму $\left\|\hat{Q}^+\right\|_2$, а потому и $\left\|\hat{Q}^+\right\|_F$, а потому (согласно (1.17)) и $\left\|\hat{Q}^+ q\right\|_2$.

Найдем этот минимум. Для этого распишем

$$\left\|\hat{Q}^+ q\right\|_2^2 = q^*\left(\hat{Q}^+\right)^*\hat{Q}^+ q = q^*\left(\hat{Q}\hat{Q}^*\right)^{-1} q = q^*\left(QQ^* - qq^*\right)^{-1} q = q^*\left(I - qq^*\right)^{-1} q.$$

По формуле Шермана–Моррисона получаем

$$\left\|\hat{Q}^+ q\right\|_2^2 = \frac{q^* q}{1 - q^* q}. \tag{1.18}$$

Минимум достигается на столбце минимальной длины, причем, поскольку всего столбцов $n+1$, то найдется такой, что

$$q^*q = \|q\|_2^2 \leqslant \frac{\|Q\|_F^2}{n+1} = \frac{r}{n+1}. \tag{1.19}$$

Подставляя (1.19) в (1.18), с учетом $\hat{A}^+a = \hat{Q}^+q$ получаем (1.10). □

Следствие 1.7. Неравенство (1.10) достигается, когда строки $\begin{bmatrix}\hat{A} & a\end{bmatrix}$ ортонормированы, а столбцы равны по норме. Например, подойдут r любых разных строк матрицы Фурье размера $(n+1)\times(n+1)$.

Те же свойства распространяются и на подматрицы локально максимального r-проективного объема.

Лемма 1.4. *Пусть подматрица $\hat{A} \in \mathbb{C}^{m\times n}$ обладает локально максимальным r-проективным объемом в матрице $A \in \mathbb{C}^{m\times N}$. Тогда для любого столбца $a \in \mathbb{C}^{m\times 1}$ вне подматрицы $\hat{A}$ выполнено неравенство*

$$\left\|\hat{A}_r^+ a\right\|_2 \leqslant \sqrt{\frac{r}{n-r+1}}. \tag{1.20}$$

Суммируя квадраты норм по всем столбцам,

$$\left\|\hat{A}_r^+ A\right\|_F \leqslant \sqrt{r + \frac{r}{n-r+1}(N-n)} = \sqrt{r\frac{N-r+1}{n-r+1}}. \tag{1.21}$$

Доказательство. Пусть $U \in \mathbb{C}^{m\times r}$ – матрица левых сингулярных векторов $\hat{A}$. Рассмотрим $B = U^*A$. Для любой подматрицы $\tilde{B} \in \mathbb{C}^{r\times n}$ верно неравенство

$$\mathcal{V}\left(\tilde{B}\right) = \mathcal{V}\left(U^*\tilde{A}\right) \leqslant \mathcal{V}_r\left(\tilde{A}\right),$$

где $\tilde{A}$ – соответствующая $\tilde{B}$ подматрица матрицы A. С другой стороны, проективный объем $\hat{B}$, в отличие от всех остальных подматриц уменьшится не мог:

$$\mathcal{V}\left(\tilde{B}\right) = \mathcal{V}\left(U^*\tilde{A}\right) = \mathcal{V}_r\left(\tilde{A}\right),$$

а потому $\hat{B}$ обладает локально максимальным объемом в B. Доказательство далее полностью совпадает с доказательством леммы 1.3. □

Глава 2. Существование столбцовых и крестовых аппроксимаций высокой точности

2.1. Точность по норме Чебышева

2.1.1. Верхние оценки. Аппроксимация тензоров

Для оценки погрешности по норме Чебышева применяется следующая лемма, которая выражает погрешность через отношение объемов.

Лемма 2.1. *Пусть $A \in \mathbb{C}^{M\times N}$, $C \in \mathbb{C}^{M\times n}$ её столбцы, $R \in \mathbb{C}^{m\times N}$ её строки, $\hat{A} \in \mathbb{C}^{m\times n}$ – подматрица на пересечении ранга не ниже r. Пусть $\tilde{C} \in \mathbb{C}^{M\times(n+1)}$ – произвольное расширение столбцов C еще одним столбцом A, а $\tilde{A} \in \mathbb{C}^{(m+1)\times(n+1)}$ – произвольное расширение $\hat{A}$ еще одной строкой и одним столбцом A. Тогда*

1. *Если $n = r$, то*
$$\left\|A - CC^{+}A\right\|_C \leqslant \max_{\tilde{C}} \left\|\tilde{C} - CC^{+}\tilde{C}\right\|_2 = \max_{\tilde{C}} \frac{\mathcal{V}\left(\tilde{C}\right)}{\mathcal{V}\left(C\right)}. \tag{2.1}$$
2. *Если $m = n = r$, то*
$$\left\|A - C\hat{A}^{-1}R\right\|_C = \max_{\tilde{A}} \frac{\mathcal{V}\left(\tilde{A}\right)}{\mathcal{V}\left(\hat{A}\right)}. \tag{2.2}$$
3. *Если $n = r$, то*
$$\left\|A - C\hat{A}^{+}R\right\|_C \leqslant \max_{\tilde{A}} \frac{\mathcal{V}\left(\tilde{A}\right)}{\mathcal{V}\left(\hat{A}\right)}. \tag{2.3}$$
4. *В общем случае,*
$$\left\|A - C\hat{A}_r^{+}R\right\|_C \leqslant \max_{\tilde{A}} \frac{\mathcal{V}_{r+1}\left(\tilde{A}\right)}{\mathcal{V}_r\left(\hat{A}\right)}. \tag{2.4}$$

Доказательство. Рассмотрим блочное разбиение расширения $\tilde{C} \in \mathbb{C}^{M\times(r+1)}$, содержащего столбец c, на котором погрешность по 2-норме $\left\|\tilde{C} - CC^{+}\tilde{C}\right\|_2 = \|c - CC^{+}c\|_2$ максимальна:

$$\tilde{C} = [C\ c]\,.$$

Рассмотрим QR разложение $\tilde{C}$ в блочном виде:

$$[C\ c] = [Q_1\ q]\begin{bmatrix} R_{11} & * \\ 0 & r_{22} \end{bmatrix} = QR,$$

так что $C = Q_1R_{11}$. 2-норма погрешности последнего столбца равна

$$\left\|c - CC^{+}c\right\|_2 = \|c - QQ^{*}c\|_2 = \|qr_{22}\|_2 = |r_{22}|\,. \tag{2.5}$$

С другой стороны, умножение R_{11} на Q_1 или R на Q не меняет сингулярных чисел, поэтому

$$\frac{\mathcal{V}\left(\tilde{C}\right)}{\mathcal{V}\left(C\right)} = \frac{\mathcal{V}\left(R\right)}{\mathcal{V}\left(R_{11}\right)} = \frac{|\det R|}{|\det R_{11}|} = \frac{|r_{22} \det R_{11}|}{|\det R_{11}|} = |r_{22}| \,. \tag{2.6}$$

Объединяя (2.6) с (2.5), получаем (2.1).

Теперь докажем (2.2).

Рассмотрим блочное разбиение расширения $\tilde{A} \in \mathbb{C}^{(r+1)\times(r+1)}$, содержащего элемент d, на котором достигается C-норма погрешности:

$$\tilde{A} = \begin{bmatrix} \hat{A} & c \\ b^* & d \end{bmatrix} \in \mathbb{C}^{(r+1)\times(r+1)} \tag{2.7}$$

Это случай скелетной аппроксимации, которая эквивалентна неполному LU разложению. Полное LU разложение можно представить в блочном виде как

$$L = \begin{bmatrix} L_{11} & 0 \\ * & l_{22} \end{bmatrix} \in \mathbb{C}^{(r+1)\times(r+1)}, \quad U = \begin{bmatrix} U_{11} & * \\ 0 & u_{22} \end{bmatrix} \in \mathbb{C}^{(r+1)\times(r+1)},$$

где погрешность неполного LU разложения равна $l_{22}u_{22} = d - b^*\hat{A}^{-1}c$. Так как определитель треугольной матрицы есть произведение её диагональных элементов, получаем

$$\det \tilde{A} = \det L \det U = \det L_{11} l_{22} \cdot \det U_{11} u_{22} = \det L_{11} \det U_{11} \cdot l_{11} u_{11} = \det \hat{A} \left(d - b^* \hat{A}^{-1} c\right).$$

В итоге

$$\left\|A - C\hat{A}^{-1}R\right\|_C = \left|d - b^*\hat{A}^{-1}c\right| = \frac{\left|\det \tilde{A}\right|}{\left|\det \hat{A}\right|} = \frac{\mathcal{V}\left(\tilde{A}\right)}{\mathcal{V}\left(\hat{A}\right)}.$$

Теперь докажем (2.3).

Рассмотрим блочное разбиение $\tilde{A}$:

$$\tilde{A} = \begin{bmatrix} \hat{A} & c \\ b^* & d \end{bmatrix} \in \mathbb{C}^{(m+1)\times(r+1)} \tag{2.8}$$

Погрешность в столбце c оценивается по формуле (2.1) для $\hat{C} = \hat{A}$ и $\tilde{C} = \begin{bmatrix} \hat{A} & c \end{bmatrix}$. После этого остается заметить, что $\mathcal{V}\left(\begin{bmatrix} \hat{A} & c \end{bmatrix}\right) \leqslant \mathcal{V}\left(\tilde{A}\right)$, поскольку объем есть произведение сингулярных чисел, а сингулярные числа подматрицы всегда не выше сингулярных чисел матрицы, а потому

$$\left\|c - \hat{A}\hat{A}^{+}c\right\|_C \leqslant \frac{\mathcal{V}\left(\begin{bmatrix} \hat{A} & c \end{bmatrix}\right)}{\mathcal{V}\left(\hat{A}\right)} \leqslant \frac{\mathcal{V}\left(\tilde{A}\right)}{\mathcal{V}\left(\hat{A}\right)}.$$

Погрешность в строке b^* и внутри $\hat{A}$ нулевая:

$$\begin{bmatrix} \hat{A} \\ b^* \end{bmatrix} - \begin{bmatrix} \hat{A} \\ b^* \end{bmatrix} \hat{A}^{+}\hat{A} = \begin{bmatrix} \hat{A} \\ b^* \end{bmatrix} - \begin{bmatrix} \hat{A} \\ b^* \end{bmatrix} = 0.$$

Осталось лишь оценить погрешность в элементе d. Для этого домножим $\tilde{A}$ слева на $\tilde{A}$ справа на $\begin{bmatrix} U^* & 0 \\ 0 & 1 \end{bmatrix} \in \mathbb{C}^{(r+1)\times(m+1)}$, где $U \in \mathbb{C}^{m\times r}$ – левые сингулярные векторы $\hat{A}$, соответствующие ненулевым сингулярным числам. Получим блочное разбиение

$$\begin{bmatrix} U^*\hat{A} & U^*c \\ b^* & d \end{bmatrix} \in \mathbb{C}^{(r+1)\times(r+1)},$$

эквивалентное предыдущему случаю (2.7). Так как сингулярные числа не возросли (мы умножили на матрицу с ортонормированными строками), погрешность в нем не выше $\mathcal{V}\left(\tilde{A}\right)/\mathcal{V}\left(U^*\hat{A}\right) = \mathcal{V}\left(\tilde{A}\right)/\mathcal{V}\left(\hat{A}\right)$. Однако, от такого преобразования погрешность в элементе d не изменилась, так как

$$b^*\left(U^*\hat{A}\right)^+ U^*c = b^*\left(U^*U\Sigma V\right)^+ U^*c = b^*\left(\Sigma V\right)^+ U^*c = b^*\left(U\Sigma V\right)^+ c = b^*\hat{A}^+c.$$

Таким образом, мы доказали, что погрешность во всех блоках не выше $\mathcal{V}\left(\tilde{A}\right)/\mathcal{V}\left(\hat{A}\right)$.

Осталось лишь доказать (2.4).

Рассмотрим блочное разбиение $\tilde{A}$:

$$\tilde{A} = \begin{bmatrix} \hat{A} & c \\ b^* & d \end{bmatrix}$$

Рассмотрим сингулярное разложение $\hat{A}_r = U\Sigma V$, $U \in \mathbb{C}^{m\times r}$, $\Sigma \in \mathbb{C}^{r\times r}$, $V \in \mathbb{C}^{r\times n}$. Домножим $\tilde{A}$ справа на $\begin{bmatrix} V^* & 0 \\ 0 & 1 \end{bmatrix} \in \mathbb{C}^{(n+1)\times(r+1)}$, что не увеличит её сингулярных чисел. Получим блочное разбиение

$$\widehat{A} = \begin{bmatrix} \hat{A}V^* & c \\ b^*V^* & d \end{bmatrix} \in \mathbb{C}^{(m+1)\times(r+1)}$$

эквивалентное предыдущему случаю (2.8). Так как сингулярные числа не возросли и их всего $r+1$, погрешность в нем не выше

$$\mathcal{V}\left(\widehat{A}\right)/\mathcal{V}\left(\hat{A}V^*\right) = \mathcal{V}_{r+1}\left(\tilde{A}\right)/\mathcal{V}\left(\hat{A}_rV^*\right) = \mathcal{V}_{r+1}\left(\tilde{A}\right)/\mathcal{V}_r\left(\hat{A}\right).$$

Однако, от такого преобразования погрешность в блоках c и d не изменилась: в блоке c она равна

$$\hat{A}V^*\left(\hat{A}V^*\right)^+ c = \hat{A}_rV^*\left(\hat{A}_rV^*\right)^+ c = U\Sigma VV^*\left(U\Sigma\right)^+ c = U\Sigma V\left(U\Sigma V\right)^+ c = \hat{A}_r\hat{A}_r^+c = \hat{A}\hat{A}_r^+c,$$

а в блоке d

$$b^*V^*\left(\hat{A}_rV^*\right)^+ c = b^*V^*\left(U\Sigma VV^*\right)^+ c = b^*V^*\left(U\Sigma\right)^+ c = b^*\left(U\Sigma V\right)^+ c = b^*\hat{A}_r^+c.$$

Таким образом, мы доказали, что погрешность в блоках c и d не выше $\mathcal{V}_{r+1}\left(\tilde{A}\right)/\mathcal{V}_r\left(\hat{A}\right)$. Аналогично, умножая $\tilde{A}$ слева на блок $\begin{bmatrix} U^* & 0 \\ 0 & 1 \end{bmatrix}$, получаем ту же самую оценку для блока b^*. Осталась лишь погрешность в самой $\hat{A}$, которая также не выше:

$$\left\|\hat{A} - \hat{A}\hat{A}_r^+\hat{A}\right\|_C = \left\|\hat{A} - \hat{A}_r\right\|_C \leqslant \left\|\hat{A} - \hat{A}_r\right\|_2 = \frac{\mathcal{V}_{r+1}\left(\hat{A}\right)}{\mathcal{V}_r\left(\hat{A}\right)} \leqslant \frac{\mathcal{V}_{r+1}\left(\tilde{A}\right)}{\mathcal{V}_r\left(\hat{A}\right)}.$$

□

С помощью данной леммы можно, в частности, получить оценки для подматриц локально максимального объема, полученные в [9, 10] и обобщенные в [15]. В теорему ниже мы также добавили доказательство для столбцовых аппроксимаций.

Теорема 2.1 ([15])**.** *Пусть $C \in \mathbb{C}^{M\times r}$ – подматрица локально максимального объема в матрице $A \in \mathbb{C}^{M\times N}$ ранга не ниже $r+1$. Тогда*

$$\|A - CC^+A\|_C \leqslant \max_{\tilde{C}} \sqrt{\frac{r+1}{\sum\limits_{i=1}^{n+1} \sigma_i^{-2}\left(\tilde{C}\right)}} \leqslant \sqrt{\frac{r+1}{k-r+1}} \max_{\tilde{C}} \left\|\tilde{C} - \tilde{C}_k\right\|_2 \leqslant \sqrt{\frac{r+1}{k-r+1}} \|A - A_k\|_2, \quad (2.9)$$

где $\tilde{C} \in \mathbb{C}^{M\times(r+1)}$ – произвольное расширение C одним столбцом.

Пусть $\hat{A} \in \mathbb{C}^{m\times r}$ – подматрица локально максимального объема в матрице $A \in \mathbb{C}^{M\times N}$ ранга не ниже $r+1$. Тогда

$$\|A - C\hat{A}^+R\|_C \leqslant \max_{\tilde{A}} \sqrt{\frac{(r+1)(m+1)}{(m-r+1)\sum\limits_{i=1}^{r+1} \sigma_i^{-2}(A)}} \leqslant \sqrt{\frac{(r+1)(m+1)}{(r-k+1)(m-r+1)}} \max_{\tilde{A}} \left\|\tilde{A} - \tilde{A}_k\right\|_2, \quad (2.10)$$

где $\tilde{A} \in \mathbb{C}^{(m+1)\times(r+1)}$ – произвольное расширение $\hat{A}$ одной строкой и одним столбцом.

Требование на rank A здесь чисто техническое: если rank $A = r$, то погрешность нулевая. При этом ранг не ниже r, поскольку требуется существование подматрицы локально максимального объема, которая по определению 1.6 полного ранга.

Доказательство. Рассматривая произвольное расширение одним столбцом от столбцов C до $\tilde{C} \in \mathbb{C}^{M\times(r+1)}$ и одним столбцом и строкой от $\hat{A}$ до $\tilde{A} \in \mathbb{C}^{(r+1)\times(m+1)}$, согласно лемме 2.1, получаем

$$\|A - CC^+A\|_C \leqslant \frac{\mathcal{V}\left(\tilde{C}\right)}{\mathcal{V}(C)} \quad (2.11)$$

и

$$\|A - C\hat{A}^+R\|_C \leqslant \frac{\mathcal{V}\left(\tilde{A}\right)}{\mathcal{V}\left(\hat{A}\right)}, \quad (2.12)$$

где $\tilde{C} \in \mathbb{C}^{M\times(r+1)}$ – расширение C одним столбцом, а $\tilde{A} \in \mathbb{C}^{(m+1)\times(r+1)}$ – расширение $\hat{A}$ одним столбцом и одной строкой.

Согласно лемме 1.1, сумма произведений сингулярных чисел $\tilde{C}$ выражается через сумму квадратов объемов подматриц из r столбцов. Оценивая каждый из них сверху через $\mathcal{V}(C)$, получаем

$$\mathcal{V}^2\left(\tilde{C}\right)\sum_{k=1}^{r+1}\sigma_k^{-2}(\tilde{C}) = \sum_{1\leqslant i_1<\ldots<i_r}\sigma_{i_1}^2(\tilde{C})\ldots\sigma_{i_r}^2(\tilde{C}) = \sum_{\mathcal{I},|\mathcal{I}|=r}\mathcal{V}^2\left(\tilde{C}_{:,\mathcal{I}}\right) \leqslant (r+1)\mathcal{V}^2(C) \tag{2.13}$$

Вместе с (2.11), получаем

$$\|A - CC^+A\|_C \leqslant \frac{\mathcal{V}\left(\tilde{C}\right)}{\mathcal{V}(C)} \leqslant \sqrt{\frac{r+1}{\sum\limits_{k=1}^{r+1}\sigma_k^{-2}(\tilde{C})}} \leqslant \sqrt{\frac{r+1}{\sum\limits_{k=1}^{r+1}\sigma_k^{-2}(A)}}.$$

Теперь докажем оценку для крестовой аппроксимации. Для этого заметим, что каждый минор размера $r\times r$ входит ровно в $m-r+1$ различных подматриц размера $m\times r$. Так как согласно следствию 1.1 квадрат объема каждой такой подматрицы есть сумма квадратов объемов её $r\times r$ подматриц, получаем

$$\begin{aligned}
\mathcal{V}^2\left(\tilde{A}\right)\sum_{k=1}^{r+1}\sigma_k^{-2}(\tilde{A}) &= \sum_{1\leqslant i_1<\ldots<i_r}\sigma_{i_1}^2(\tilde{A})\ldots\sigma_{i_r}^2(\tilde{A}) \\
&= \sum_{\mathcal{I},|\mathcal{I}|=r}\sum_{\mathcal{J},|\mathcal{J}|=r}\mathcal{V}^2\left(\tilde{A}_{\mathcal{I},\mathcal{J}}\right) \\
&= \frac{1}{m-r+1}\sum_{\mathcal{I},|\mathcal{I}|=m}\sum_{\mathcal{J},|\mathcal{J}|=r}\mathcal{V}^2\left(\tilde{A}_{\mathcal{I},\mathcal{J}}\right) \\
&\leqslant \frac{(m+1)(r+1)}{m-r+1}\mathcal{V}^2\left(\hat{A}\right),
\end{aligned} \tag{2.14}$$

где последнее неравенство следует из того, что всего в $\tilde{A}$ ровно $(m+1)(r+1)$ различных миноров размера $m\times r$, и каждый по объему не больше $\mathcal{V}\left(\hat{A}\right)$.

Вместе с (2.12), получаем

$$\|A - C\hat{A}^+R\|_C \leqslant \frac{\mathcal{V}\left(\tilde{A}\right)}{\mathcal{V}(A)} \leqslant \sqrt{\frac{(m+1)(r+1)}{(m-r+1)\sum\limits_{k=1}^{r+1}\sigma_k^{-2}(\tilde{A})}}.$$

□

Замечание 2.1. В случае ρ-локально максимального объема все оценки меняются не более, чем в ρ раз, поскольку максимум во столько раз изменятся оценки на объемы подматриц в (2.13) и (2.14).

Следствие 2.1. Пусть $Z, \operatorname{rank} Z = k$ – наилучшее приближение матрицы A по норме Чебышева. Тогда

$$\sum_{i=k+1}^{r+1} \sigma_i^2\left(\tilde{A}\right) = \left\|\tilde{A} - \tilde{A}_k\right\|_F^2 \leqslant \left\|\tilde{A} - \tilde{Z}\right\|_F^2 \leqslant (r+1)(m+1)\left\|\tilde{A} - \tilde{Z}\right\|_C^2 \leqslant (r+1)(m+1)\|A - Z\|_C^2 .$$

Оставив $\sum\limits_{i=k+1}^{r+1} \sigma_i^{-2}\left(\tilde{A}\right)$ вместо суммы по сингулярным числам самой матрицы A в (2.10), получаем

$$\begin{aligned}\left\|A - C\hat{A}^{+}R\right\|_C &\leqslant \frac{(r+1)(m+1)}{\sqrt{m-r+1}} \sqrt{\frac{1}{\sum\limits_{i=k+1}^{r+1} \sigma_i^{-2}\left(\tilde{A}\right) \sum\limits_{j=k+1}^{r+1} \sigma_j^2\left(\tilde{A}\right)}}\,\|A - Z\|_C \\ &\leqslant \frac{(r+1)(m+1)}{\sqrt{m-r+1}(r-k+1)}\,\|A - Z\|_C .\end{aligned}$$

Для $m = r = k$ получаем

$$\|A - C\hat{A}^{-1}R\|_C \leqslant (k+1)^2\,\|A - Z\|_C .$$

Для $m = r, r = 2k - 1$ получаем

$$\|A - C\hat{A}^{-1}R\|_C \leqslant 4k\,\|A - Z\|_C .$$

Для $m = 2r - 1, r = 3k - 1$ получаем

$$\|A - C\hat{A}^{+}R\|_C \leqslant 3\sqrt{3k-1}\,\|A - Z\|_C .$$

Теорема 2.2 ([15])**.** *Пусть $\hat{A} \in \mathbb{C}^{r\times r}$ – диагональная (главная) подматрица локально максимального объема в эрмитовой неотрицательно определенной матрице $A \in \mathbb{C}^{M\times N}$ ранга не ниже $r + 1$. Тогда*

$$\|A - C\hat{A}^{-1}C^*\|_C \leqslant \max_{\tilde{A}} \frac{r+1}{\sum\limits_{k=1}^{r+1} \sigma_k^{-1}\left(\tilde{A}\right)} \leqslant \frac{r+1}{r-k+1} \max_{\tilde{A}} \left\|\tilde{A} - \tilde{A}_k\right\|_2 , \tag{2.15}$$

где $\tilde{A} \in \mathbb{C}^{(m+1)\times(r+1)}$ – произвольное расширение $\hat{A}$ одной строкой и одним столбцом.

Доказательство. Как и в теореме 2.1, воспользуемся

$$\|A - C\hat{A}^{-1}C^*\|_C = \frac{\mathcal{V}\left(\tilde{A}\right)}{\mathcal{V}\left(\hat{A}\right)}. \tag{2.16}$$

Так как выражение $C\hat{A}^{-1}C^*$ задает неполное разложение Холецкого, максимум остатка обязан быть на диагонали (как и любой максимальный по объему минор), а потому достаточно рассмотреть диагональную (а потому положительно определенную, когда ошибка ненулевая) матрицу расширения $\tilde{A} \in \mathbb{C}^{(r+1)\times(r+1)}$.

Рассматривая произвольное разложение вида $\tilde{A} = \tilde{B}^*\tilde{B}$ (например, Холецкого), мы получим, что $\hat{A} = \hat{B}^*\hat{B}$, где $\hat{B} \in \mathbb{C}^{(r+1)\times r}$ задает какие-то r столбцов $\tilde{B}$. Тогда

$$\frac{\mathcal{V}\left(\tilde{A}\right)}{\mathcal{V}\left(\hat{A}\right)} = \frac{\mathcal{V}\left(\tilde{B}\tilde{B}^*\right)}{\mathcal{V}\left(\hat{B}\hat{B}^*\right)} = \frac{\mathcal{V}^2\left(\tilde{B}\right)}{\mathcal{V}^2\left(\hat{B}\right)} \leqslant \frac{r+1}{\sum\limits_{k=1}^{r+1}\sigma_k^{-2}(\tilde{B})}, \tag{2.17}$$

где последнее неравенство доказывается точно так же, как и столбцовый случай теоремы 2.1. Так как $\sigma_k^{-2}(\tilde{B}) = \sigma_k^{-1}(\tilde{A}) \geqslant \sigma_k^{-1}(A)$, то из (2.17) и (2.16) вытекает (2.15). □

Следствие 2.2. Если Z, $\operatorname{rank} Z = k$ есть наилучшее приближение A по норме Чебышева, то

$$\left\|A - C\hat{A}^{-1}C^*\right\|_C \leqslant \frac{(r+1)^2}{(r-k+1)^{3/2}}\|A - Z\|_C .$$

Для $r = 4k - 1$ получаем

$$\|A - C\hat{A}^{-1}C^*\|_C \leqslant 16\sqrt{k/27}\,\|A - Z\|_C .$$

Таким образом, для симметричных неотрицательно определенных матриц завышение ранга не приводит к росту оценки погрешности. С другой стороны, для произвольных матриц такой рост возможен, а потому, даже если есть возможность строить аппроксимации большего ранга, проективный объем дает лучшие гарантии для размера погрешности.

Недостатком данных оценок является тот факт, что для уменьшения коэффициента погрешности требуется использовать аппроксимацию ранга $r > k$. Тем не менее, хотя (2.15) является оценкой ранга, большего k, она говорит о том, что для эрмитовых неотрицательно определенных матриц на $r > k$ строках и столбцах можно построить аппроксимацию, по размеру погрешности не хуже, чем дает проективный объем, который мы рассмотрим далее.

Теорема 2.3. *Пусть $\hat{A} \in \mathbb{C}^{m\times r}$ – подматрица локально максимального r-проективного объема в матрице $A \in \mathbb{C}^{M\times N}$ ранга не ниже r. Тогда*

$$\|A - C\hat{A}_r^+R\|_C \leqslant \sqrt{\frac{(m+1)(n+1)}{(m-r+1)(n-r+1)}}\max_{\tilde{A}}\left\|\tilde{A} - \tilde{A}_r\right\|_2 , \tag{2.18}$$

где $\tilde{A} \in \mathbb{C}^{(m+1)\times(r+1)}$ – произвольное расширение $\hat{A}$ одной строкой и одним столбцом.

Доказательство. Рассмотрим n столбцов $\tilde{A}_{:,\mathcal{I}}$ с индексами из $\mathcal{I}$, $|\mathcal{I}| = n$, обладающих максимальным r-проективным объемом в $\tilde{A}$.

Рассмотрим матрицу $Z = \tilde{A}_r$ с сингулярным разложением $Z = U\Sigma V$, $U \in \mathbb{C}^{(m+1)\times r}$, $\Sigma \in \mathbb{C}^{r\times r}$, $V \in \mathbb{C}^{r\times(n+1)}$.

Заметим, что для произвольных столбцов $\mathcal{J}$, $|\mathcal{J}| = n$ верно неравенство

$$\mathcal{V}_r\left(Z_{:,\mathcal{J}}\right) = \mathcal{V}_r\left(UU^*\tilde{A}_{:,\mathcal{J}}\right) \leqslant \mathcal{V}_r\left(\tilde{A}_{:,\mathcal{J}}\right) \leqslant \mathcal{V}_r\left(\tilde{A}_{:,\mathcal{I}}\right). \tag{2.19}$$

С другой стороны, используя лемму 1.1, с учетом того, что все сингулярные числа Z после r-го нулевые,

$$\mathcal{V}_r^2\left(\tilde{A}\right) = \mathcal{V}_r^2\left(Z\right) = \sum_{\mathcal{K},|\mathcal{K}|=r} \mathcal{V}^2\left(Z_{:,\mathcal{K}}\right). \tag{2.20}$$

Каждая подматрица $Z_{:,\mathcal{K}}$ входит в $n-r+1$ различных подматриц $Z_{:,\mathcal{J}} \in \mathbb{C}^{(m+1)\times n}$. Поэтому

$$\sum_{\mathcal{K},|\mathcal{K}|=r} \mathcal{V}^2\left(Z_{:,\mathcal{K}}\right) = \frac{1}{n-r+1} \sum_{\mathcal{J},|\mathcal{J}|=n} \mathcal{V}_r^2\left(Z_{:,\mathcal{J}}\right) \leqslant \frac{n+1}{n-r+1} \mathcal{V}_r^2\left(\tilde{A}_{:,\mathcal{I}}\right), \tag{2.21}$$

где мы воспользовались тем, что в сумме всего $n+1$ подматрица, и, согласно (2.19), объем каждой не больше, чем у $\tilde{A}_{:,\mathcal{I}}$. Объединяя (2.20) и (2.21) получаем

$$\mathcal{V}_r^2\left(\tilde{A}\right) \leqslant \frac{n+1}{n-r+1} \mathcal{V}_r^2\left(\tilde{A}_{:,\mathcal{I}}\right)$$

Теперь заметим, что так как $\hat{A}$ – подматрица максимального проективного объема, то её проективный объем не меньше максимального проективного объема среди $m \times n$ подматриц подматрицы $\tilde{A}_{:,\mathcal{I}}$. Используя те же рассуждения для $\Phi = \left(\tilde{A}_{:,\mathcal{I}}\right)_r$ вместо Z, получаем неравенство

$$\mathcal{V}_r^2\left(\tilde{A}_{:,\mathcal{I}}\right) = \mathcal{V}_r^2\left(\Phi\right) = \sum_{\mathcal{K},|\mathcal{K}|=r} \mathcal{V}^2\left(\Phi_{\mathcal{K},:}\right) = \frac{1}{m-r+1} \sum_{\mathcal{J},|\mathcal{J}|=m} \mathcal{V}_r^2\left(\Phi_{\mathcal{J},:}\right) \leqslant \frac{m+1}{m-r+1} \mathcal{V}_r^2\left(\hat{A}\right).$$

Используя лемму 2.1, пункт 4, получаем

$$\begin{aligned}
\left\|A - C\hat{A}_r^+ R\right\|_C &\leqslant \frac{\mathcal{V}_{r+1}\left(\tilde{A}\right)}{\mathcal{V}_r\left(\hat{A}\right)} \\
&= \frac{\mathcal{V}_r\left(\tilde{A}\right)}{\mathcal{V}_r\left(\hat{A}\right)} \sigma_{r+1}\left(\tilde{A}\right) \\
&= \frac{\mathcal{V}_r\left(\tilde{A}\right)}{\mathcal{V}_r\left(\hat{A}\right)} \left\|\tilde{A} - \tilde{A}_r\right\|_2 \\
&\leqslant \sqrt{\frac{n+1}{n-r+1}} \frac{\mathcal{V}_r\left(\tilde{A}_{:,\mathcal{I}}\right)}{\mathcal{V}_r\left(\hat{A}\right)} \left\|\tilde{A} - \tilde{A}_r\right\|_2 \\
&\leqslant \sqrt{\frac{(m+1)(n+1)}{(m-r+1)(n-r+1)}} \left\|\tilde{A} - \tilde{A}_r\right\|_2.
\end{aligned}$$

□

Следствие 2.3. Если Z, $\operatorname{rank} Z = r$ есть наилучшее приближение A по норме Чебышева, то

$$\left\|A - C\hat{A}_r^+ R\right\|_C \leqslant \frac{(m+1)(n+1)}{\sqrt{(m-k+1)\,(n-k+1)}} \left\|A - Z\right\|_C.$$

Для $m = n = 2r-1$ получаем

$$\left\|A - C\hat{A}_r^+ R\right\|_C \leqslant 4r \left\|A - Z\right\|_C.$$

Можно получить оценку еще лучше, если напрямую использовать наилучшую аппроксимацию для построения ядра G. Однако, из-за использования наилучшей аппроксимации, такой результат уже не будет иметь той же практической ценности, что и предыдущие, для которых, как будет показано в разделе 4, существуют эффективные методы поиска подматриц локально максимального объема и проективного объема. Доказательство следующей теоремы основано на тех же рассуждениях, что были применены в [8] к спектральной норме. Для него нам также понадобится определение τ-псевдообращения.

Определение 2.1. Матрица вида

$$B_\tau^+ = U\Sigma_\tau^+ V, \quad \sigma_k\left(B_\tau^+\right) = \begin{cases} 1/\sigma_k(B), & \sigma_k(B) > \tau, \\ 0, & \sigma_k(B) \leqslant \tau, \end{cases}$$

называется τ-псевдообратной матрицей матрицы $B = U\Sigma V$.

Аналогично определяется матрица $B_\tau = U\Sigma_\tau V$, для которой

$$\sigma_k(B_\tau) = \begin{cases} \sigma_k(B), & \sigma_k(B) > \tau, \\ 0, & \sigma_k(B) \leqslant \tau, \end{cases}$$

Теорема 2.4. *Для произвольной матрицы $A \in \mathbb{C}^{M\times N}$ и любых m, n и $r \leqslant \min(m,n)$ существует крестовое приближение, основанное на m строках $R \in \mathbb{C}^{m\times N}$ и n столбцах $C \in \mathbb{C}^{M\times n}$ и некотором ядре $G \in \mathbb{C}^{n\times m}$ ранга не выше r, для которого*

$$\|A - CGR\|_C \leqslant \left(1 + \left(\sqrt[4]{\frac{mr}{n-r+1}} + \sqrt[4]{\frac{nr}{m-r+1}}\right)^2\right) \min_{Z,\operatorname{rank} Z=r} \|A - Z\|_C . \tag{2.22}$$

Доказательство. Пусть наилучшая аппроксимация матрицы A по норме Чебышева ранга r достигается на матрице Z. Рассмотрим сингулярное разложение

$$Z = U\Sigma V, \quad U \in \mathbb{C}^{M\times r}, \quad \Sigma \in \mathbb{C}^{r\times r}, \quad V \in \mathbb{C}^{r\times N}.$$

Пусть $F = A - Z$, столбцы $F_C \in \mathbb{C}^{M\times n}$ матрицы F соответствуют столбцам C, а строки $F_R \in \mathbb{C}^{m\times N}$ – строкам R. Пусть в U и V соответствующим строкам и столбцам соответствуют подматрицы $\hat{U} \in \mathbb{C}^{m\times r}$ и $\hat{V} \in \mathbb{C}^{r\times n}$. Тогда

$$CGR = \left(U\Sigma\hat{V} + F_C\right) G \left(\hat{U}\Sigma V + F_R\right) = U\Sigma\hat{V}G\hat{U}\Sigma V + E,$$

где

$$\begin{aligned} E &= \left(U\Sigma\hat{V} + F_C\right) GF_R + F_C G\left(\hat{U}\Sigma V + F_R\right) - F_C G F_R \\ &= U\hat{U}^+\hat{U}\Sigma\hat{V}GF_R + F_C G\hat{U}\Sigma\hat{V}\hat{V}^+V + F_C G F_R \\ &= U\hat{U}^+\hat{Z}GF_R + F_C G\hat{Z}\hat{V}^+V + F_C G F_R, \end{aligned}$$

где мы ввели подматрицу $\hat{Z} = \hat{U}\Sigma\hat{V}$.

С ее помощью мы также можем записать

$$Z = U\Sigma V = U\hat{U}^+\hat{U}\Sigma\hat{V}\hat{V}^+V = U\hat{U}^+\hat{Z}\hat{V}^+V$$

и

$$CGR = U\Sigma\hat{V}G\hat{U}\Sigma V + E = U\hat{U}^+\hat{Z}G\hat{Z}\hat{V}^+V + E.$$

Итого

$$\begin{aligned} A - CGR &= F + Z - U\hat{U}^+\hat{Z}G\hat{Z}\hat{V}^+V - E \\ &= F + U\hat{U}^+\left(\hat{Z} - \hat{Z}G\hat{Z}\right)\hat{V}^+V - U\hat{U}^+\hat{Z}GF_R - F_CG\hat{Z}\hat{V}^+V - F_CGF_R. \end{aligned} \tag{2.23}$$

Выберем $G = \hat{Z}_\tau^+$. Тогда

$$\begin{aligned} \left\|\hat{Z} - \hat{Z}G\hat{Z}\right\|_2 &\leqslant \tau, \\ \left\|\hat{Z}G\right\|_2 &\leqslant 1, \\ \left\|G\hat{Z}\right\|_2 &\leqslant 1, \\ \|G\|_2 &\leqslant \tau^{-1}. \end{aligned}$$

Обозначим через $\varepsilon = \|F\|_C = \|A - Z\|_C$ наименьшую погрешность аппроксимации ранга r. Выберем строки и столбцы, соответствующие подматрицам $\hat{U}$ и $\hat{V}$ локально максимального объема, так что, согласно лемме 1.3,

$$\max_i \left|U_{i,:}\hat{U}^+\right|_2 \leqslant \sqrt{\frac{r}{m-r+1}}, \quad \max_j \left|\hat{V}^+V_{:,j}\right|_2 \leqslant \sqrt{\frac{r}{n-r+1}}.$$

Оценивая остальные слагаемые в правой части (2.23), получаем

$$\begin{aligned} \left\|U\hat{U}^+\left(\hat{Z} - \hat{Z}G\hat{Z}\right)\hat{V}^+V\right\|_C &= \max_{i,j}\left|U_{:,i}\hat{U}^+\left(\hat{Z} - \hat{Z}G\hat{Z}\right)\hat{V}^+V_{:,j}\right| \\ &\leqslant \sqrt{\frac{r}{m-r+1}}\left\|\hat{Z} - \hat{Z}G\hat{Z}\right\|_2\sqrt{\frac{r}{n-r+1}} \\ &\leqslant \sqrt{\frac{r}{m-r+1}}\cdot\sqrt{\frac{r}{n-r+1}}\cdot\tau, \\ \left\|U\hat{U}^+\hat{Z}GF_R\right\|_C &\leqslant \sqrt{\frac{r}{m-r+1}}\left\|\hat{Z}G\right\|_2\max_j\left\|(F_R)_{:,j}\right\|_2 \\ &\leqslant \sqrt{\frac{rn}{m-r+1}}\varepsilon, \\ \left\|F_CG\hat{Z}\hat{V}^+V\right\|_C &\leqslant \sqrt{\frac{rm}{n-r+1}}\varepsilon, \\ \|F_CGF_R\|_C &\leqslant \varepsilon^2\sqrt{mn}/\tau. \end{aligned}$$

Суммируя их все, получаем

$$\|A - CGR\|_C \leqslant \varepsilon + \sqrt{\frac{r}{m-r+1}}\cdot\sqrt{\frac{r}{n-r+1}}\cdot\tau + \sqrt{\frac{rn}{m-r+1}}\varepsilon + \sqrt{\frac{rm}{n-r+1}}\varepsilon + \sqrt{mn}\frac{\varepsilon^2}{\tau}.$$

При $\tau = \sqrt[4]{\frac{mn(m-r+1)(n-r+1)}{r^2}}$ получаем (2.22). □

Используя TT-cross (крестовый тензорный поезд) [37], можно доказать существование аппроксимаций высокой точности по норме Чебышева для тензоров. Для краткости ограничимся случаем проективного объема.

Определение 2.2. Пусть $A \in \mathbb{C}^{N_1 \times \ldots \times N_d}$ – d-мерный тензор с элементами $A\,(i_1, \ldots, i_d)$.

Крестовым тензорным поездом называется тензор вида

$$\tilde{A} = A\left(J^{\leqslant 0}, i_1, J^{\geqslant 2}\right) \prod_{k=1}^{d-1} \left(\left[A\left(J^{\leqslant k}, J^{\geqslant k+1}\right) \right]^{+}_{\tau_k} A\left(J^{\leqslant k}, i_{k+1}, J^{\geqslant k+2}\right) \right),$$

где $A\left(J^{\leqslant k}, J^{\geqslant k+1}\right)$ – некоторая подматрица матрицы развертки $A\,(i_1 \ldots i_k, i_{k+1} \ldots i_d)$, заданная мультииндексами $J^{\leqslant k}$ и $J^{\geqslant k+1}$, выбирающими её строки и столбцы из мультииндексов $i_1 \ldots i_k$ и $i_{k+1} \ldots i_d$ соответственно, а $A\left(J^{\leqslant k+1}, i_{k+1}, J^{\geqslant k+2}\right)$ – трехмерные подтензоры, заданные мультииндексами $J^{\leqslant k}$, i_k и $J^{\geqslant k+2}$ соответственно. При этом первое и последнее множество мультииндексов содержат единственный элемент $J^{\leqslant 0} = J^{\geqslant d+1} = \{1\}$, так что первый и последний трехмерные тензоры являются матрицами.

Заметим, что здесь ядро G всегда имеет вид τ-псевдообратной к некоторой подматрице одной из матриц развертки. При умножении суммирование идет по общим индексам, заданным множествами J.

При $d = 2$ определение 2.2 задает крестовое разложение вида

$$A\left(i_1, J^2\right) \left[A\left(J^1, J^2\right) \right]^{+}_{\tau} A\left(J^1, i_2\right) = C \hat{A}^{+}_{\tau} R$$

для некоторой подматрицы $\hat{A}$, заданной индексами строк J^1 и столбцов J^2.

Теорема 2.5. *Пусть Z – произвольное приближение ранга r тензора $A \in \mathbb{C}^{N_1 \ldots N_d}$. Тогда существует крестовый тензорный поезд $\tilde{A}$ такой, что размеры всех подматриц в нем не больше n, а их ранг (после τ-псевдообращения) не больше r, и такой, что*

$$\left\| A - \tilde{A} \right\|_C \leqslant \frac{\left(4\sqrt{\frac{rn}{n-r+1}}\right)^{\lceil \log_2 d \rceil} - 1}{4\sqrt{\frac{rn}{n-r+1}} - 1} \cdot \frac{(n+1)^2}{n-r+1} \left\| A - Z \right\|_C \tag{2.24}$$

Доказательство. Выберем $k = \lceil d/2 \rceil$ и рассмотрим матрицу развертки

$$B = A\left(J^{\leqslant 0} i_1 \ldots i_k, i_{k+1} \ldots i_d J^{\geqslant d+1}\right),$$

где мы сразу ввели фиктивные индексы $J^{\leqslant 0}$ и $J^{\geqslant d+1}$.

Построим её разложение вида

$$B = C\left(\hat{A}\right)^{+}_{\tau_k} R + E,$$

где E – матрица погрешности, а $\hat{A}$ соответствует подматрице $A\left(J^{\leqslant k}, J^{\geqslant k+1}\right)$ с мультииндексами строк $J^{\leqslant k}$ и столбцов $J^{\geqslant k+1}$.

Оценку будем доказывать индукцией по d. Если $d = 2$ (база индукции), то текущий тензор имеет размерность $d+2 = 4$ с учетом фиктивных индексов, а погрешность аппроксимации равна

$$\varepsilon_1 = \|E\|_C \tag{2.25}$$

(мы оценим её позже). Рассмотрим теперь общий случай $d > 2$. Тогда столбцы C и строки R можно представить в виде тензоров

$$C\left(J^{\leqslant 0}, i_1, \ldots, i_k, J^{\geqslant k+1}\right)$$

и

$$R\left(J^{\leqslant k}, i_{k+1}, \ldots, i_d, J^{\geqslant k+1}\right)$$

размерности не выше $k + 2 = \lceil d/2 \rceil + 2$ и $(d-k) + 2 = \lceil d/2 \rceil \leqslant \lceil d/2 \rceil + 2$. Таким образом, размерность снизилась как минимум до $2^{\lceil \log_2 d \rceil - 1} + 2$. Применяя крестовую аппроксимацию рекурсивно к C и R, мы снизим размерность до $2 + 2 = 4$ (что соответствует базе индукции) за $\lceil \log_2 d \rceil - 1$ шагов. Погрешность на последнем $\lceil \log_2 d \rceil$-м шаге обозначим через $\varepsilon_{\lceil \log_2 d \rceil}$.

Пусть по предположению индукции для C и R нами получены оценки погрешности

$$\begin{aligned} C &= \tilde{C} + E_C, \quad \|E_C\|_C \leqslant \varepsilon_{\lceil \log_2 d \rceil - 1}, \\ R &= \tilde{R} + E_R, \quad \|E_R\|_C \leqslant \varepsilon_{\lceil \log_2 d \rceil - 1}, \end{aligned}$$

тогда

$$\begin{aligned} \varepsilon_{\lceil \log_2 d \rceil} &= \left\|A - \tilde{A}\right\|_C \\ &= \left\|B - \tilde{C}\left(\hat{A}\right)^+_{\tau_k} \tilde{R}\right\|_C \\ &\leqslant \left\|\tilde{C}\left(\hat{A}\right)^+_{\tau_k} \tilde{R} - C\left(\hat{A}\right)^+_{\tau_k} R\right\|_C + \|E\|_C \\ &= \left\|E_C\left(\hat{A}\right)^+_{\tau_k} E_R + E_C\left(\hat{A}\right)^+_{\tau_k} R + C\left(\hat{A}\right)^+_{\tau_k} E_R\right\|_C + \|E\|_C \\ &\leqslant \left\|E_C\left(\hat{A}\right)^+_{\tau_k} E_R\right\|_C + \left\|E_C\left(\hat{A}\right)^+_{\tau_k} R\right\|_C + \left\|C\left(\hat{A}\right)^+_{\tau_k} E_R\right\|_C + \|E\|_C . \end{aligned} \tag{2.26}$$

Выберем в качестве $\hat{A} \in \mathbb{C}^{n \times n}$ подматрицу в текущей развертке локально максимального r-проективного объема. Погрешность $\|E\|_C = \left\|B - C\left(\hat{A}\right)^+_{\tau_k} R\right\|_C$ оценим, используя оценку для r-проективного объема из следствия 2.3

$$\left\|B - C\hat{A}_r^+ R\right\|_C \leqslant \frac{(n+1)^2}{n-r+1} \|A - Z\|_C$$

и оценку на 2-нормы столбцов (и строк), которые ограничены по лемме 1.4

$$\max_i \left\|C_{i,:}\hat{A}_r^+\right\|_2 \leqslant \sqrt{\frac{r}{n-r+1}}, \quad \max_j \left\|\hat{A}_r^+ R_{:,j}\right\|_2 \leqslant \sqrt{\frac{r}{n-r+1}}.$$

Тогда при $\tau_k \geqslant \sigma_{r+1}\left(\hat{A}\right)$

$$\begin{aligned}
\|E\|_C &= \left\|B - C\left(\hat{A}\right)_{\tau_k}^+ R\right\|_C \\
&\leqslant \left\|B - C\hat{A}_r^+ R\right\|_C + \left\|C\left(\hat{A}\right)_{\tau_k}^+ R - C\hat{A}_r^+ R\right\|_C \\
&= \left\|B - C\hat{A}_r^+ R\right\|_C + \left\|C\hat{A}_r^+\left(\hat{A}\right)_{\tau_k}\hat{A}_r^+ R - C\hat{A}_r^+\hat{A}_r\hat{A}_r^+ R\right\|_C \\
&\leqslant \left\|B - C\hat{A}_r^+ R\right\|_C + \max_i \left\|C_{i,:}\hat{A}_r^+\right\|_2 \left\|\left(\hat{A}\right)_{\tau_k} - \hat{A}_r\right\|_2 \max_j \left\|\hat{A}_r^+ R_{:,j}\right\|_2 \\
&\leqslant \frac{(n+1)^2}{n-r+1}\|A - Z\|_C + \tau_k \frac{r}{n-r+1}.
\end{aligned} \tag{2.27}$$

Аналогично оценим остальные слагаемые в (2.26):

$$\begin{aligned}
\left\|E_C\left(\hat{A}\right)_{\tau_k}^+ R\right\|_C &= \left\|E_C\left(\hat{A}\right)_{\tau_k}^+ A_r A_r^+ R\right\|_C \\
&\leqslant \sqrt{n}\,\|E_C\|_C \left\|\left(\hat{A}\right)_{\tau_k}^+ A_r\right\|_2 \max_j \left\|A_r^+ R_{:,j}\right\|_2 \\
&\leqslant \sqrt{n}\varepsilon_{\lceil\log_2 d\rceil - 1} \cdot 1 \cdot \sqrt{\frac{r}{n-r+1}} \\
&= \sqrt{\frac{rn}{n-r+1}}\varepsilon_{\lceil\log_2 d\rceil - 1},
\end{aligned} \tag{2.28}$$

$$\left\|C\left(\hat{A}\right)_{\tau_k}^+ E_R\right\|_C \leqslant \sqrt{\frac{rn}{n-r+1}}\varepsilon_{\lceil\log_2 d\rceil - 1}, \tag{2.29}$$

$$\left\|E_C\left(\hat{A}\right)_{\tau_k}^+ E_R\right\|_C \leqslant \sqrt{n}\varepsilon_{\lceil\log_2 d\rceil - 1} \cdot \left\|\left(\hat{A}\right)_{\tau_k}^+\right\|_2 \cdot \sqrt{n}\varepsilon_{\lceil\log_2 d\rceil - 1} \leqslant \varepsilon^2_{\lceil\log_2 d\rceil - 1} n/\tau_k. \tag{2.30}$$

Выберем $\tau_k = \max\left(\sigma_{r+1}\left(\hat{A}\right), \sqrt{\frac{n(n-r+1)}{r}}\varepsilon_{\lceil\log_2 d\rceil - 1}\right)$ и подставим (2.27)-(2.30) в (2.26). После упрощений получим

$$\varepsilon_{\lceil\log_2 d\rceil} \leqslant 4\sqrt{\frac{rn}{n-r+1}}\varepsilon_{\lceil\log_2 d\rceil - 1}. \tag{2.31}$$

Для ε_1 (2.25) выберем $\tau = \sigma_{r+1}\left(\hat{A}\right)$ и получим согласно следствию 2.3

$$\varepsilon_1 = \|E\|_C \leqslant \frac{(n+1)^2}{n-r+1}\|A - Z\|_C.$$

Таким образом, рекуррентное соотношение (2.31) задает геометрическую прогрессию. Суммируя первые $\lceil\log_2 d\rceil$ её элементов, получаем (2.24). □

Следствие 2.4. При $n = r$ получаем

$$\left\|A - \tilde{A}\right\|_C \leqslant \frac{(4r)^{\lceil \log_2 d \rceil} - 1}{4r - 1} (r+1)^2 \|A - Z\|_C .$$

При $n = 2r - 1$ получаем

$$\left\|A - \tilde{A}\right\|_C \leqslant \frac{\left(4\sqrt{2r-1}\right)^{\lceil \log_2 d \rceil} - 1}{4\sqrt{2r-1} - 1} \cdot 4r \|A - Z\|_C .$$

Следствие 2.5. Вместо следствия 2.3 можно использовать теорему 2.4, что позволит улучшить оценку еще примерно в $\sqrt{r}$ раз. В частности, при $n = 2r - 1$ получаем

$$\left\|A - \tilde{A}\right\|_C \leqslant \frac{\left(4\sqrt{2r-1}\right)^{\lceil \log_2 d \rceil} - 1}{4\sqrt{2r-1} - 1} \left(4\sqrt{2r-1} + 1\right) \|A - Z\|_C .$$

Заметим, что доказательство использует лишь тот факт, что развертки тензора Z имеют ранг не выше r, то есть оценка справедлива не только для канонического разложения ранга r, но и для любого тензора, ранг соответствующих разверток которого не превосходит r, что является существенно более слабым условием. Таким образом, точность крестовой тензорной аппроксимации, в частности, отличается максимум на тот же коэффициент от любой аппроксимации с помощью тензорных поездов [70] с тем же максимальным рангом r. Кроме того, в доказательстве, естественно, не обязательно использовать один и тот же ранг r. В общем случае итоговую погрешность можно оценить, например, взяв максимум среди двух коэффициентов ошибки, накопленных в строках и столбцах соответственно, на каждом из $\lceil \log_2 d \rceil$ шагов. И таким образом получить произведение $\lceil \log_2 d \rceil$ различных рангов в итоговой оценке.

2.1.2. Нижние оценки. Аппроксимация единичной матрицы

Оценки сверху для столбцовой аппроксимации могут быть также использованы для доказательства нижних оценок аппроксимации единичной матрицы по норме Чебышева.

И, наоборот, верхние оценки аппроксимации единичной матрицы дают нижние оценки точности столбцовых (и, как частного случая, крестовых) аппроксимаций по норме Чебышева.

Точность аппроксимации единичной $N \times N$ матрицы в действительном случае эквивалентна поиску значения поперечника N-мерного октаэдра

$$d_r\left(l_1^N, l_\infty^N\right) = \min_{Z \in \mathbb{R}^{N \times N},\ \mathrm{rank}\ Z = r} \|I - Z\|_C .$$

Так как в данной работе мы рассматриваем (часто более общий) комплексный случай, будем далее предполагать возможность выбора комплексной матрицы Z. Везде далее в данном подразделе будем предполагать, что Z обозначает наилучшее приближение по C-норме.

Поскольку наилучшая аппроксимация для $N = r + 1$ имеет погрешность $1/(r+1)$, получаем следующее утверждение.

Утверждение 2.1.

$$\sup_{A\in\mathbb{C}^{M\times N}}\min_{\substack{C\in\mathbb{C}^{M\times r}\\ W\in\mathbb{C}^{r\times N}}}\max_{Z,\ \mathrm{rank}\,Z=r}\frac{\|A-CW\|_C}{\|A-Z\|_C}\geqslant r+1.$$

В случае использования n столбцов нам понадобятся более точные оценки. Так как на практике $n\sim r$, то имеет смысл рассматривать оценки для случая $N\sim r$, а затем выбрать $N=r+1$.

Теорема 2.6 ([71], теорема 2.20)**.** *Для любого $k=p^{2m}\leqslant r$, p – произвольное простое число, $m\in\mathbb{N}$ существует симметричная неотрицательно определенная матрица $Z\in\mathbb{R}^{N\times N}$ ранга k такая, что*

$$\|I-Z\|_C\leqslant\frac{1}{1+\frac{\sqrt{k}}{\lceil 2\log_k N\rceil-1}}.$$

Впервые аналогичная оценка была получена в [72].

При $p=2$, допустимое значение N можно увеличить в $\log_2\sqrt{r}+1$ раз путем использования ортогональных комбинаций из 1 и -1 вместо 1 в факторах Z (в комплексном случае N можно увеличить в p^m раз за счет использования Фурье соответствующего размера). То же справедливо, если степень двойки достаточно близка к p^m. Кроме того, можно построить аналогичную оценку для общего случая, когда k является квадратом произвольного числа, однако это также не дает преимущества с точки зрения асимптотики.

Выбрав $N=n+1$, получаем следующее утверждение.

Утверждение 2.2. *Пусть $k=p^{2m}$ – наибольшее целое число такое, что $k\leqslant r$, p – простое, $m\in\mathbb{N}$. Тогда*

$$\sup_{A\in\mathbb{C}^{M\times N}}\min_{\substack{C\in\mathbb{C}^{M\times r}\\ W\in\mathbb{C}^{r\times N}}}\max_{Z,\ \mathrm{rank}\,Z=r}\frac{\|A-CW\|_C}{\|A-Z\|_C}\geqslant 1+\frac{\sqrt{k}}{\lceil 2\log_k(n+1)\rceil-1}.$$

Используя те же рассуждения, что и в классической работе Шеннона [73], легко также получить следующие оценки, которые являются наилучшими известными для случая $r\sim\log^\alpha N$.

Теорема 2.7 ([73])**.** *Существует симметричная неотрицательно определенная матрица $Z_{\mathbb{R}}\in\mathbb{R}^{N\times N}$ ранга r такая, что*

$$\|I-Z_{\mathbb{R}}\|_C\leqslant\frac{1}{1+\frac{1}{\sqrt{1-e^{-\frac{2\ln N}{r}}}}},$$

а также эрмитова неотрицательно определенная матрица $Z_{\mathbb{C}}\in\mathbb{C}^{N\times N}$ ранга r такая, что

$$\|I-Z_{\mathbb{C}}\|_C\leqslant\frac{1}{1+\frac{1}{\sqrt{1-e^{-\frac{\ln N}{r}}}}}.\tag{2.32}$$

Доказательство. Запишем $Z_{\mathbb{C}}=B^*B$, $B\in\mathbb{C}^{r\times N}$. Зададим нормы всех столбцов B единичными. Фиксируем угол φ, и будем набирать столбцы как векторы на сфере так, чтобы угол между

направлениями для любой пары векторов оставался не меньше φ. Для этого достаточно каждый раз удалять из сферы соответствующий конус с углом φ, и выбирать новый столбец из оставшихся точек сферы. Так как относительная мера конуса на комплексной сфере равна $(\sin\varphi)^{2(r-1)}$, то мера остатка не меньше нуля, пока мы мне набрали как минимум $(\sin\varphi)^{-2(r-1)}$ столбцов. Другими словами,

$$N \geqslant (\sin\varphi)^{-2(r-1)}\,.$$

Эту оценку можно интерпретировать и как границу на угол φ при фиксированных N и r. На диагонали погрешность нулевая, поэтому достаточно проверить внедиагональные элементы. Для них получаем следующую точность аппроксимации

$$\|I - Z_{\mathbb{C}}\|_C = \max_{i,j} \left|B_i^* B_j\right| \leqslant \cos\varphi.$$

Таким образом,

$$\begin{aligned} N &\geqslant \left(1 - \cos^2\varphi\right)^{-(r-1)}, \\ \cos^2\varphi &\leqslant 1 - (N)^{\frac{1}{r-1}}, \\ \|I - Z_{\mathbb{C}}\|_C &\leqslant \cos\varphi \leqslant \sqrt{1 - e^{-\frac{\ln N}{r}}}. \end{aligned}$$

Чтобы получить оценку (2.32), достаточно взвесить матрицу $Z_{\mathbb{C}}$ в $\frac{\|I-Z_{\mathbb{C}}\|_C}{1+\|I-Z_{\mathbb{C}}\|_C}$. В этом случае погрешность диагональных элементов совпадет с новой оценкой погрешности внедиагональных элементов.

В действительном случае относительная мера не превосходит объема двух полусфер с радиусом $\sin\varphi$, накрывающих конус, к объему всей сферы, то есть отношение будет $\sin^{r-1}\varphi$ вместо $\sin^{2(r-1)}\varphi$, что приведет к появлению 2-ки в числителе экспоненты. □

В дальнейшем нам понадобится отдельно рассматривать оценки для действительного и комплексного поля. В связи с этим далее введем обозначение $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, которое будет говорить о том, что соответствующие утверждения верны как для действительного, так и для комплексного поля.

Теорема 2.8. *Для произвольного приближения $Z \in \mathbb{F}^{N\times N}$, $\operatorname{rank} Z = r$ единичной матрицы $I \in \mathbb{F}^{N\times N}$ верна оценка*

$$\|I - Z\|_C \geqslant \frac{1}{1 + \sup\limits_{\substack{V \in \mathbb{F}^{r\times N} \\ VV^* = I}} \min\limits_{\substack{n \in \mathbb{N} \\ r \leqslant n \leqslant N-1}} \sqrt{n} \min\limits_{\substack{\hat{V} \in \mathbb{F}^{r\times n} \\ \operatorname{rank}\hat{V} = r}} \min\limits_{\substack{v \in \mathbb{F}^{r\times 1} \\ v \subset V \\ v \not\subset \hat{V}}} \left\|\hat{V}^+ v\right\|_2}. \tag{2.33}$$

Доказательство. Рассмотрим сокращенное сингулярное разложение матрицы Z:

$$Z = U\Sigma V, \quad U \in \mathbb{F}^{N\times r}, \quad \Sigma \in \mathbb{F}^{r\times r}, \quad V \in \mathbb{F}^{r\times N}, \tag{2.34}$$

и с его помощью построим столбцовое приближение $C\hat{V}^+V$, где столбцы $C \in \mathbb{F}^{N\times n}, C \subset I$, а $\hat{V} \in \mathbb{F}^{r\times n}, \hat{V} \subset V$ соответствующая тем же столбцам подматрица в матрице V. Заметим, что разность $I - C\hat{V}^+V$ равна 1 на всех диагональных элементах вне столбцов C. Выберем произвольный такой i-й диагональный элемент. Для него получаем равенство

$$\left|\left(I - C\hat{V}^+V\right)_{ii}\right| = 1. \tag{2.35}$$

Далее, обозначим $E = I - Z$. Обозначим $C_E \subset E$ и $C_Z \subset Z$ соответствующие C столбцы в матрицах E и Z. Тогда

$$C = C_E + C_Z. \tag{2.36}$$

Кроме того, согласно (2.34) и определению $\hat{V}$,

$$C_Z = U\Sigma\hat{V}. \tag{2.37}$$

Подставим (2.36) и (2.37) в (2.35):

$$\begin{aligned}
1 &= \left|\left(I - C\hat{V}^+V\right)_{ii}\right| \\
&= \left|\left(Z + E - C_E\hat{V}^+V - C_Z\hat{V}^+V\right)_{ii}\right| \\
&= \left|\left(E - C_E\hat{V}^+V + U\Sigma V - U\Sigma\hat{V}\hat{V}^+V\right)_{ii}\right| \\
&= \left|\left(E - C_E\hat{V}^+V\right)_{ii}\right| \\
&\leqslant |E_{ii}| + \left|\left(C_E\hat{V}^+V\right)_{ii}\right| \\
&\leqslant \|E\|_C + \left|\left(C_E\hat{V}^+V\right)_{ii}\right|.
\end{aligned} \tag{2.38}$$

Последнее слагаемое есть модуль скалярного произведения i-й строки C_E и i-го столбца $\hat{V}^+V$. Так как в строке всего n элементов, каждый из которых не выше $\|E\|_C$, её норма не превосходит $\sqrt{n}\,\|E\|_C$. Кроме того, обозначим i-й столбец V через v. В итоге из (2.38) получаем

$$1 \leqslant \|E\|_C + \sqrt{n}\,\|E\|_C \left\|\hat{V}^+v\right\|_2, \tag{2.39}$$

что дает следующее ограничение на $\|E\|_C$:

$$\|E\|_C \geqslant \frac{1}{1 + \sqrt{n}\left\|\hat{V}^+v\right\|_2}.$$

Взяв минимумы по n, подматрице $\hat{V}$ и по индексу i столбца v, получим

$$\|I - P_r\|_C \geqslant \frac{1}{1 + \min\limits_{\substack{n\in\mathbb{N}\\ r\leqslant n\leqslant N-1}} \sqrt{n} \min\limits_{\substack{\hat{V}\in\mathbb{F}^{r\times n}\\ \operatorname{rank}\hat{V}=r}} \min\limits_{\substack{v\in\mathbb{F}^{r\times 1}\\ v\subset V\\ v\not\subset\hat{V}}} \left\|\hat{V}^+v\right\|_2},$$

где V – матрица правых сингулярных векторов Z. Заменив её на матрицу, для которой знаменатель максимален, получим (2.33). □

Заметим, что из леммы 1.3 следует

$$\begin{array}{lll} \sup & \min & \max \\ V \in \mathbb{C}^{r\times N} & \hat{V} \in \mathbb{C}^{r\times n} & v \in \mathbb{C}^{r\times 1} \\ VV^* = I & \operatorname{rank}\hat{V} = r & v \subset V \\ & & v \not\subset \hat{V} \end{array} \quad \left\|\hat{V}^+ v\right\|_2 \leqslant \sqrt{\frac{r}{n-r+1}}. \tag{2.40}$$

Следствие 2.6. Если таким образом аппроксимировать произвольную матрицу A, то вместо (2.39) с учетом (2.40) получим

$$\|A - CW\|_C \leqslant \left(1 + \sqrt{\frac{rn}{n-r+1}}\right) \min_{Z,\operatorname{rank} Z=r} \|A - Z\|_C .$$

При $n = r$ получаем

$$\|A - CW\|_C \leqslant (r+1) \min_{Z,\operatorname{rank} Z=r} \|A - Z\|_C ,$$

что совпадает с нижней оценкой утверждения 2.1.

Другие нижние оценки, полученные далее на основе теоремы 2.8, также порождают соответствующие верхние оценки для точности столбцовых аппроксимаций.

Из теоремы 2.8 вместе с (2.40) следует следующая оценка.

Следствие 2.7. Для произвольного приближения $Z \in \mathbb{C}^{N\times N}$, $\operatorname{rank} Z = r$ единичной матрицы $I \in \mathbb{C}^{N\times N}$ верна оценка

$$\|A - Z\|_C \geqslant \frac{1}{1 + \sqrt{r\frac{N-1}{N-r}}}. \tag{2.41}$$

Замечание 2.2. Для эрмитовых неотрицательно определенных Z в точности та же самая оценка была получена в работе [74] в контексте сферических кодов, и, похоже, остается наилучшей нижней оценкой для $N \leqslant r^2$.

Доказательство.

$$
\begin{aligned}
&\sup_{\substack{V\in\mathbb{C}^{r\times N}\\ VV^*=I}} \min_{\substack{n\in\mathbb{N}\\ r\leqslant n\leqslant N-1}} \sqrt{n} \min_{\hat{V}\in\mathbb{C}^{r\times n},\ \operatorname{rank}\hat{V}=r} \min_{\substack{v\in\mathbb{C}^{r\times 1}\\ v\subset V\\ v\not\subset \hat{V}}} \left\|\hat{V}^+ v\right\|_2 \leqslant \\
&\leqslant \min_{\substack{n\in\mathbb{N}\\ r\leqslant n\leqslant N-1}} \sqrt{n} \sup_{V\in\mathbb{C}^{r\times N},\ VV^*=I} \min_{\hat{V}\in\mathbb{C}^{r\times n},\ \operatorname{rank}\hat{V}=r} \min_{\substack{v\in\mathbb{C}^{r\times 1}\\ v\subset V\\ v\not\subset \hat{V}}} \left\|\hat{V}^+ v\right\|_2 \\
&\leqslant \min_{\substack{n\in\mathbb{N}\\ r\leqslant n\leqslant N-1}} \sqrt{n} \sup_{V\in\mathbb{C}^{r\times N},\ VV^*=I} \min_{\hat{V}\in\mathbb{C}^{r\times n},\ \operatorname{rank}\hat{V}=r} \max_{\substack{v\in\mathbb{C}^{r\times 1}\\ v\subset V\\ v\not\subset \hat{V}}} \left\|\hat{V}^+ v\right\|_2 \qquad (2.42)\\
&\leqslant \min_{\substack{n\in\mathbb{N}\\ r\leqslant n\leqslant N-1}} \sqrt{n}\cdot\sqrt{\frac{r}{n-r+1}} \\
&= \sqrt{r\frac{N-1}{N-r}},
\end{aligned}
$$

где последнее неравенство соответствует (2.40), полученному из леммы 1.3. Подстановка (2.42) в (2.33) дает оценку (2.41). □

Заметим, что данная оценка в точности достигается на следующих примерах: $r = 1$ и N произвольное или $r = N - 1$ и N произвольное или $r = 2$ и $N = 4$.

Отсюда также можно вывести, что в действительном случае $1/2 - \|I - Z\|_C = O\left(re^{-\ln N/r}\right)$, однако данный результат, хоть и превосходит по асимптотике результат ниже, нас интересовать не будет, а его вывод слишком громоздкий.

Насколько нам известно, при $r \sim N - r$ результат следствия 2.7 лучше всех известных оценок снизу, в том числе для действительных аппроксимаций. Кроме того, поскольку наилучшая известная оценка снизу для $r \ll N$ [75] основана на оценке для $r = N/2$, полученный результат позволяет улучшить коэффициент, приблизив его к реально наблюдаемой погрешности (см. рисунок 2.1) и распространив оценку также и на случай комплексных

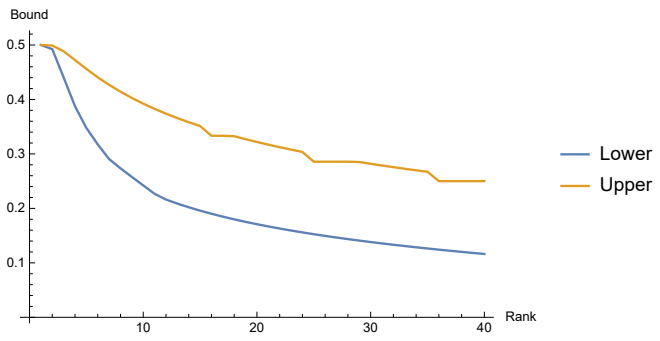


Рис. 2.1: Нижние (Lower) и верхние (Upper) оценки погрешности аппроксимации в C-норме единичной матрицы размера $N = 128$ с помощью матриц ранга r.

аппроксимаций.

Сформулируем результат для $r \ll N$ в том виде, в каком он использовался при построении рисунка 2.1. Доказательство ничем не отличается от приведенного в [75], а потому мы его опустим.

Теорема 2.9 ([75])**.** *Для произвольной аппроксимации $Z \in \mathbb{C}^{N\times N}$ ранга r единичной матрицы размера $N \times N$ верно неравенство*

$$\min_{Z,\ \operatorname{rank} Z=r} \|I - Z\|_C \geqslant \sup_{\substack{k,\\ C^k_{r+k-1}<N}} \frac{1}{1 + \left(C^k_{r+k-1}\frac{N-1}{N-C^k_{r+k-1}}\right)^{1/(2k)}}.$$

Наконец заметим, что если использовать $S\left(\hat{V}S\right)^+$ вместо $\hat{V}^+$, то получится более общая оценка

$$\|I - Z\|_C \geqslant \frac{1}{1 + \sup\limits_{\substack{V\in\mathbb{F}^{r\times N}\\ VV^*=I}} \min\limits_{\substack{\hat{V}\in\mathbb{F}^{r\times(N-1)}\\ \operatorname{rank}\hat{V}=r\\ v\subset V\\ v\not\subset\hat{V}}} \min\limits_{\substack{S\in\mathbb{F}^{(N-1)\times(N-1)}\\ \operatorname{rank}\hat{V}S\geqslant r}} \left\|S\left(\hat{V}S\right)^+ v\right\|_1}.$$

В действительном случае при $N > r(r+1)/2$ можно набрать $N-1$ столбцов, задающих минимальный охватывающий эллипсоид. При этом [76] матрица S будет диагональная, $\operatorname{tr} S = r$, $\|S\|_C \leqslant 1$, $\left\|\left(\hat{V}S\right)^+ v\right\|_2 \leqslant 1$. Подставив эти оценки, получим

$$\|I - Z\|_C \geqslant \frac{1}{1+\sqrt{r}}, \tag{2.43}$$

что дает точную оценку в действительном случае при $r = 2$ и $N = 4$. Для комплексного случая эллипсоид задается уже r^2 значениями, а потому оценку (2.43) получаем только для $N > r^2$.

Результат (2.43) обобщает некоторые оценки работы [77] на случай несимметричных/неэрмитовых аппроксимаций. В частности, в [77] показано, что для эрмитовых неотрицательно определенных приближений оценка (2.43) строго выполняется при $N > r^2 + r$ (здесь она доказана для $N > r^2$), а для симметричных неотрицательно определенных при $N > r(r+1)/2 + r/2$ (здесь она доказана для $N > r(r+1)/2$).

2.2. Точность по спектральной норме

Оценки данного и следующего разделов тесно связаны со свойствами подматриц унитарных матриц. Впервые такая связь была обнаружена в [8] в случае квадратных подматриц и затем применена к аппроксимациям на основе подматриц произвольного размера в [78].

Определение 2.3 (t-функция)**.**

$$t(r,n,N) = \sup_{\substack{U\in\mathbb{C}^{r\times N},\\ UU^*=I}} \min_{\substack{\hat{U}\in\mathbb{C}^{r\times n},\\ \operatorname{rank}\hat{U}=r}} \left\|\hat{U}^+\right\|_2 .$$

Как будет показано далее, оценки крестовых и столбцовых аппроксимаций часто можно выразить через t-функцию, где N – размер матрицы, n – размер подматрицы, а r – ранг аппроксимации.

2.2.1. Верхние оценки

Теорема 2.10. *Для произвольной матрицы $A \in \mathbb{C}^{M\times N}$, произвольных положительных $n \leqslant N$ и $r \leqslant n$ найдутся столбцы $C \in \mathbb{C}^{M\times n}$ и матрица $W \in \mathbb{C}^{n\times N}$,* $\operatorname{rank} W = r$ *такие, что*

$$\|A - CW\|_2 \leqslant t(r,n,N)\,\|A - A_r\|_2 .$$

Похожий результат, с множителем $1 + t(r,r,N)$ для r столбцов, был доказан в [79].

Доказательство. Рассмотрим сокращенное сингулярное разложение A_r:

$$A_r = U\Sigma V,\ U \in \mathbb{C}^{M\times r},\ \Sigma \in \mathbb{C}^{r\times r},\ V \in \mathbb{C}^{r\times N}.$$

Пусть $A = A_r + E = U\Sigma V + E$, тогда $C = U\Sigma\hat{V} + C_E$ для некоторой подматрицы $\hat{V} \in \mathbb{C}^{r\times n}$. Выберем $W = \hat{V}^+V$. Тогда

$$A - CW = U\Sigma V + E - U\Sigma\hat{V}\hat{V}^+V - C_E\hat{V}^+V = E - C_E\hat{V}^+V.$$

Пусть $P_C \in \mathbb{C}^{N\times n}$ – столбцы единичной матрицы, соответствующие столбцам C. Тогда $C_E = EP_C$, так что

$$\|A - CW\|_2 = \left\|E - EP_C\hat{V}^+V\right\|_2 = \left\|E\left(I - P_C\hat{V}^+V\right)\right\|_2 \leqslant \|E\|_2\left\|I - P_C\hat{V}^+V\right\|_2 .$$

$I - P_C\hat{V}^+V$ также есть некоторый проектор, а потому

$$\left\|I - P_C\hat{V}^+V\right\|_2 = \left\|P_C\hat{V}^+V\right\|_2 = \left\|\hat{V}^+\right\|_2 .$$

Выбрав подматрицу $\hat{V}$, которая минимизирует $\left\|\hat{V}^+\right\|_2$, получаем

$$\|A - CW\|_2 \leqslant \|E\|_2\left\|\hat{V}^+\right\|_2 \leqslant t(r,n,N)\,\|E\|_2 = t(r,n,N)\,\|A - A_r\|_2 .$$

□

Таким образом, достижимая точность столбцовой аппроксимации полностью определяется t-функцией. Поиск же подматриц с минимальной $\left\|\hat{U}^+\right\|_2$ является одной из задач выбора подмножества столбцов (subset selection). Соответствующие методы, дающие оценки на величину

$t(r,n,N)$, будут нами рассмотрены в главе 4. Пока что можно сразу заметить, что исходя из свойств подматриц локально максимального объема (следствие 1.6) получаем

$$t\left(r,n,N\right) \leqslant 1 + \frac{r}{n-r+1}\left(N-n\right). \tag{2.44}$$

В разделе 4.6 нами будет получена оценка

$$t\left(r,n,N\right) \leqslant \frac{\sqrt{N}}{\sqrt{n+1}-\sqrt{r}} + \sqrt{\frac{1}{N(n+1)}}. \tag{2.45}$$

Таким образом, при больших n коэффициент погрешности аппроксимации по спектральной норме не выше порядка $\sqrt{N/n}$. На практике множитель такого порядка появляется, например, в случае, когда все сингулярные числа после r-го равны, так что $\|A-A_r\|_F = \sqrt{N-r}\,\|A-A_r\|_2$.

Поскольку все оценки на спектральную норму содержат коэффициент порядка $\sqrt{N}$, в то время как для нормы Фробениуса такой коэффициент отсутствует, иногда может иметь смысл строить оценки на спектральную норму с помощью оценок на норму Фробениуса (как в следствии 1.6). В результате получим

$$t^2\left(r,n,N\right) \leqslant \sup_{\substack{U\in\mathbb{C}^{r\times N},\\ UU^*=I}} \min_{\substack{\hat{U}\in\mathbb{C}^{r\times n},\\ \operatorname{rank}\hat{U}=r}} \left\|\hat{U}^+\right\|_F^2 - r + 1,$$

а также

$$\frac{\|A-CW\|_2}{\|A-A_r\|_2} \leqslant \frac{\|A-CW\|_F}{\|A-A_r\|_F}\cdot\frac{\|A-A_r\|_F}{\|A-A_r\|_2} \leqslant \frac{\|A-CW\|_F}{\|A-A_r\|_F}\sqrt{\min\left(M,N\right)-r}.$$

Однако, как мы убедимся далее, такие оценки (при $M=N\gg n, r\gg 1$) оказываются лучше (2.45) только при $n < r+4$.

Заметим также, что в случае $N > M$ более точной оказывается оценка из [68] вида

$$\|A-CW\|_2 \leqslant \frac{\sqrt{\operatorname{rank}A-r}+\sqrt{n+1}}{\sqrt{n+1}-\sqrt{r}} \leqslant \frac{\sqrt{\min\left(M,N\right)-r}+\sqrt{n+1}}{\sqrt{n+1}-\sqrt{r}}. \tag{2.46}$$

В последующих результатах мы будем опираться на t-функцию, однако стоит учитывать, что если размеры матрицы N и M существенно отличаются ($N \gg M$), вместо нее следует подставлять правую часть (2.46).

Как уже было показано в разделе 1.3, формула (1.4), для CC^+A аппроксимации оценка по спектральной норме будет не хуже. Оценим теперь погрешность $\left(CC^+A\right)_r$ аппроксимации.

Следствие 2.8. Для произвольной матрицы $A\in\mathbb{C}^{M\times N}$, произвольных положительных $n\leqslant N$ и $r\leqslant n$ найдутся столбцы $C\in\mathbb{C}^{M\times n}$ такие, что

$$\left\|A-\left(CC^+A\right)_r\right\|_2 \leqslant \sqrt{1+t^2(r,n,N)}\,\|A-A_r\|_2$$

Доказательство. Выберем столбцы как в теореме 2.10. Тогда, используя QR разложение $C = QR$,

$$\begin{aligned}\left\|A - \left(CC^+A\right)_r\right\|_2^2 &\leqslant \left\|A - CC^+A\right\|_2^2 + \left\|CC^+A - \left(CC^+A\right)_r\right\|_2^2 \\ &= \left\|A - CC^+A\right\|_2^2 + \left\|QQ^*A - Q\left(Q^*A\right)_r\right\|_2^2 \\ &= \left\|A - CC^+A\right\|_2^2 + \left\|Q^*A - \left(Q^*A\right)_r\right\|_2^2 \\ &\leqslant \left\|A - CC^+A\right\|_2^2 + \|A - A_r\|_2^2 \\ &\leqslant \|A - CW\|_2^2 + \|A - A_r\|_2^2 \\ &\leqslant \left(1 + t^2(r,n,N)\right)\|A - A_r\|_2^2 .\end{aligned}$$

□

Перейдем теперь к крестовым аппроксимациям. Здесь проще всего получить оценки для тех видов аппроксимаций, которые оптимальны по норме Фробениуса.

Теорема 2.11. *Для произвольной матрицы $A \in \mathbb{C}^{M\times N}$, произвольных положительных $m \leqslant M$, $n \leqslant N$ и $r \leqslant \min(m,n)$ найдутся строки $R \in \mathbb{C}^{m\times N}$ и столбцы $C \in \mathbb{C}^{M\times n}$ такие, что*

$$\left\|A - CC^+AR^+R\right\|_2 \leqslant \sqrt{t^2(r,n,N) + t^2(r,m,M)}\,\|A - A_r\|_2$$

и

$$\left\|A - \left(CC^+AR^+R\right)_r\right\|_2 \leqslant \sqrt{1 + t^2(r,n,N) + t^2(r,m,M)}\,\|A - A_r\|_2 .$$

Доказательство. Выберем столбцы C в A и столбцы R^* в A^* согласно теореме 2.10. Используя вариант матричной теоремы Пифагора для спектральной нормы, получаем

$$\begin{aligned}\left\|A - CC^+AR^+R\right\|_2^2 &= \left\|A - CC^+A + CC^+A - CC^+AR^+R\right\|_2^2 \\ &\leqslant \left\|A - CC^+A\right\|_2^2 + \left\|CC^+\left(A - AR^+R\right)\right\|_2^2 \\ &\leqslant \left\|A - CC^+A\right\|_2^2 + \left\|A^* - R^*\left(R^*\right)^+A^*\right\|_2^2 \\ &\leqslant \left(t^2(r,n,N) + t^2(r,m,M)\right)\|A - A_r\|_2^2 .\end{aligned}$$

И, аналогично следствию 2.8, используя $C = Q_1R_1$ и $R = R_2^*Q_2^*$,

$$\begin{aligned}\left\|A - \left(CC^+AR^+R\right)_r\right\|_2^2 &\leqslant \left\|A - CC^+AR^+R\right\|_2^2 + \left\|CC^+AR^+R - \left(CC^+AR^+R\right)_r\right\|_2^2 \\ &= \left\|A - CC^+AR^+R\right\|_2^2 + \left\|Q_1Q_1^*AQ_2Q_2^* - Q_1\left(Q_1^*AQ_2\right)_rQ_2^*\right\|_2^2 \\ &= \left\|A - CC^+A\right\|_2^2 + \left\|Q_1^*AQ_2^* - \left(Q_1^*AQ_2\right)_r\right\|_2^2 \\ &\leqslant \left\|A - CC^+AR^+R\right\|_2^2 + \|A - A_r\|_2^2 \\ &\leqslant \left(1 + t^2(r,n,N) + t^2(r,m,M)\right)\|A - A_r\|_2^2 .\end{aligned}$$

□

Замечание 2.3. Напомним, что в случае, когда M и N не совпадают, одно из значений t-функции (соответствующее $\max(M,N)$) можно заменить на правую часть (2.46), так что коэффициент в теореме 2.11 будет зависеть только от $\min(M,N)$, но не от $\max(M,N)$.

2.2.2. Нижние оценки

Прежде всего заметим, что в плане нижних оценок по спектральной норме (и по норме Фробениуса) уже известны многие почти точные границы точности. Они записываются в виде

$$\|A - CW\|_{2,F} \geqslant C(r,n,N)\,\|A - A_r\|_{2,F}\,, \tag{2.47}$$

где $\|A - A_r\|_{2,F}$ есть погрешность сокращенного сингулярного разложения A в спектральной либо фробениусовой норме. Таким образом, $C(r,n,N)$ описывает во сколько раз погрешность столбцовой аппроксимации может быть выше погрешности наилучшей аппроксимации ранга r, заданной сокращенным сингулярным разложением.

Оценки по спектральной норме вида (2.47) были получены в [68]:

$$\left\|A - CC^{+}A\right\|_2 \geqslant \sqrt{\frac{N+\varepsilon}{n+\varepsilon}}\,\|A - A_r\|_2\,.$$

Стоит отметить, что хотя данная оценка справедлива и для аппроксимации ранга меньше n, для аппроксимаций ранга r наилучший вид аппроксимации может не совпадать с $(CC^{+}A)_r$.

Ранее нами была определена t-функция (см. определение 2.3). Как показывает теорема 2.10, данная функция тесно связана с точностью столбцовых аппроксимаций.

Чтобы построить точные нижние оценки через t-функцию, нам понадобится решить две задачи. Во-первых, необходимо ограничить погрешность столбцовой аппроксимации снизу через $t(r,n,N)$. Во-вторых, нам необходимы нижние оценки самой функции $t(r,n,N)$. Последние могут быть получены с использованием следующей леммы.

Лемма 2.2.

$$t(r,n,N) = \sup_{\substack{A\in\mathbb{R}^{r\times N},\\ \operatorname{rank} A=r}}\ \min_{\substack{\hat{A}\in\mathbb{R}^{r\times n},\\ \operatorname{rank}\hat{A}=r}} \left\|\hat{A}^{+}A\right\|_2 = \sup_{\substack{A\in\mathbb{R}^{r\times N},\\ \operatorname{rank} A=r}}\ \min_{\substack{\hat{A}\in\mathbb{R}^{r\times n},\\ \operatorname{rank}\hat{A}=r}} \frac{\left\|\hat{A}^{+}\right\|_2}{\|A^{+}\|_2}. \tag{2.48}$$

Доказательство. Рассмотрим (сокращенное) LQ разложение $A = LQ$ с $Q \in \mathbb{R}^{r\times N}$. Тогда

$$\left\|\hat{A}^{+}A\right\|_2 = \left\|\hat{Q}^{+}L^{-1}LQ\right\|_2 = \left\|\hat{Q}^{+}Q\right\|_2 = \left\|\hat{U}^{+}U\right\|_2 = \left\|\hat{U}^{+}\right\|_2.$$

Первое равенство леммы тогда доказывается взятием супремума по A и минимума по $\hat{A}$. Затем, выбрав $A = Q$ в (2.48), получаем

$$\sup_{\substack{A\in\mathbb{C}^{r\times N},\\ \operatorname{rank} A=r}}\ \min_{\substack{\hat{A}\in\mathbb{C}^{r\times n},\\ \operatorname{rank}\hat{A}=r}} \frac{\left\|\hat{A}^{+}\right\|_2}{\|A^{+}\|_2} \geqslant \sup_{\substack{Q\in\mathbb{C}^{r\times N},\\ QQ^{*}=I}}\ \min_{\substack{\hat{Q}\in\mathbb{C}^{r\times n},\\ \operatorname{rank}\hat{Q}=r}} \frac{\left\|\hat{Q}^{+}\right\|_2}{\|Q^{+}\|_2} = \sup_{\substack{Q\in\mathbb{C}^{r\times N},\\ QQ^{*}=I}}\ \min_{\substack{\hat{Q}\in\mathbb{C}^{r\times n},\\ \operatorname{rank}\hat{Q}=r}} \left\|\hat{Q}^{+}\right\|_2 = t(r,n,N). \tag{2.49}$$

С другой стороны,

$$\sup_{\substack{A\in\mathbb{C}^{r\times N},\\ \operatorname{rank} A=r}}\ \min_{\substack{\hat{A}\in\mathbb{C}^{r\times n},\\ \operatorname{rank}\hat{A}=r}} \left\|\hat{A}^{+}A\right\|_2 \geqslant \sup_{\substack{A\in\mathbb{C}^{r\times N},\\ \operatorname{rank} A=r}}\ \min_{\substack{\hat{A}\in\mathbb{C}^{r\times n},\\ \operatorname{rank}\hat{A}=r}} \left(\left\|\hat{A}^{+}\right\|_2 \sigma_r(A)\right) = \sup_{\substack{A\in\mathbb{C}^{r\times N},\\ \operatorname{rank} A=r}}\ \min_{\substack{\hat{A}\in\mathbb{C}^{r\times n},\\ \operatorname{rank}\hat{A}=r}} \frac{\left\|\hat{A}^{+}\right\|_2}{\|A^{+}\|_2}. \tag{2.50}$$

Вместе уравнения (2.49) и (2.50) доказывают второе равенство леммы. □

Далее построим примеры матрицы A, такие что $\hat{A}^+A$ легко оценить.

Утверждение 2.3.

$$t(r,n,N) \geqslant \sqrt{\frac{N-r+1}{n-r+1}}.$$

Доказательство. Рассмотрим матрицу

$$A = \begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{r\times N}.$$

Все её подматрицы $n \times r$ полного ранга равны с точностью до перестановок столбцов, так что достаточно рассмотреть $\hat{A}$, состоящую из первых n столбцов.

Тогда

$$\hat{A}^+A = \begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \vdots & \ddots & \ddots & \vdots \\ \vdots & 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & \frac{1}{n-r+1} & \cdots & \cdots & \frac{1}{n-r+1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{n-r+1} & \cdots & \cdots & \frac{1}{n-r+1} \end{bmatrix} \in \mathbb{R}^{n\times N}.$$

Спектральная норма этой матрицы есть норма $(N-r+1)\times(n-r+1)$ блока с элементами $1/(n-r+1)$, так что

$$\left\|\hat{A}^+A\right\|_2 = \sqrt{(N-r+1)(n-r+1)\cdot\frac{1}{(n-r+1)^2}} = \sqrt{\frac{N-r+1}{n-r+1}}.$$

По лемме 2.2, данный пример дает нижнюю границу $t(r,n,N)$. □

Оценка для $r = n$ может быть немного улучшена. Для этого воспользуемся следующей леммой.

Лемма 2.3.

$$t(r,r,r+1) = \sqrt{r+1}$$

Доказательство. $t(r,r,r+1) \leqslant \sqrt{r+1}$ следует из (2.44) при $n = r$ и $N = r+1$. Для получения нижней границы рассмотрим матрицу

$$A = \begin{bmatrix} 1 & 0 & \cdots & 0 & 1 \\ 0 & \ddots & 0 & \vdots & \vdots \\ \vdots & 0 & 1 & 0 & 1 \\ 0 & \cdots & 0 & 1 & 1 \end{bmatrix} \in \mathbb{R}^{r\times(r+1)}, \tag{2.51}$$

которая содержит единичную матрицу и заканчивается столбцом из единиц. Поскольку один столбец есть обновление ранга 1, то только одно из сингулярных чисел перестает быть равным единице. Следовательно,

$$\left\|\hat{A}^{-1}A\right\|_2^2 = \left\|\hat{A}^{-1}A\right\|_F^2 - (r-1) = r+1,$$

где $\hat{A} = I \in \mathbb{R}^{r\times r}$ – подматрица в первых r столбцах.

Аналогично, если один из столбцов в $\hat{A}$ заменить на последний, то $\hat{A}^{-1}A$ будет снова содержать единичную матрицу и столбец из 1 и -1. Например, если

$$\hat{A} = \begin{bmatrix} 1 & 0 & \cdots & 1 \\ 0 & \ddots & 0 & \vdots \\ \vdots & 0 & 1 & 1 \\ 0 & \cdots & 0 & 1 \end{bmatrix} \in \mathbb{R}^{r\times r},$$

то

$$\hat{A}^{-1}A = \begin{bmatrix} 1 & 0 & \cdots & -1 & 0 \\ 0 & \ddots & 0 & \vdots & \vdots \\ \vdots & 0 & 1 & -1 & 0 \\ 0 & \cdots & 0 & 1 & 1 \end{bmatrix}.$$

Таким образом, $\left\|\hat{A}^{-1}A\right\|_2^2 = r+1$ выполнено для любой подматрицы $\hat{A}$, что заканчивает доказательство. □

Последняя лемма позволяет немного улучшить оценку для $n = r$. Схожая оценка была получена в [8].

Утверждение 2.4.

$$t(r,r,N) \geqslant \sqrt{(r+1)\left\lfloor \frac{N}{r+1} \right\rfloor} = \sqrt{N - N \bmod (r+1)}.$$

Доказательство. Рассмотрим матрицу

$$B = [A\ A\ \cdots A\ 0] \in \mathbb{R}^{r\times N},$$

где A – матрица из предыдущей леммы (2.51). Возможные значения $\hat{B} \in \mathbb{R}^{r\times r}$ те же, что и возможные значения $\hat{A} \in \mathbb{R}^{r\times r}$, поскольку B состоит из копий A, и мы не можем использовать одинаковые столбцы при выборе подматрицы $\hat{B}$, поскольку тогда подматрица $\hat{B}$ будет вырожденной. Используя $BB^T = \left\lfloor \frac{N}{r+1} \right\rfloor AA^T$, получаем

$$\begin{aligned}\left\|\hat{B}^{-1}B\right\|_2^2 = \left\|\hat{A}^{-1}B\right\|_2^2 = \left\|\hat{A}^{-1}BB^T\hat{A}^{-T}\right\|_2 = \left\lfloor \frac{N}{r+1} \right\rfloor \left\|\hat{A}^{-1}AA^T\hat{A}^{-T}\right\|_2 \\ = \left\lfloor \frac{N}{r+1} \right\rfloor \left\|\hat{A}^{-1}A\right\|_2^2 = (r+1)\left\lfloor \frac{N}{r+1} \right\rfloor.\end{aligned}$$

□

Осталось доказать основную часть: как $t(r, n, N)$ связана со столбцовыми аппроксимациями матриц. Сформулируем эту связь в виде теоремы.

Теорема 2.12.

$$\sup_{A \in \mathbb{C}^{N \times N}} \min_{C,W} \frac{\|A - CW\|_2}{\|A - A_r\|_2} \geqslant t(r, n, N). \tag{2.52}$$

Доказательство. Выберем

$$A = \left[\begin{array}{c|c} U_1 & U_2 \\ \hline \varepsilon I_n & 0 \\ 0 & \varepsilon I_{N-n} \end{array}\right] \in \mathbb{C}^{(N+r) \times N}, \tag{2.53}$$

где $[U_1\ U_2] = U \in \mathbb{C}^{r \times N}$ – произвольная матрица с ортонормированными строками, $U_1 \in \mathbb{C}^{r \times n}$. Для единичных матриц нижний индекс обозначает их размер. Без ограничения общности считаем, что наилучшая столбцовая CW аппроксимация достигается с помощью первых n столбцов A, которые в данном случае равны $C = \left[\begin{array}{c} U_1 \\ \hline \varepsilon I_n \\ 0 \end{array}\right] \in \mathbb{C}^{(N+r) \times n}$.

Собственные числа $A^*A = U^*U + \varepsilon^2 I_N$ равны либо $1 + \varepsilon^2$, либо ε^2, откуда $\|A - A_r\|_2 = \varepsilon$.

Без ограничения общности можно считать, что для столбцов $W_2 \in \mathbb{C}^{r \times (N-n)}$ матрицы W, которые соответствуют $U_2 \in \mathbb{C}^{r \times (N-n)}$, справедливо разложение

$$W_2 = U_1^+ (U_2 + \varepsilon E_1) + (I - U_1^+ U_1) E_2, \quad W_2 \in \mathbb{C}^{n \times (N-n)},$$

где E_1 и E_2 – некоторые пока неизвестные матрицы. Здесь мы просто разбили W_2 на сумму двух слагаемых, где первое является линейной комбинацией столбцов U_1^*, а второе линейной комбинацией ортогональных им столбцов.

Покажем, что подматрица U_1 обязана быть полного ранга при достаточно малом ε. Для этого оценим погрешность $A - CW$ в верхнем правом блоке (2.53). Её норма равна

$$\begin{aligned} \|U_2 - U_1 W_2\|_2 &= \left\|U_2 - U_1 U_1^+ (U_2 + \varepsilon E_1) + U_1 (I_r - U_1^+ U_1) E_2\right\|_2 \\ &= \left\|U_2 - U_1 U_1^+ (U_2 + \varepsilon E_1)\right\|_2 \\ &= \left\|\left(I_r - U_1 U_1^+\right) U_2 - \varepsilon U_1 U_1^+ E_1\right\|_2 . \end{aligned} \tag{2.54}$$

Если $\operatorname{rank} U_1 < r$, то U_2 не лежит в линейной оболочке столбцов U_1^* (иначе $\operatorname{rank} U = \operatorname{rank} U_1 < r$). Поэтому $\left(I_r - U_1 U_1^+\right) U_2 \neq 0$, и, уменьшая ε, можно получить сколь угодно большую относительную погрешность, а потому такая аппроксимация при достаточно малом ε не будет оптимальной. Получаем, что $\operatorname{rank} U_1 = r$, и выражение (2.54) можно упростить:

$$\|U_2 - U_1 W_2\|_2 = \left\|\left(I_r - U_1 U_1^+\right) U_2 - \varepsilon U_1 U_1^+ E_1\right\|_2 = \|(I_r - I_r) U_2 - \varepsilon E_1\|_2 = \varepsilon \|E_1\|_2 .$$

Если $\|E_1\|_2 > t(r,n,N)$, то погрешность во всей матрице не меньше, и утверждение теоремы доказано. Поэтому далее предполагаем $\|E_1\|_2 \leqslant t(r,n,N)$, то есть E_1 ограничено величиной, не зависящей от ε.

Далее в качестве оценки снизу рассмотрим погрешность правого нижнего блока матрицы A. Тогда

$$\begin{aligned}\|A - CW\|_2^2 &\geqslant \left\| \begin{bmatrix} -\varepsilon I_n W_2 \\ \varepsilon I_{N-n} \end{bmatrix} \right\|_2^2 \\ &= \varepsilon^2 + \|\varepsilon W_2\|_2^2 \\ &\geqslant \varepsilon^2 + \varepsilon^2 \|U_1^+ (U_2 + \varepsilon E_1)\|_2^2 \\ &= \varepsilon^2 \|U_1^+\|_2^2 - O(\varepsilon^3).\end{aligned}$$

Устремляя ε к нулю, мы получим (2.52).

На данный момент мы доказали утверждение для $(N+r) \times N$ матрицы A. Определим теперь матрицу $Q \in \mathbb{C}^{(N+r)\times N}$, $Q^*Q = I_N$, задающую базис столбцов матрицы A. Тогда утверждение теоремы справедливо и для матрицы $B = Q^*A$. Действительно, проекция как самой A, так и её столбцов C на подпространство, ортогональное столбцам Q будет нулевой, а потому

$$\|A - CW\|_2 = \|Q^*A - Q^*CW\|_2 = \|B - C_B W\|_2,$$

где $C_B = Q^*C \in \mathbb{R}^{N\times n}$ – соответствующие C столбцы матрицы $B = Q^*A$. Кроме того,

$$\|A - A_r\|_2 = \|Q^* (A - A_r)\|_2 = \|B - B_r\|_2 = \varepsilon,$$

а значит и для матрицы $B \in \mathbb{C}^{N\times N}$, n её произвольных столбцов C_B и произвольной матрицы W справедливо

$$\frac{\|B - C_B W\|_2}{\|B - B_r\|_2} \geqslant \|U_1^+\|_2 - O(\varepsilon).$$

Устремив ε к нулю и взяв супремум по U, получим утверждение теоремы. □

Замечание 2.4. Теорема не требует $\operatorname{rank} W = r$.

Следствие 2.9. При $n = r$ и $A \in \mathbb{C}^{(r+1)\times(r+1)}$ получаем

$$\begin{aligned}\sup_{A\in\mathbb{C}^{N\times N}} \min_C \frac{\|A - CC^+A\|_F}{\|A - A_r\|_F} &\geqslant \sup_{A\in\mathbb{C}^{(r+1)\times(r+1)}} \min_C \frac{\|A - CC^+A\|_F}{\|A - A_r\|_F} \\ &= \sup_{A\in\mathbb{C}^{(r+1)\times(r+1)}} \min_C \frac{\|A - CC^+A\|_2}{\|A - A_r\|_2} \\ &\geqslant t(r,r,r+1) \\ &= \sqrt{r+1},\end{aligned}$$

где мы воспользовались тем, что $A - CC^+A$ – матрица ранга 1 (содержит ненулевые элементы только в одном столбце), так что её спектральная и фробениусова нормы совпадают. Таким образом, мы получили наилучшую оценку снизу по норме Фробениуса для r столбцов, аналогичную (2.80).

В качестве следствия теорем 2.10 и 2.12 получаем, что при $N \leqslant M$ нижняя и верхняя оценки в терминах $t(r, n, N)$ совпадают.

Следствие 2.10.

$$t(r, n, \min(M, N)) \leqslant \sup_{A \in \mathbb{R}^{M \times N}} \min_{C,W} \frac{\|A - CW\|_2}{\|A - A_r\|_2} \leqslant t(r, n, N).$$

В частности, при $N \leqslant M$,

$$\sup_{A \in \mathbb{R}^{M \times N}} \min_{C,W} \frac{\|A - CW\|_2}{\|A - A_r\|_2} = t(r, n, N).$$

Замечание 2.5. Нижняя оценка в теореме 2.12 использует произвольную W, тогда как верхняя оценка доказана при дополнительном условии $\operatorname{rank} W = r$. Поэтому данный результат не зависит от того, берем ли мы минимум по всем W или только по W ранга r.

2.3. Точность по норме Фробениуса

2.3.1. Верхние оценки

Впервые оптимальные оценки по норме Фробениуса для столбцовых аппроксимаций были получены в [43], где авторами был использован вероятностный подход. А именно, было доказано существование столбцов $C \in \mathbb{C}^{M \times r}$ таких, что

$$\left\|A - CC^{+}A\right\|_F \leqslant \sqrt{r+1}\,\|A - A_r\|_F\,.$$

Там же, используя геометрическую интерпретацию, было доказано существование матрицы $A \in \mathbb{R}^{(r+1)\times(r+1)}$, для которой

$$\left\|A - CC^{+}A\right\|_F \geqslant (1-\varepsilon)\sqrt{r+1}\,\|A - A_r\|_F$$

для произвольного $\varepsilon > 0$, что говорит об оптимальности полученного коэффициента.

Их идея была развита в [46] на случай $n > r$ столбцов. Здесь мы получим ту же оценку, используя более простое доказательство, аналогичное описанному в [43].

Теорема 2.13 ([46])**.** *Пусть $A \in \mathbb{C}^{M \times N}$. Тогда для любого n найдутся столбцы $C \in \mathbb{C}^{M \times n}$ такие, что для любого $r \leqslant n$ справедлива оценка*

$$\left\|A - CC^{+}A\right\|_F \leqslant \sqrt{\frac{n+1}{n-r+1}}\,\|A - A_r\|_F\,. \tag{2.55}$$

Доказательство. Будем выбирать столбцы C с вероятностью

$$P(C) = \frac{\mathcal{V}^2(C)}{\sum\limits_{\mathcal{I},|\mathcal{I}|=n} \mathcal{V}^2\left(A_{:,\mathcal{I}}\right)}. \tag{2.56}$$

Оценим квадрат нормы Фробениуса ошибки как сумму квадратов норм погрешностей в каждом столбце A. Погрешность в каждом отдельном столбце получим, рассмотрев соответствующие расширения $\tilde{C} \in \mathbb{C}^{M\times(n+1)}$ столбцов C одним новым столбцом и применив первый пункт леммы 2.1:

$$\left\|A - CC^{+}A\right\|_F^2 = \sum_{\tilde{C}\supset C, \tilde{C}\in\mathbb{C}^{M\times(n+1)}} \frac{\mathcal{V}^2\left(\tilde{C}\right)}{\mathcal{V}^2\left(C\right)}.$$

Теперь мы можем оценить матожидание погрешности, подставив вероятности (2.56):

$$\begin{aligned}
\mathbb{E}_C \left\|A - CC^{+}A\right\|_F^2 &= \sum_{\tilde{C}\supset C, \tilde{C}\in\mathbb{C}^{M\times(n+1)}} \frac{\mathcal{V}^2\left(\tilde{C}\right)}{\mathcal{V}^2\left(C\right)} P\left(C\right) \\
&= \sum_{\tilde{C}\supset C, \tilde{C}\in\mathbb{C}^{M\times(n+1)}} \frac{\mathcal{V}^2\left(\tilde{C}\right)\mathcal{V}^2\left(C\right)}{\mathcal{V}^2\left(C\right) \sum\limits_{\mathcal{I},|\mathcal{I}|=n} \mathcal{V}^2\left(A_{:,\mathcal{I}}\right)} \\
&= \frac{\sum\limits_{\tilde{C}\supset C, \tilde{C}\in\mathbb{C}^{M\times(n+1)}} \mathcal{V}^2\left(\tilde{C}\right)}{\sum\limits_{\mathcal{I},|\mathcal{I}|=n} \mathcal{V}^2\left(A_{:,\mathcal{I}}\right)}.
\end{aligned}$$

Столбцы $\tilde{C} = A_{:,\mathcal{J}}, |\mathcal{J}| = n+1$ содержат $n+1$ различных столбцов C, а потому будут посчитаны $n+1$ раз. Таким образом,

$$\mathbb{E}_C \left\|A - CC^{+}A\right\|_F^2 = (n+1)\frac{\sum\limits_{\mathcal{J},|\mathcal{J}|=n+1} \mathcal{V}^2\left(A_{:,\mathcal{J}}\right)}{\sum\limits_{\mathcal{I},|\mathcal{I}|=n} \mathcal{V}^2\left(A_{:,\mathcal{I}}\right)}. \tag{2.57}$$

Разделим (2.57) на $\|A - A_r\|_F^2 = \sum\limits_{k\geqslant r+1} \sigma_k^2\left(A\right)$ и применим лемму 1.1 для числителя и знаменателя в правой части (2.57). Получим

$$\frac{\mathbb{E}_C \left\|A - CC^{+}A\right\|_F^2}{\|A - A_r\|_F^2} = (n+1)\frac{\sum\limits_{i_1<\ldots<i_{n+1}} \sigma_{i_1}^2\left(A\right)\cdot\ldots\cdot\sigma_{i_{n+1}}^2\left(A\right)}{\sum\limits_{i_1<\ldots<i_n} \sigma_{i_1}^2\left(A\right)\cdot\ldots\cdot\sigma_{i_n}^2\left(A\right) \sum\limits_{k\geqslant r+1} \sigma_k^2\left(A\right)}. \tag{2.58}$$

Числитель и знаменатель правой части (2.58) состоят из суммы произведение сингулярных чисел A по $n+1$ штуке. Каждое произведение $\sigma_{i_1}^2\left(A\right)\cdot\ldots\cdot\sigma_{i_{n+1}}^2$ в числителе содержит $n+1$ сингулярных чисел. Заметим, что оно встретиться в знаменателе хотя бы $n-r+1$ раз: когда $k = i_{r+1}, k = i_{r+2}, \ldots, k = i_{n+1}$ ($k = i_r$ может отсутствовать, если $i_r = r$, так как $k \geqslant r+1$). Таким образом, знаменатель хотя бы в $n-r+1$ раз больше числителя, а потому

$$\frac{\mathbb{E}_C \left\|A - CC^{+}A\right\|_F^2}{\|A - A_r\|_F^2} \leqslant \frac{n+1}{n-r+1}.$$

Раз в среднем мы получили коэффициент не больше $\frac{n+1}{n-r+1}$, то значит, что он также ограничен для каких-то конкретных столбцов C, что доказывает (2.55). □

Замечание 2.6. При доказательстве можно также учесть произведения в знаменателе, содержащие n различных сингулярных чисел, что приведет к оценке

$$\left\|A - CC^{+}A\right\|_F^2 \leqslant \frac{n+1}{n-r+1} \cdot \frac{1}{1+\frac{n-r}{2(\min(M,N)-n)}} \|A - A_r\|_F^2$$

при $n < \operatorname{rank} A$.

Таким образом, коэффициент лучше, когда набранное число столбцов n близко к общему числу столбцов N. Похожий результат мы увидим и при построении нижних оценок.

Недостатком данной теоремы является то, что ранг аппроксимации оказывается выше требуемого ранга r. Чтобы получить столбцовую аппроксимацию ранга r, воспользуемся следующей теоремой.

Теорема 2.14 ([43]). *Пусть дана матрица $A \in \mathbb{C}^{M\times N}$ и k столбцов $\hat{C} \in \mathbb{C}^{M\times k}$. Тогда для любого n и r найдутся столбцы $C \in \mathbb{C}^{M\times n}$, содержащие столбцы $\hat{C}$ и такие, что*

$$\left\|A - \left(CC^{+}A\right)_r\right\|_F \leqslant \|A - A_r\|_F + \frac{r}{n-k}\left\|A - \hat{C}\hat{C}^{+}A\right\|_F.$$

Данный результат основан на доборе $n-k$ столбцов исходя из leverage scores: с вероятностью, пропорциональной квадрату их нормы в проекции на k исходных столбцов. Тогда оценка теоремы 2.14 достигается при усреднении по такому вероятностному распределению.

Заметим, что leverage scores полезны лишь в теории как раз благодаря оценкам на усреднение. На практике же обычно всегда выгодней взять самый длинный столбец как в QR разложении с выбором ведущих столбцов, чем специально (локально) ухудшать выбор, так что самый длинный столбец выбирается лишь с вероятностью, пропорциональной его длине.

Даже преимущество в виде быстрого набора столбцов (столбцы через leverage scores набираются одновременно) компенсируется тем фактом, что и при жадном наборе всегда можно решить одновременно набрать несколько столбцов, хотя это, безусловно, обычно приводит к меньшей точности аппроксимации, чем при последовательном наборе. Сложность такого набора после вычисления длин будет в среднем $O\left(N + n\log n\right)$ при использовании быстрой сортировки. То же можно сказать об использовании leverage scores в задаче поиска сильно невырожденных подматриц, третий источник в таблице 4.1.

Вместе с теоремой 2.13, она позволяет построить столбцовую аппроксимацию ранга r с n столбцами, близкую к оптимальной.

Утверждение 2.5. *Пусть $A \in \mathbb{C}^{M\times N}$. Тогда для любых n и r найдутся столбцы $C \in \mathbb{C}^{M\times n}$ такие, что*

$$\left\|A - \left(CC^{+}A\right)_r\right\|_F \leqslant \min\left(\sqrt{2+\frac{r}{n-r+1}}, \sqrt{1+\frac{r}{n+r+1-\sqrt{1+4r(n+1)}}}\right)\|A - A_r\|_F. \quad (2.59)$$

Доказательство. Рассмотрим сначала случай, когда r и n близки. Тогда близкую к оптимальной оценку можно получить следующим образом, используя QR разложение $C = QR$:

$$\begin{aligned}
\left\|A - \left(CC^+A\right)_r\right\|_F^2 &\leqslant \left\|A - CC^+A\right\|_F^2 + \left\|CC^+A - \left(CC^+A\right)_r\right\|_F^2 \\
&= \left\|A - CC^+A\right\|_F^2 + \left\|QQ^*A - Q\left(Q^*A\right)_r\right\|_F^2 \\
&= \left\|A - CC^+A\right\|_F^2 + \left\|Q^*A - \left(Q^*A\right)_r\right\|_F^2 \\
&\leqslant \left\|A - CC^+A\right\|_F^2 + \|A - A_r\|_F^2 \\
&\leqslant \left(1 + \frac{n+1}{n-r+1}\right)\|A - A_r\|_F^2 = \left(2 + \frac{r}{n-r+1}\right)\|A - A_r\|_F^2 .
\end{aligned}$$

К сожалению, такой подход не позволяет достичь коэффициента, сколь угодно близкого к единице. Чтобы получить близкий к единице коэффициент погрешности, воспользуемся теоремой 2.13 для $k-1$ столбца, а затем применим теорему 2.14, чтобы набрать оставшиеся $n-k-1$ столбцов. Тогда мы сможем гарантировать точность аппроксимации

$$\begin{aligned}
\left\|A - \left(CC^+A\right)_r\right\|_F &\leqslant \sqrt{1 + \frac{r}{n-k+1}\cdot\frac{k}{k-r}}\,\|A - A_r\|_F \\
&= \sqrt{1 + \frac{r}{n+1-k}\cdot\frac{1}{1-r/k}}\,\|A - A_r\|_F \\
&= \sqrt{1 + \frac{r}{n+1-k-r(n+1)/k+r}}\,\|A - A_r\|_F \\
&= \sqrt{1 + \frac{r}{n+r+1-k-r(n+1)/k}}\,\|A - A_r\|_F .
\end{aligned} \tag{2.60}$$

Выберем целое k, минимизирующее правую часть (2.60). Это будет $k = \left\lceil\sqrt{(n+1)\,r}\right\rceil$ или $k = \left\lfloor\sqrt{(n+1)\,r}\right\rfloor$. С учетом дополнительного требования $k - 1 \geqslant r$, такие k можно использовать при $n \geqslant r+2$. Величина

$$k + \frac{r(n+1)}{k} \tag{2.61}$$

является выпуклой функцией, а потому её максимум по x на отрезке $[x;\ x+1]$ (что не меньше максимума при целых k) достигается, когда значения на концах отрезка равны, то есть

$$x + \frac{r(n+1)}{x} = x + 1 + \frac{r(n+1)}{x+1}$$

Решив квадратное уравнение на x и подставив решение в (2.61), находим, что

$$k + \frac{r(n+1)}{k} \leqslant \sqrt{1 + 4r(n+1)}.$$

В итоге

$$\left\|A - \left(CC^+A\right)_r\right\|_F \leqslant \left(1 + \frac{r}{n+r+1-\sqrt{1+4r(n+1)}}\right)\|A - A_r\|_F .$$

При $n = r$ правая часть в минимуме в (2.59) не определена, а при $n = r+1$ левая часть всегда меньше правой, что позволяет избавиться от ограничения $n \geqslant r+2$. □

Замечание 2.7. Данного результата уже достаточно для достижения крестовых аппроксимаций с постоянным коэффициентом при числе строк и столбцов в несколько раз больше ранга r. Из теоремы 2.13 и матричной теоремы Пифагора получаем

$$\begin{aligned}\left\|A - CC^+AR^+R\right\|_F^2 &= \left\|A - CC^+A\right\|_F + \left\|CC^+\left(A - AR^+R\right)\right\|_F^2 \\ &\leqslant \left\|A - CC^+A\right\|_F + \left\|A - AR^+R\right\|_F^2 \\ &\leqslant \left(\frac{n+1}{n-r+1} + \frac{m+1}{m-r+1}\right)\|A - A_r\|_F^2 .\end{aligned}$$

Если требуется получить оценку на аппроксимацию ранга r, можно воспользоваться

$$\left\|CC^+AR^+R - \left(CC^+AR^+R\right)_r\right\|_F \leqslant 1,$$

что приведет к

$$\left\|A - \left(CC^+AR^+R\right)_r\right\|_F \leqslant \sqrt{1 + \frac{n+1}{n-r+1} + \frac{m+1}{m-r+1}}\,\|A - A_r\|_F ,$$

или воспользоваться только что доказанным утверждением вместе с аппроксимацией вида $\left(CC^+A\right)_r R^+R$ при $m \leqslant n$, что приводит к

$$\begin{aligned}\left\|A - \left(CC^+AR^+R\right)_r\right\|_F &\leqslant \left\|A - \left(CC^+A\right)_r R^+R\right\|_F \\ &\leqslant \sqrt{1 + \frac{r}{n+r+1-\sqrt{1+4r(n+1)}} + \frac{m+1}{m-r+1}}\,\|A - A_r\|_F .\end{aligned}$$

Мы можем говорить, что полученные оценки столбцовой аппроксимации близки к оптимальным, поскольку коэффициенты при больших n и $\min(M,N) \to \infty$ имеют вид $1 + r/n + o\left(1/n\right)$, что в том же пределе $N \to \infty$ совпадает с доказанной в следующем подразделе оценкой снизу.

В [80] показано, что можно построить CC^+A аппроксимацию ранга r с коэффициентом ошибки не более $\sqrt{r+1}$, который является оптимальным с точки зрения нижней границы, доказанной там же. Однако их подход требует от $O\left(rT_{SVD} + rMN^2\right)$ до $O\left(MN^3 r \log N\right)$ операций (при константе матричного умножения 3) в детерминированном случае (в зависимости от требований к точности и устойчивости) и $O\left(rM^2N\right)$ для рандомизированного алгоритма, дающего оценку с $\sqrt{r+1}$ в среднем (T_{SVD} – время сингулярного разложения $M \times N$ матрицы). Кроме того, заметим, что хотя оптимальный коэффициент погрешности достижим за полиномиальное время, поиск наилучших столбцов в целом является NP-сложной задачей [81] (для поиска оптимальных по норме Фробениуса столбцов и строк крестовой аппроксимации доказана UG-сложность [82]).

Таким образом, важно отметить, что хотя задача поиска оптимальных столбцов

$$\left\|A - CC^+A\right\|_F \to \min_C$$

является NP-сложной, задача поиска столбцов, гарантирующих достижение наилучшей оценки

$$\left\|A - CC^{+}A\right\|_F \leqslant \left(\sup_B \min_{C_B} \frac{\left\|B - C_B C_B^{+} B\right\|_F}{\|B - B_r\|_F}\right) \|A - A_r\|_F = \sqrt{r+1}\,\|A - A_r\|_F$$

для r столбцов решается за полиномиальное время.

Во всех случаях стоимость алгоритмов чрезмерна велика. Здесь мы построим метод получения оценки с коэффициентом $\sqrt{r+1}$ от наилучшей, используя только одно сингулярное разложение, а потому имеющему полную сложность $O\left(MN \min\left(M, N\right)\right)$.

Теорема 2.15. *Пусть даны матрицы* $A, Z \in \mathbb{C}^{M\times N}$, $\operatorname{rank} Z = r$. *Тогда за* $O\left(MNr\right)$ *можно найти такие строки* R *и столбцы* C *матрицы* A, *для которых одновременно*

$$\left\|A - CC^{+}A\right\|_F \leqslant \|A - CW\|_F \leqslant \sqrt{r+1}\,\|A - Z\|_F \tag{2.62}$$

и

$$\left\|A - CC^{+}A\right\|_2^2 \leqslant \|A - CW\|_2^2 \leqslant \|A - Z\|_2^2 + r\,\|A - Z\|_F^2 \leqslant \left(1 + r\left(N - r\right)\right)\|A - Z\|_2^2. \tag{2.63}$$

Доказательство. Воспользуемся строками $V \in \mathbb{C}^{r\times N}$ матрицы правых сингулярных векторов Z и ортогонализуем к ним матрицу A:

$$\tilde{A} = A - AV^{*}V.$$

Из условия задачи следует, что $\left\|\tilde{A}\right\|_F \leqslant \|A - Z\|_F$. Такая инициализация занимает $O\left(MNr\right)$ операций.

Приближение будем строить с $W = \hat{V}^{-1}V$, где $\hat{V} \in \mathbb{C}^{r\times N}$ соответствует столбцам C. Если столбцам C при этом соответствуют столбцы $\tilde{C}$ матрицы $\tilde{A}$, то

$$\begin{aligned}\|A - CW\|_{2,F} &= \left\|\tilde{A} + AV^{*}V - \tilde{C}W - AV^{*}\hat{V}W\right\|_{2,F} \\ &= \left\|\tilde{A} + AV^{*}V - \tilde{C}W - AV^{*}\hat{V}\hat{V}^{-1}V\right\|_{2,F} \\ &= \left\|\tilde{A} - \tilde{C}W\right\|_{2,F}.\end{aligned}$$

Столбцы можно набирать по одному, что позволяет доказывать оценку по индукции. А именно, после взятия нового столбца число столбцов в $\tilde{A}$ и U уменьшается на 1 (так как этот столбец уже использован, и для него ничего не нужно пересчитывать), и число строк V уменьшается на 1 (так как к выбранному столбцу происходит ортогонализация, что на единицу понижает размерность).

Пусть осталось еще k строк в V, то есть $V \in \mathbb{C}^{k\times(N-r+k)}$ и $\tilde{A} \in \mathbb{C}^{M\times(N-r+k)}$. Выберем новый столбец $\tilde{C}_j \in \mathbb{C}^M$ из условия

$$j = \arg\min_i \left\|\tilde{C}_i\right\|_2 / \|v_i\|_2, \tag{2.64}$$

$v_j \in \mathbb{C}^k$ – выбранный столбец V. Заметим, что $\mathbb{E}_j \left\|\tilde{C}_j\right\|_2^2 = \left\|\tilde{A}\right\|_2^2 / (N - r + k)$ и $\mathbb{E}_j \left\|v_j\right\|_2^2 = k/(N - r + k)$, а потому если j соответствует минимуму (2.64), то

$$\left\|\tilde{C}_j w_j\right\|_2 = \left\|\tilde{C}_j v_j^+ V\right\|_2 = \left\|\tilde{C}_j v_j^+\right\|_2 = \left\|\tilde{C}_j\right\|_2 / \left\|V_j\right\|_2 \leqslant 1/\sqrt{k}.$$

При этом достаточно найти j для которого $\left\|\tilde{C}_j\right\|_2 / \left\|v_j\right\|_2 \leqslant 1/\sqrt{k}$. Такой столбец всегда можно найти, так как всегда можно найти отношение, не превосходящее отношение средних неотрицательных величин.

Тогда ошибка на следующем шаге будет оцениваться как (с учетом ортогональности строк $\tilde{A}$ и V)

$$\left\|\tilde{A} - \tilde{C}_j w_j\right\|_{2,F}^2 \leqslant \left\|\tilde{A}\right\|_{2,F}^2 + \left\|\tilde{C}_j w_j\right\|_2^2 \leqslant \left\|\tilde{A}\right\|_{2,F}^2 + \frac{1}{k}\left\|\tilde{A}\right\|_F^2.$$

Таким образом, ошибка на следующем шаге определяется ошибкой на предыдущем шаге. Подставляя предыдущие $k - 1$ шагов, получаем

$$\left\|\tilde{A} - \tilde{C}_{j_1,\ldots,j_k} W_{j_1,\ldots,j_k}\right\|_{2,F}^2 \leqslant \|A - Z\|_{2,F}^2 + \frac{k}{r - k + 1}\|A - Z\|_F^2.$$

При $k = r$ получаем выражения (2.62) и (2.63). Каждый шаг занимает $O(MN)$, а потому полная сложность метода $O(MNr)$. □

Замечание 2.8. Если $\operatorname{rank} A \geqslant r$, то всегда можно за то же время найти столбцы C полного ранга, чтобы гарантировать, что $\operatorname{rank}(CC^+A) = r$. Для этого достаточно удалить из столбцов C линейно зависимые (что не увеличит погрешность), а затем добавить вместо них произвольные линейно независимые от уже выбранных (что также не увеличит погрешность).

Замечание 2.9. В пределе $\sigma_r(A)/\sigma_{r+1}(A) \to \infty$ данный алгоритм совпадает с алгоритмом дерандомизации из [43]. В таком пределе будут совпадать критерии набора столбцов, а значит алгоритмы будут набирать те же самые столбцы. Самый быстрый вариант дерандомизации был получен в [47] и занимает $O\left(rM^2N\right)$ операций. Однако, если его применять к матрице R после QR разложение $A = QR$, сложность сокращается до $O\left((M + Nr)\min^2(M, N)\right)$.

Заметим, что $[I\ B] = \hat{U}^+U$, а потому может далее использоваться для расширения в алгоритмах поиска подматриц локально максимального объема [60]. Квадраты элементов c также (с точностью до 1) связаны с квадратами норм столбцов матрицы U.

Теорема 2.16. *Пусть даны матрицы* $A, Z \in \mathbb{C}^{M\times N}$, $\operatorname{rank} Z = r$. *Тогда за* $O(MNr)$ *можно найти такие строки* R *и столбцы* C *матрицы* A, *для которых одновременно*

$$\left\|A - CC^+AR^+R\right\|_F \leqslant \sqrt{2r + 2}\,\|A - Z\|_F \tag{2.65}$$

и

$$\left\|A - CC^+AR^+R\right\|_2 \leqslant \sqrt{2 + 2r(N - r)}\|A - Z\|_2. \tag{2.66}$$

Если $\operatorname{rank} A \geqslant r$*, то также найти такие (возможно, другие) строки* R *и столбцы* C*, для которых одновременно*

$$\left\|A - C\hat{A}^{-1}R\right\|_F \leqslant (r+1)\,\|A - Z\|_F \tag{2.67}$$

и

$$\left\|A - C\hat{A}^{-1}R\right\|_2 \leqslant \sqrt{1 + r\,(r+2)\,(N-r)}\|A - Z\|_2, \tag{2.68}$$

где $\hat{A} \in \mathbb{C}^{r\times r}$ *– подматрица на пересечении строк* R *и столбцов* C*.*

Доказательство. Неравенства (2.65)-(2.66) следуют из ортогональности оценок для аппроксимаций по строкам и столбцам в случае CC^+AR^+R приближения:

$$\begin{aligned}\left\|A - CC^+AR^+R\right\|_{2,F}^2 &= \left\|A - CC^+A + CC^+A - CC^+AR^+R\right\|_{2,F}^2 \\ &\leqslant \left\|A - CC^+A\right\|_{2,F}^2 + \left\|CC^+\left(A - AR^+R\right)\right\|_{2,F}^2 \\ &\leqslant \left\|A - CC^+A\right\|_{2,F}^2 + \left\|A - AR^+R\right\|_{2,F}^2,\end{aligned} \tag{2.69}$$

где мы воспользовались утверждением 1.1.

Для выбора столбцов C можно напрямую воспользоваться теоремой 2.15. Для выбора строк R можно применить теорему 2.15 к A^T. Подставив оценки из нее в (2.69), получим (2.65) и (2.66).

Для доказательства неравенства (2.67) введем строковую аппроксимацию $\Phi = WR = U\hat{U}^{-1}R$, где $U \in \mathbb{C}^{N\times r}$ – матрица левых сингулярных векторов Z. Если её ранг меньше r, согласно замечанию 2.8, вместо этого выбрать соответствующую тем же строкам аппроксимацию $\Phi = WR = AR^+R$ полного ранга. Для строковой аппроксимации, как и для столбцовой, также справедливы оценки теоремы 2.15. Повторяя доказательство теоремы 2.15, получаем

$$\begin{aligned}\|A - \Phi\|_F &= \|A - WR\|_F \\ &\leqslant \sqrt{r+1}\,\|A - UU^*A\|_F \\ &\leqslant \sqrt{(r+1)\,(\min(M,N) - r)}\,\|A - UU^*A\|_2 \\ &\leqslant \sqrt{(r+1)\,(\min(M,N) - r)}\,\|A - Z\|_2\,.\end{aligned}$$

С другой стороны, применяя теорему 2.15 теперь для $Z = \Phi$, получаем, используя (2.63):

$$\left\|A - C\hat{V}_\Phi^{-1}V_\Phi\right\|_F \leqslant \sqrt{r+1}\,\|A - \Phi\|_F \leqslant (r+1)\,\|A - Z\|_F \tag{2.70}$$

и

$$\begin{aligned}\left\|A - C\hat{V}_\Phi^{-1}V_\Phi\right\|_2^2 &\leqslant \|A - \Phi\|_2^2 + r\,\|A - \Phi\|_F^2 \\ &\leqslant (1 + r\,(\min(M,N) - r))\,\|A - Z\|_2^2 + r\,(r+1)\,(\min(M,N) - r)\,\|A - Z\|_2^2 \\ &= (1 + r\,(r+2)\,(\min(M,N) - r))\,\|A - Z\|_2^2,\end{aligned} \tag{2.71}$$

где $V_\Phi \in \mathbb{C}^{r\times N}$ – подматрица правых сингулярных векторов матрицы $U\hat{U}^{-1}R = \Phi = U_\Phi\Sigma_\Phi V_\Phi$.

Наконец, заметим, что

$$\hat{V}_\Phi^{-1} V_\Phi = \left(\hat{U}_\Phi \Sigma_\Phi \hat{V}_\Phi\right)^{-1} \hat{U}_\Phi \Sigma_\Phi V_\Phi = \left(\hat{W}\hat{A}\right)^{-1} \hat{W} R = \hat{A}^{-1}\hat{W}^{-1}\hat{W}R = \hat{A}^{-1}R,$$

поскольку $\hat{V}_\Phi = \hat{W}\hat{A}$ обратима по построению, а потому $\hat{W} \in \mathbb{C}^{r\times r}$, соответствующая строкам R, и $\hat{A}$ обязаны быть полного ранга.

Таким образом,

$$A - C\hat{V}_\Phi^{-1} V_\Phi = A - C\hat{A}^{-1}R,$$

и из (2.70) и (2.71) следует (2.67) и (2.68). □

Следствие 2.11. При $Z = A_r$ сокращенному сингулярному разложению матрицы A ранга r, получаем оценку

$$\left\|A - C\hat{A}^{-1}R\right\|_F \leqslant (r+1)\left\|A - A_r\right\|_F. \tag{2.72}$$

В [47] показано, как подматрицу, удовлетворяющую (2.72) можно найти за $O\left(M^3 N r\right)$ операций. В теореме 2.16 для получения наилучшей оценки по норме Фробениуса требуется построить $Z = A_r$, а потому полная стоимость составляет $O\left(MN\min\left(M,N\right)\right)$.

Заметим, что наш метод соответствует методу из [80], когда Z получена из сингулярного разложения, а первые r сингулярных чисел стремятся к бесконечности. В этом случае оценка для нормы Фробениуса в среднем по объему не изменится, зато при вычислении коэффициентов полиномов из сингулярных чисел останутся только те множители, где присутствуют первые r сингулярных чисел. И при вычислении условного матожидания при фиксированном новом столбце, в правой части вместо отношения различных произведений сингулярных чисел будет присутствовать норма Фробениуса, что позволяет пересчитывать только её. Таким образом, наш алгоритм полностью соответствует максимизации условного матожидания, когда первые r сингулярных чисел бесконечно велики.

Чтобы получить оценку для числа столбцов, большего ранга аппроксимации, воспользуемся идеей, аналогичной теореме 2.13.

Теорема 2.17. *Пусть $A \in \mathbb{C}^{M\times N}$, $\operatorname{rank} A \geqslant n$. Тогда для любого n найдутся столбцы $C \in \mathbb{C}^{M\times n}$, строки $R \in \mathbb{C}^{n\times N}$ и подматрица на их пересечении $\hat{A} \in \mathbb{C}^{n\times n}$ такие, что для любого $r \leqslant n$ справедлива оценка*

$$\left\|A - C\hat{A}^{-1}R\right\|_F \leqslant \frac{n+1}{\sqrt{n-r+1}}\left\|A - A_r\right\|_F. \tag{2.73}$$

Доказательство. Будем действовать аналогично доказательству теоремы 2.13. Выберем подматрицу $\hat{A}$ с вероятностью

$$P\left(\hat{A}\right) = \frac{\mathcal{V}^2\left(\hat{A}\right)}{\sum\limits_{\mathcal{I},|\mathcal{I}|=n}\sum\limits_{\mathcal{J},|\mathcal{I}|=n}\mathcal{V}^2\left(A_{\mathcal{I},\mathcal{J}}\right)}. \tag{2.74}$$

Оценим квадрат нормы Фробениуса ошибки как сумму квадратов погрешностей в каждом элементе A. Погрешность в каждом элементе получим, рассмотрев соответствующие расширения $\tilde{A} \in \mathbb{C}^{(n+1)\times(n+1)}$ подматрицы $\hat{A}$ одним новым столбцом и одной новой строкой и применив второй пункт леммы 2.1:

$$\left\|A - C\hat{A}^{-1}R\right\|_F^2 = \sum_{\tilde{A}\supset\hat{A},\tilde{A}\in\mathbb{C}^{(n+1)\times(n+1)}} \frac{\mathcal{V}^2\left(\tilde{A}\right)}{\mathcal{V}^2\left(\hat{A}\right)}.$$

После подстановки вероятностей получим

$$\mathbb{E}_C\left\|A - C\hat{A}^{-1}R\right\|_F^2 = \frac{\sum\limits_{\tilde{A}\supset\hat{A},\tilde{A}\in\mathbb{C}^{(n+1)\times(n+1)}} \mathcal{V}^2\left(\tilde{A}\right)}{\sum\limits_{\mathcal{I},|\mathcal{I}|=n}\sum\limits_{\mathcal{J},|\mathcal{I}|=n} \mathcal{V}^2\left(A_{\mathcal{I},\mathcal{J}}\right)}.$$

На этот раз подматрицы $\tilde{A}$ содержат $(n+1)^2$ различных подматриц $\hat{A}$, а потому они будут посчитаны $(n+1)^2$ раз. Разделив всё на $\|A - A_r\|_F^2 = \sum\limits_{k\geqslant r+1}\sigma_k^2(A)$ и применив лемму 1.1, получим в среднем

$$\frac{\mathbb{E}_C\left\|A - C\hat{A}^{-1}R\right\|_F^2}{\|A - A_r\|_F^2} = (n+1)^2\frac{\sum\limits_{i_1<\ldots<i_{n+1}}\sigma_{i_1}^2(A)\cdot\ldots\cdot\sigma_{i_{n+1}}^2(A)}{\sum\limits_{i_1<\ldots<i_n}\sigma_{i_1}^2(A)\cdot\ldots\cdot\sigma_{i_n}^2(A)\sum\limits_{k\geqslant r+1}\sigma_k^2(A)}, \tag{2.75}$$

что отличается от (2.58) ровно в $n+1$ раз. Так как квадрат матожидания отличается от (2.58) не более, чем в $n+1$ раз, то оценка погрешности будет отличаться от (2.55) не более, чем в $\sqrt{n+1}$ раз, что доказывает (2.73). □

Замечание 2.10. Использование не квадратной подматрицы в данном случае только ухудшит оценку: коэффициент будет $\sqrt{\frac{(m+1)(n+1)}{\min(m,n)-r+1}}$. Улучшение коэффициента, как и для теоремы 2.13, возможно только если значение $\min(M,N)$ близко к n.

Для крестовой аппроксимации результаты, позволяющие вывести наилучшие оценки, были получены в [83]. А именно, был доказан следующий результат.

Теорема 2.18 ([83]). *Пусть $A \in \mathbb{C}^{M\times N}$, $A = Z + F$, $\operatorname{rank} Z = r$. Тогда для произвольной подматрицы $\hat{V} \in \mathbb{C}^{r\times n}$ полного ранга матрицы $V \in \mathbb{C}^{r\times N}$ правых сингулярных векторов Z и соответствующих ей столбцов $C \in \mathbb{C}^{M\times n}$ найдется матрица $S \in \mathbb{C}^{n\times n}$, $\operatorname{rank} S\left(\hat{V}S\right)^+ = r$, такая, что*

$$\begin{aligned}\left\|A - CS\left(\hat{V}S\right)^+V\right\|_F^2 &\leqslant \lim_{k\to\infty}\left\|Z' - C'C'^+Z'\right\|_F^2 + \|F'\|_F^2\\ &= \lim_{k\to\infty}\left(\left\|A' - C'C'^+A'\right\|_F^2 + \left\|C'C'^+F'\right\|_F^2\right)\\ &= \lim_{k\to\infty}\left\|A' - C'C'^+Z'\right\|_F^2,\end{aligned} \tag{2.76}$$

где $Z' = k(Z + FZ^+Z)$, $F' = F(I - Z^+Z)$, а C' соответствуют столбцам C в матрице $A' = Z' + F'$.

При этом матрица S представима в виде

$$S = \left(\hat{V}^*\hat{V} + \lambda_0 \left(C - AV^*\hat{V}\right)^* \left(C - AV^*\hat{V}\right)\right)^{-1/2} \tag{2.77}$$

для тех λ_0, что допускают обратимость.

Чтобы определение S не требовало обратимости, вместо (2.77) можно использовать оптимальную матрицу S. Одним из возможных будет набор решений вида

$$S = \left(\left(C - C\hat{V}^+\hat{V}\right)^+ \left(C\hat{V}^+\hat{V} - AV^*\hat{V}\right)\right)^+ + \left(I - \hat{V}^+\hat{V}\right) + \varepsilon\hat{V}^+\hat{V}.$$

С помощью другого набора можно задать прямоугольную матрицу S', оставив лишь r столбцов:

$$S' = \hat{V}^+ + \left(C - C\hat{V}^+\hat{V}\right)^+ \left(C\hat{V}^+ - AV^*\right). \tag{2.78}$$

При каких-то ненулевых ε всегда выполнено $\operatorname{rank} S = n$, S и S' не зависят от k или от веса вносимой в C погрешности, если погрешность ортогональна $\hat{V}$. Таким образом, с новой матрицей S условие теоремы будет справедливо даже в отсутствии обратимости, если далее нам удастся доказать его в случае, когда условие на обратимость выполняется, что мы и сделаем.

Нам нужно лишь оценить правую часть (2.76), чтобы получить итоговую оценку столбцовой аппроксимации.

Теорема 2.19. *Пусть $A \in \mathbb{C}^{M\times N}$, $A = Z + F$, $\operatorname{rank} Z = r$. Тогда для любого n найдутся столбцы $C \in \mathbb{C}^{M\times n}$, которые соответствуют некоторой подматрице $\hat{V} \in \mathbb{C}^{r\times n}$ матрицы $V \in \mathbb{C}^{r\times N}$ правых сингулярных векторов Z, и матрица $S \in \mathbb{C}^{n\times n}$ (ранга n) такие, что*

$$\left\|A - CS\left(\hat{V}S\right)^+ V\right\|_F \leqslant \|F\|_F \cdot \sup_A \min_C \frac{\|A - (CC^+A)_r\|_F}{\|A - A_r\|_F}. \tag{2.79}$$

Таким образом, левая часть (2.79) позволяет достичь наилучшего возможного коэффициента аппроксимации. Оценка данного коэффициента нами получена в теореме 2.14, однако мы не записываем её здесь в явном виде, так как эта оценка может быть неточной, хоть и является асимптотически точной, с погрешностью не более $O\left(r^{3/2}/n^{3/2}\right)$.

Доказательство. Воспользуемся теоремой 2.18 для конечного k. Далее в этом доказательстве опустим штрихи, так что $A = Z + F$, $Z = kZ_0$, $\operatorname{rank} Z_0 = r$, $Z_0F^* = ZF^* = 0$. Благодаря ортогональности Z и F при достаточно большом k получаем $Z = A_r$. Осталось сравнить при данном достаточно большом k построенную аппроксимацию вида CC^+Z с наилучшей аппроксимацией, имеющей вид $(CC^+A)_r = (CC^+Z + CC^+F)_r$.

Для r столбцов проектор $P_Z = Z_CZ_C^+ = Z_{0C}Z_{0C}^+ = Z_0Z_0^+ = ZZ^+$ отличается от $P_CP_Z = CC^+Z_{0C}Z_{0C}^+ = (Z_{0C} + F_C/k)(Z_C + F_C/k)^+ Z_CZ_C^+$ на величину $O(1/k)$, поэтому в нашем случае справедливо равенство

$$CC^+ = ZZ^+ + ZZ^+ \cdot O(1/k) + \left(I - ZZ^+\right)(X + O(1/k)),$$

где $X \in \mathbb{C}^{M\times M}$ описывает проектор CC^+ в пространстве, ортогональном ZZ^+ при $k \to \infty$.

Таким образом, используя ортогональность $Z^+F = 0$, получаем

$$CC^+F = ZZ^+ \cdot O\left(1/k\right) \cdot F + \left(I - ZZ^+\right)\left(X + O\left(1/k\right)\right)F.$$

Другими словами, CC^+F ортогонален CC^+Z и в строках, и в столбцах с точностью до величины $ZZ^+ \cdot O\left(1/k\right) \cdot F = O\left(1/k\right)$, а потому

$$\left(CC^+Z + CC^+F\right)_r = CC^+Z + O\left(1/k\right),$$

а значит в пределе $k \to \infty$ погрешности $A - CC^+Z$ и $A - \left(CC^+A\right)_r$ совпадают. То есть, возвращаясь к штрихам,

$$\begin{aligned}\left\|A - CS\left(\hat{V}S\right)^+ V\right\|_F &\leqslant \lim_{k\to\infty}\left\|A' - C'C'^+Z'\right\|_F \\ &= \lim_{k\to\infty}\left\|\left(A' - C'C'^+A'\right)_r\right\|_F \\ &\leqslant \lim_{k\to\infty}\left\|A' - A'_r\right\|_F \cdot \sup_A \min_C \frac{\|A - \left(CC^+A\right)_r\|_F}{\|A - A_r\|_F}.\end{aligned}$$

Осталось вспомнить, что при достаточно большом k верно равенство $A'_r = Z'$, а потому

$$\left\|A' - A'_r\right\|_F = \|A' - Z'\|_F = \|F'\|_F = \left\|F\left(I - ZZ^+\right)\right\|_F \leqslant \|F\|_F.$$

□

Замечание 2.11. $\operatorname{rank}\hat{V} = r$ не является для нас ограничением при выборе оптимальных столбцов, поскольку при $k \to \infty$ аппроксимация CC^+Z будет иметь ранг не ниже r (иначе погрешность стремится к бесконечности), а начиная с какого-то k

$$\operatorname{rank}\hat{V} = \operatorname{rank} Z_C = \operatorname{rank} CC^+Z_C = \operatorname{rank} CC^+Z = r.$$

Замечание 2.12. На практике наилучшей известной оценки точности столбцовой (и далее крестовой) аппроксимации можно достичь, построив $\left(CC^+A\right)_r$ аппроксимацию, соответствующую пределу $k \to \infty$. Для этого достаточно соответствующим образом (аналогично случаю $n = r$, который соответствует теореме 2.15) модифицировать алгоритм поиска CC^+A аппроксимации [46] (или его дерандомизированную версию), а затем набрать столбцы, модифицировав соответствующим образом leverage scores в [43], что даст в итоге оценку утверждения 2.5. В итоге такой выбор (в среднем достигающий полученной оценки: он использует рандомизированные алгоритмы) потребует $O\left(\min\left(M^2, N^2\right)N\sqrt{rn}\right)$ операций.

Теорема 2.20. *Пусть $A \in \mathbb{C}^{M\times N}$, $\operatorname{rank} A \geqslant r$. Тогда для любых m, n и r найдутся столбцы $C \in \mathbb{C}^{M\times n}$, строки $R \in \mathbb{C}^{m\times N}$, подматрица на их пересечении $\hat{A} \in \mathbb{C}^{m\times n}$ и ортогональный проектор $P \in \mathbb{C}^{m\times m}$ ранга r такие, что*

$$\left\|A - C\left(\hat{A}P\right)^+ R\right\|_F \leqslant \|A - A_r\|_F \cdot \left(\sup_A \min_C \frac{\|A - \left(CC^+A\right)_r\|_F}{\|A - A_r\|_F}\right)\left(\sup_A \min_R \frac{\|A - \left(AR^+R\right)_r\|_F}{\|A - A_r\|_F}\right).$$

Так как при больших n коэффициент погрешности здесь порядка $1 + r/n$, этот результат улучшает (асимптотически) наилучшую известную верхнюю оценку из [80] с коэффициентом $1 + 40\frac{r}{n-4r}$ примерно в 40 раз.

Доказательство. Сначала построим строковую аппроксимацию WR, $\operatorname{rank} W = r$ матрицы A (например, на основе теоремы 2.19 или используя $WR = (AR^+R)_r$) такую, что

$$\|A - WR\|_F \leqslant \|A - A_r\|_F \cdot \sup_A \min_R \frac{\|A - (AR^+R)_r\|_F}{\|A - A_r\|_F}$$

и $\operatorname{rank} WR = r$. Условие на ранг всегда можно выполнить, поскольку если $\operatorname{rank} AR^+R < r$, то строки R линейно зависимы, и можно удалить любую из них и добавить строку, увеличивающую ранг и уменьшающую погрешность (так как новый проектор $I - RR^+$ ранга $N - r$ будет проектировать на подпространство, содержащееся в исходном проекторе).

Теперь обозначим $Z = WR$ и воспользуемся теоремой 2.19.

Сразу заметим, что согласно замечанию 2.11, наилучшая столбцовая аппроксимация с использованием Z будет соответствовать столбцам ранга r в Z, то есть

$$\operatorname{rank} Z_C = \operatorname{rank} \left(W\hat{A}\right) = r.$$

Отсюда, в частности, следует, что $W = \hat{A}\hat{A}^+W^+$.

Введем (сокращенное) сингулярное разложение $Z = U\Sigma V$. Тогда для аппроксимации с использованием S' (2.78):

$$\begin{aligned}
CS' \left(\hat{V}S'\right)^+ V &= CS' \left(\hat{V}^+\hat{V}S'\right)^+ \hat{V}^+V \\
&= CS' \left(\hat{V}^+\Sigma^{-1}U^*U\Sigma\hat{V}S'\right)^+ \hat{V}^+V \\
&= CS' \left(U\Sigma\hat{V}S'\right)^+ U\Sigma\hat{V}\hat{V}^+V \\
&= CS' \left(U\Sigma\hat{V}S'\right)^+ U\Sigma V \\
&= CS' \left(Z_CS'\right)^+ Z \\
&= CS' \left(W\hat{A}S'\right)^+ WR \\
&= CS' \left(W^+W\hat{A}S'\right)^+ R \\
&= CS' \left(\hat{A}\left(\hat{A}^+W^+W\hat{A}\right)S'\right)^+ R \\
&= C \left(\hat{A}\left(\hat{A}^+W^+W\hat{A}\right)S'S'^+\right)^+ R \\
&= C \left(\hat{A}P\right)^+ R,
\end{aligned}$$

где $P = \left(\hat{A}^+W^+W\hat{A}\right)S'S'^+$.

Согласно теореме 2.19, получаем

$$\left\|A - C\left(\hat{A}P\right)^{+}R\right\|_F \leqslant \|A - WR\|_F \cdot \sup_A \min_C \frac{\|A - (CC^{+}A)_r\|_F}{\|A - A_r\|_F}$$
$$\leqslant \|A - A_r\|_F \cdot \left(\sup_A \min_C \frac{\|A - (CC^{+}A)_r\|_F}{\|A - A_r\|_F}\right)\left(\sup_A \min_R \frac{\|A - (AR^{+}R)_r\|_F}{\|A - A_r\|_F}\right).$$

□

2.3.2. Нижние оценки

Начнем с известных результатов. Во-первых, напомним, что в [43] построены матрицы $A \in \mathbb{R}^{(r+1)\times(r+1)}$, такие что

$$\left\|A - CC^{+}A\right\|_F \geqslant (1-\varepsilon)\sqrt{r+1}\,\|A - A_r\|_F \tag{2.80}$$

для любого $\varepsilon > 0$. Там же был представлен алгоритм, достигающий точности $\sqrt{r+1}$, а потому данная оценка является точной.

Для $n > r$ столбцов в [46], была получена следующая оценка:

$$\left\|A - CC^{+}A\right\|_F \geqslant \sqrt{\frac{N-n}{N-r}\left(1 + \frac{r}{n} - o(1)\right)}\,\|A - A_r\|_F,$$

где $M = N \vdots r$, а $o(1)$ можно сделать сколь угодно малым.

Для аппроксимации ранга r на n столбцах наилучшая известная нам оценка получена в [84] и может быть записана в виде

$$\left\|A - \left(CC^{+}A\right)_r\right\|_F \geqslant \sqrt{1 + \frac{r}{2n}}\,\|A - A_r\|_F,$$

где в матрице A было $M = 4nr$ строк и $N = (4n+1)r$ столбцов.

Здесь мы избавимся от ограничений на размеры матрицы A, получив схожие оценки с помощью аналога t-функции. Аналогично спектральной норме, можно определить t_F-функцию для нормы Фробениуса.

Определение 2.4.

$$t_F(r,n,N) = \sup_{\substack{U\in\mathbb{C}^{r\times N},\\ UU^{*}=I}} \min_{\substack{\hat{U}\in\mathbb{C}^{r\times n},\\ \operatorname{rank}\hat{U}=r}} \left\|\hat{U}^{+}\right\|_F.$$

Мы снова воспользуемся эквивалентным определением для построения оценок.

Лемма 2.4.

$$t_F(r,n,N) = \sup_{\substack{A\in\mathbb{C}^{r\times N},\\ \operatorname{rank}A=r}} \min_{\substack{\hat{A}\in\mathbb{R}^{r\times n},\\ \operatorname{rank}\hat{A}=r}} \left\|\hat{A}^{+}A\right\|_F.$$

Доказательство. Доказательство то же самое, что и в первой части леммы 2.2. □

$t_F(r,n,N)$ можно оценить снизу следующим образом.

Утверждение 2.6.

$$t_F(r, n, N) \geqslant \sqrt{r\frac{N}{n}}. \tag{2.81}$$

Доказательство. Сначала рассмотрим случай $N \vdots r$. Тогда мы можем выбрать

$$A = [I\ I\ \cdots I] \in \mathbb{R}^{r\times N}.$$

В этом случае $\|\hat{A}^+A\|_F^2 = \frac{N}{r}\|\hat{A}^+\|_F^2$. Если i-й столбец единичной матрицы встречается в $\hat{A}$ k_i раз, то i-й столбец $\hat{A}^+$ будет содержать k_i ненулевых элементов, каждый из которых равен $1/k_i$. Таким образом, при минимизации $\|\hat{A}^+\|_F^2$, какие-то $n \bmod r$ столбцов единичной матрицы будут встречаться $\left\lfloor\frac{n}{r}\right\rfloor + 1$ раз, а оставшиеся $r - n \bmod r$ будут встречаться $\left\lfloor\frac{n}{r}\right\rfloor$ раз (если $n \vdots r$, то оптимумом является равное количество каждого столбца). Такая подматрица $\hat{A}$ приводит к оценке

$$\|\hat{A}^+A\|_F^2 = \frac{N}{r}\left(\frac{n \bmod r}{\left\lfloor\frac{n}{r}\right\rfloor + 1} + \frac{r - n \bmod r}{\left\lfloor\frac{n}{r}\right\rfloor}\right).$$

Используя равенство $r\left\lfloor\frac{n}{r}\right\rfloor = n - n \bmod r$, приходим к оценке

$$\begin{aligned}\|\hat{A}^+A\|_F^2 &= \frac{rN\,(n + r - 2\,(n \bmod r))}{(n - n \bmod r)\,(n + r - n \bmod r)} \\ &= r\frac{N}{n} + \frac{rN\,(n \bmod r)\,(r - n \bmod r)}{n\,(n - n \bmod r)\,(n + r - n \bmod r)}.\end{aligned} \tag{2.82}$$

Теперь рассмотрим случай $N \bmod r > 0$. Выберем

$$A = [I\ I\ \cdots\ I\ B] \in \mathbb{R}^{r\times N}.$$

с подматрицей $B \in \mathbb{R}^{r\times(N \bmod r)}$, состоящей из различных столбцов единичной матрицы. Тогда оптимальный выбор $\hat{A}$ не изменится (в ней все равно будет по $\left\lfloor\frac{n}{r}\right\rfloor + 1$ каких-то столбцов и по $\left\lfloor\frac{n}{r}\right\rfloor$ других столбцов). Поэтому мы можем воспользоваться уравнением (2.82) для первых $N - N \bmod r$ столбцов (подставив в него $N - N \bmod r$ вместо N) и добавив $\|\hat{A}^+B\|_F^2$ в явном виде:

$$\begin{aligned}\|\hat{A}^+A\|_F^2 &= r\frac{N - N \bmod r}{n} + \frac{r\,(N - N \bmod r)\,(n \bmod r)\,(r - n \bmod r)}{n\,(n - n \bmod r)\,(n + r - n \bmod r)} + \|\hat{A}^+B\|_F^2 \\ &= r\frac{N}{n} + \frac{rN\,(n \bmod r)\,(r - n \bmod r)}{n\,(n - n \bmod r)\,(n + r - n \bmod r)} \\ &\quad - \frac{r\,(N \bmod r)\,(n + r - 2\,(n \bmod r))}{(n - n \bmod r)\,(n + r - n \bmod r)} + \|\hat{A}^+B\|_F^2.\end{aligned} \tag{2.83}$$

Для оценки последнего слагаемого рассмотрим два случая: $N \bmod r \leqslant n \bmod r$ и $N \bmod r > n \bmod r$. В первом случае оптимальным выбором будет выбор столбцов из B $\left\lfloor\frac{n}{r}\right\rfloor + 1$ раз в $\hat{A}$, что приводит к

$$\|\hat{A}^+B\|_F^2 = \frac{N \bmod r}{\left\lfloor\frac{n}{r}\right\rfloor + 1} = \frac{r\,(N \bmod r)}{r\left\lfloor\frac{n}{r}\right\rfloor + r} = \frac{r\,(N \bmod r)}{n + r - n \bmod r}.$$

Тогда

$$\begin{aligned}\|\hat{A}^{+}A\|_F^2 &= r\frac{N}{n} + \frac{rN\,(n \bmod r)\,(r - n \bmod r)}{n\,(n - n \bmod r)\,(n + r - n \bmod r)} \\ &- \frac{r\,(N \bmod r)\,(n + r - 2\,(n \bmod r))}{(n - n \bmod r)\,(n + r - n \bmod r)} + \frac{r\,(N \bmod r)}{n + r - n \bmod r} \\ &= r\frac{N}{n} + \frac{rN\,(n \bmod r)\,(r - n \bmod r)}{n\,(n - n \bmod r)\,(n + r - n \bmod r)} \\ &- \frac{r\,(N \bmod r)\,(r - n \bmod r)}{(n - n \bmod r)\,(n + r - n \bmod r)} \\ &= r\frac{N}{n} + \frac{r\,(r - n \bmod r)}{(n - n \bmod r)\,(n + r - n \bmod r)}\left(\frac{N}{n}\,(n \bmod r) - N \bmod r\right) \\ &\geqslant r\frac{N}{n}\end{aligned}$$

с учетом предположений $N/n \geqslant 1$ и $N \bmod r \leqslant n \bmod r$.

Наконец, рассмотрим случай $N \bmod r > n \bmod r$. Тогда

$$\begin{aligned}\|\hat{A}^{+}B\|_F^2 &= \frac{n \bmod r}{\left\lfloor \frac{n}{r} \right\rfloor + 1} + \frac{N \bmod r - n \bmod r}{\left\lfloor \frac{n}{r} \right\rfloor} \\ &= \frac{r\,(n \bmod r)}{n + r - n \bmod r} + \frac{r\,(N \bmod r - n \bmod r)}{n - n \bmod r} \\ &= \frac{r\,((N \bmod r)\,(n + r - n \bmod r) - r\,(n \bmod r))}{(n - n \bmod r)\,(n + r - n \bmod r)}.\end{aligned}$$

Подстановка в (2.83) приводит к

$$\begin{aligned}\|\hat{A}^{+}A\|_F^2 &= r\frac{N}{n} + \frac{rN\,(n \bmod r)\,(r - n \bmod r)}{n\,(n - n \bmod r)\,(n + r - n \bmod r)} \\ &- \frac{r\,(N \bmod r)\,(n + r - 2\,(n \bmod r))}{(n - n \bmod r)\,(n + r - n \bmod r)} \\ &+ \frac{r\,((N \bmod r)\,(n + r - n \bmod r) - r\,(n \bmod r))}{(n - n \bmod r)\,(n + r - n \bmod r)} \\ &= r\frac{N}{n} + \frac{rN\,(n \bmod r)\,(r - n \bmod r)}{n\,(n - n \bmod r)\,(n + r - n \bmod r)} \\ &- \frac{r\,(r - N \bmod r)\,(n \bmod r)}{(n - n \bmod r)\,(n + r - n \bmod r)} \\ &= r\frac{N}{n} + \frac{r\,(n \bmod r)}{(n - n \bmod r)\,(n + r - n \bmod r)}\left(\frac{N}{n}\,(r - n \bmod r) - (r - N \bmod r)\right) \\ &> r\frac{N}{n}\end{aligned}$$

в предположении $N \bmod r > n \bmod r$ и с учетом $N/n \geqslant 1$. Таким образом,

$$t_F(r, n, N) = \sup_{\substack{A \in \mathbb{C}^{r \times N}, \\ \operatorname{rank} A = r}} \min_{\substack{\hat{A} \in \mathbb{C}^{r \times n}, \\ \operatorname{rank} \hat{A} = r}} \|\hat{A}^{+}A\|_F \geqslant r\frac{N}{n}.$$

□

Теперь, когда у нас есть оценка t_F-функции, осталось лишь доказать аналог теоремы 2.12 для нормы Фробениуса.

Теорема 2.21.

$$\sup_{A\in\mathbb{C}^{N\times N}} \min_{C} \frac{\|A - CC^{+}A\|_F}{\|A - A_r\|_F} \geqslant \sqrt{1 + \frac{t_F^2(r,n,N) - n}{N-r}}. \tag{2.84}$$

Доказательство. Рассмотрим матрицу

$$A = \left[\begin{array}{ccc} \multicolumn{3}{c}{U} \\ \hline \varepsilon & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \varepsilon \end{array}\right] \in \mathbb{R}^{(N+r)\times N}, \tag{2.85}$$

где $U \in \mathbb{C}^{r\times N}$ – произвольная матрица с ортонормированными строками.

Все сингулярные числа матрицы $A^*A = U^*U + \varepsilon^2 I$ равны либо $1+\varepsilon^2$, либо ε^2, а потому для матрицы A

$$\|A - A_r\|_F^2 = (N-r)\varepsilon^2. \tag{2.86}$$

Рассмотрим разбиение матрицы U

$$U = [U_1\ U_2], \quad U_1 \in \mathbb{C}^{r\times n}$$

и соответствующее разбиение матрицы A:

$$A = \left[\begin{array}{c|c} U_1 & U_2 \\ \hline \varepsilon I_n & 0 \\ \hline 0 & \varepsilon I_{N-n} \end{array}\right].$$

Без ограничения общности будем считать, что для $CC^{+}A$ приближения выбраны первые n столбцов. Тогда

$$C^{+} = (C^*C)^{-1}C^* = \left(U_1^*U_1 + \varepsilon^2 I\right)^{-1} \left[\begin{array}{ccc} U_1^* & \varepsilon I & 0 \end{array}\right],$$

$$CC^{+} = \left[\begin{array}{c|c|c} U_1\left(U_1^*U_1 + \varepsilon^2 I\right)^{-1}U_1^* & \varepsilon U_1\left(U_1^*U_1 + \varepsilon^2 I\right)^{-1} & 0 \\ \hline \varepsilon\left(U_1^*U_1 + \varepsilon^2 I\right)^{-1}U_1^* & \varepsilon^2\left(U_1^*U_1 + \varepsilon^2 I\right)^{-1} & 0 \\ \hline 0 & 0 & 0 \end{array}\right]$$

и

$$CC^{+}A = \left[\begin{array}{ccc|ccc} \multicolumn{3}{c|}{U_1} & \multicolumn{3}{c}{U_1\left(U_1^*U_1 + \varepsilon^2 I\right)^{-1}U_1^*U_2} \\ \hline \varepsilon & 0 & 0 & & & \\ 0 & \ddots & 0 & \multicolumn{3}{c}{\varepsilon\left(U_1^*U_1 + \varepsilon^2 I\right)^{-1}U_1^*U_2} \\ 0 & 0 & \varepsilon & & & \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] = \left[\begin{array}{cc} B_{11} & B_{12} \\ B_{21} & B_{22} \\ B_{31} & B_{32} \end{array}\right].$$

Сразу заметим, что если матрица U_1 вырождена, то $\operatorname{rank} B_{12} < r$, и погрешность в блоке B_{12} не меньше 1 по норме Фробениуса, что не может быть оптимумом при достаточно малом ε. Поэтому далее можем считать, что $\operatorname{rank} U_1 = r$.

Погрешность $A - CC^+A$ мы оценим снизу через погрешность в блоках B_{32} и B_{22}:

$$\begin{aligned}
\|A - CW\|_F^2 &\geqslant (N-n)\,\varepsilon^2 + \|\varepsilon \left(U_1^* U_1 + \varepsilon^2 I\right)^{-1} U_1^* U_2\|_F^2 \\
&\qquad \text{заменим } \begin{bmatrix} 0 & U_2 \end{bmatrix} = U - \begin{bmatrix} U_1 & 0 \end{bmatrix} \\
&\geqslant (N-n)\,\varepsilon^2 + \varepsilon^2 \| \left(U_1^* U_1 + \varepsilon^2 I\right)^{-1} U_1^* \|_F^2 - \varepsilon^2 \| \left(U_1^* U_1 + \varepsilon^2 I\right)^{-1} U_1^* U_1 \|_F^2.
\end{aligned}$$

Заметим, что

$$\left(U_1^* U_1 + \varepsilon^2 I\right)^{-1} U_1^* = U_1^* \left(U_1 U_1^* + \varepsilon^2 I\right)^{-1} = U_1^+ + O\left(\varepsilon^2\right).$$

Тогда

$$\begin{aligned}
&\|A - CC^+A\|_F^2 \geqslant (N-n)\varepsilon^2 + \varepsilon^2 \|U_1^+ + O\left(\varepsilon^2\right)\|_F^2 - \varepsilon^2 \|U_1^+ U_1 + O\left(\varepsilon^2\right)\|_F^2 = \\
&= (N-r)\varepsilon^2 + \varepsilon^2 \|U_1^+\|_F^2 - \varepsilon^2 \|I_n\|_F^2 + O\left(\varepsilon^4\right) = \\
&\qquad \text{подставим } \varepsilon^2 \text{ из (2.86)} \\
&= \left(O(\varepsilon^2) + 1 + \frac{1}{N-r}\|U_1^+\|_F^2 - \frac{n}{N-r}\right) \|A - A_r\|_F^2.
\end{aligned}$$

Устремляя ε к нулю, получаем доказательство для $(N+r)\times N$ матриц. Чтобы получить то же утверждение для $N\times N$ матриц, достаточно, как и в теореме 2.21, рассмотреть матрицу $B \in \mathbb{C}^{N\times N}$ в подпространстве размерности N, соответствующему пространству столбцов A. □

Замечание 2.13. Данный результат можно использовать и в обратную сторону. Имея алгоритм построения столбцовой аппроксимации с достаточно высокой точностью и применяя его к матрице A вида (2.85) получаем алгоритм поиска подматрицы $\hat{U} \in \mathbb{C}^{r\times n}$ в произвольных ортонормированных строках $U \in \mathbb{C}^{r\times N}$, такой что

$$1 + \frac{\left\|\hat{U}\right\|_F^2 - n}{N-r} \leqslant \frac{\|A - CC^+A\|_F^2}{\varepsilon\,(N-r)} + O\left(\varepsilon\right).$$

Таким образом, алгоритмы построения столбцовых аппроксимаций автоматически позволяют осуществлять поиск сильно невырожденных подматриц.

Следствие 2.12.

$$\sup_{A\in\mathbb{C}^{N\times N}} \min_{C} \frac{\|A - (CC^+A)_r\|_F}{\|A - A_r\|_F} \geqslant \sqrt{1 + \frac{t_F^2(r,n,N) - r}{N-r}}. \tag{2.87}$$

Доказательство. Рассмотрим блок B_{21}, который ранее не менялся:

$$B_{21} = \varepsilon I \in \mathbb{C}^{n\times n}.$$

Любая аппроксимация ранга r матрицы A также является аппроксимацией ранга не выше r данного блока. Но любая аппроксимация этого блока содержит ошибку не меньше

$$\left\|B_{21} - B_{21,r}\right\|_F^2 \geqslant (n-r)\varepsilon,$$

поэтому квадрат погрешности возрастет как минимум не величину $\frac{n-r}{N-r}\|A - A_r\|_F^2$, что приводит к оценке (2.87). □

Для случая $n > r$ мы получаем необходимые оценки с помощью утверждения 2.6.

Теорема 2.22.

$$\sup_{A\in\mathbb{C}^{M\times N}} \min_C \frac{\|A - CC^+A\|_F}{\|A - A_r\|_F} \geqslant \sqrt{\left(1 + \frac{r}{n}\right)\frac{\min(M,N) - n}{\min(M,N) - r}}. \tag{2.88}$$

$$\sup_{A\in\mathbb{C}^{M\times N}} \min_C \frac{\|A - (CC^+A)_r\|_F}{\|A - A_r\|_F} \geqslant \sqrt{1 + \frac{r}{n}\cdot\frac{\min(M,N) - n}{\min(M,N) - r}}. \tag{2.89}$$

Доказательство. В обоих случаях достаточно доказать неравенства для квадратной матрицы $A \in \mathbb{C}^{\min(M,N)\times\min(M,N)}$, которая затем может быть расширена нулями. Неравенство (2.88) следует из подстановки (2.81) в (2.84), а неравенство (2.89) следует из подстановки (2.81) в (2.87). □

В отличие от спектральной нормы, у нас нет доказательства того, что верхние и нижние оценки совпадают. Тем не менее можно доказать, что они совпадают «в среднем». Данный результат будет получен в главе 3, теорема 3.3.

С точки зрения оценок для крестовых аппроксимаций, являющихся частным случаем столбцовых, неизвестно, существуют ли нижние оценки лучше. Для аппроксимаций симметричных неотрицательно определенных матриц нижние оценки есть, например, в [85], на основе того факта, что в этом случае наилучшая аппроксимация обязана быть симметричной (это справедливо и для спектральной нормы). Отметим, однако, что наилучшая скелетная аппроксимация не обязана быть симметричной, как показано в [47].

2.4. Точность основных видов столбцовых и крестовых аппроксимаций

Соберем полученные и известные оценки для различных видов столбцовых и крестовых аппроксимаций в таблицах 2.1 и 2.2. В таблицу 2.1 мы также включили оценки для средней ошибки, которые будут рассмотрены нами в следующей главе 3.

Оценки по спектральной норме используют оценку $t(r,n,N)$ из теоремы 4.10 раздела 4.6.

Среди известных ранее оценок отметим

Таблица 2.1: Наилучшие известные верхние оценки различных видов столбцовых и крестовых аппроксимаций. Оценки, полученные или улучшенные в данной работе, помечены жирным шрифтом.

Аппроксимация	$\frac{\left\|A-\tilde{A}\right\|_F}{\left\|A-\tilde{A}_r\right\|_F}$	$\frac{\left\|A-\tilde{A}\right\|_2}{\left\|A-\tilde{A}_r\right\|_2}$	$\frac{\left\|A-\tilde{A}\right\|_C}{\min\limits_{Z,\ \mathrm{rank}\ Z\leqslant r}\|A-Z\|_C}$
$CC^+A,\ n=r$	$\sqrt{r+1}$	$\sqrt{1+r\,(N-r)}$	$\mathbf{r+1}$
$CC^+A,\ n>r$	$\sqrt{\frac{n+1}{n-r+1}}$	$\frac{\sqrt{\mathbf{N}}}{\sqrt{\mathbf{n+1}}-\sqrt{\mathbf{r}}}+\sqrt{\frac{\mathbf{1}}{\mathbf{N(n+1)}}}$	–
$(CC^+A)_r,\ n>r$	$\sqrt{\mathbf{1}+\frac{\mathbf{r}}{\mathbf{n+r+1}-\sqrt{\mathbf{1+4r(n+1)}}}}$	$\sqrt{\mathbf{1}+\left(\frac{\sqrt{\mathbf{N}}}{\sqrt{\mathbf{n+1}}-\sqrt{\mathbf{r}}}+\sqrt{\frac{\mathbf{1}}{\mathbf{N(n+1)}}}\right)^{\mathbf{2}}}$	–
$CW_r,\ n>r$	$\sqrt{\mathbf{1}+\frac{\mathbf{r}}{\mathbf{n+r+1}-\sqrt{\mathbf{1+4r(n+1)}}}}$	$\frac{\sqrt{\mathbf{N}}}{\sqrt{\mathbf{n+1}}-\sqrt{\mathbf{r}}}+\sqrt{\frac{\mathbf{1}}{\mathbf{N(n+1)}}}$	$\mathbf{1}+\sqrt{\frac{\mathbf{nr}}{\mathbf{n-r+1}}}$
$C\hat{A}^{-1}R,\ n=r$	$\mathbf{r+1}$	$\sqrt{\mathbf{1}+\mathbf{r}\,(\mathbf{r}+\mathbf{2})\,(\mathbf{N}-\mathbf{r})}$	$(r+1)^2$
$C\hat{A}^{-1}R,\ n>r$	$\frac{\mathbf{n+1}}{\sqrt{\mathbf{n-r+1}}}$	$\sqrt{\frac{\mathbf{(n+1)^2}}{\mathbf{n-r+1}}\,(\mathbf{N}-\mathbf{n})}$	$\frac{\mathbf{(n+1)^2}}{\mathbf{n-r+1}}$
$C\left(P_r\hat{A}\right)^+R,\ n>r$	$\mathbf{1}+\frac{\mathbf{r}}{\mathbf{n+r+1}-\sqrt{\mathbf{1+4r(n+1)}}}$	$\left(\mathbf{1}+\frac{\mathbf{r}}{\mathbf{n+r+1}-\sqrt{\mathbf{1+4r(n+1)}}}\right)\sqrt{(\mathbf{N}-\mathbf{n})}$	$\frac{\mathbf{(n+1)^2}}{\mathbf{n-r+1}}$
$\mathbb{E}_{U,V}C\left(P_r\hat{A}\right)^+R,\ n>r$	$\frac{\mathbf{n+1}}{\mathbf{n-r+1}}$	$\frac{\mathbf{n+1}}{\mathbf{n-r+1}}\sqrt{\mathbf{N}-\mathbf{n}}$	$\frac{\mathbf{(n+1)^2}}{\mathbf{n-r+1}}$
$CC^+AR^+R,\ n=r$	$\sqrt{2\,(r+1)}$	$\sqrt{2+2r\,(N-r)}$	–
$CC^+AR^+R,\ n>r$	$\mathbf{1}+\frac{\mathbf{r}}{\mathbf{n+r+1}-\sqrt{\mathbf{1+4r(n+1)}}}$	$\sqrt{\mathbf{2}}\left(\frac{\sqrt{\mathbf{N}}}{\sqrt{\mathbf{n+1}}-\sqrt{\mathbf{r}}}+\sqrt{\frac{\mathbf{1}}{\mathbf{N(n+1)}}}\right)$	–
$(CC^+AR^+R)_r,\ n>r$	$\mathbf{1}+\frac{\mathbf{r}}{\mathbf{n+r+1}-\sqrt{\mathbf{1+4r(n+1)}}}$	$\sqrt{\mathbf{1}+\mathbf{2}\left(\frac{\sqrt{\mathbf{N}}}{\sqrt{\mathbf{n+1}}-\sqrt{\mathbf{r}}}+\sqrt{\frac{\mathbf{1}}{\mathbf{N(n+1)}}}\right)^{\mathbf{2}}}$	–
$CGR,\ n=r$	$\sqrt{2\,(r+1)}$	$\sqrt{2+2r\,(N-r)}$	$\mathbf{4r+1}$
$CGR,\ n>r$	$\mathbf{1}+\frac{\mathbf{r}}{\mathbf{n+r+1}-\sqrt{\mathbf{1+4r(n+1)}}}$	$\sqrt{\mathbf{1}+\mathbf{2}\left(\frac{\sqrt{\mathbf{N}}}{\sqrt{\mathbf{n+1}}-\sqrt{\mathbf{r}}}+\sqrt{\frac{\mathbf{1}}{\mathbf{N(n+1)}}}\right)^{\mathbf{2}}}$	$\mathbf{1}+\mathbf{4}\sqrt{\frac{\mathbf{nr}}{\mathbf{n-r+1}}}$

$\frac{\|A-CC^+A\|_F}{\|A-A_r\|_F} \leqslant \sqrt{r+1}$ для $n = r$ [43]
$\frac{\|A-CC^+A\|_F}{\|A-A_r\|_F} \leqslant \sqrt{\frac{n+1}{n-r+1}}$ [46]
$\frac{\|A-(CC^+A)_r\|_F}{\|A-A_r\|_F} \leqslant \left(1+\frac{2r}{n}\left(1+O\left(n^{-1/2}\right)\right)\right)$ при $n \to \infty$ [68]
$\frac{\|A-CGR\|_F}{\|A-A_r\|_F} \leqslant 1+40\frac{r}{n-4r}$ [13]
$\frac{\|A-CC^+A\|_2}{\|A-A_r\|_2} \leqslant \sqrt{1+r\,(N-r)}$ для $n = r$ [39]
$\frac{\|A-CC^+A\|_2}{\|A-A_r\|_2} \leqslant \frac{\sqrt{N}+\sqrt{n}}{\sqrt{n}-\sqrt{r}}$ [44]
$\frac{\|A-C\hat{A}^{-1}R\|_2}{\|A-A_r\|_2} \leqslant 1+r\,(N-r)$ для $n = r$ [7]
$\frac{\|A-CGR\|_2}{\|A-A_r\|_2} \leqslant 1+4\sqrt{1+r\,(N-r)}$ для $n = r$ [8]
$\frac{\left\|A-\tilde{A}\right\|_C}{\min\limits_{Z,\ \mathrm{rank}\,Z\leqslant r}\|A-Z\|_C} \leqslant (r+1)^2$ [10]

Таблица 2.2: Наилучшие известные нижние оценки столбцовых аппроксимаций. Оценки, полученные или улучшенные в данной работе отмечены жирным шрифтом.

Число столбцов	$\frac{\|A-CC^+A\|_F^2}{\|A-A_r\|_F^2}$	$\frac{\|A-(CC^+A)_r\|_F^2}{\|A-A_r\|_F^2}$	$\frac{\|A-CC^+A\|_2^2}{\|A-A_r\|_2^2}$
$n = r$	$r+1$		$\mathbf{(r+1)\left\lfloor\frac{\min(M,N)}{r+1}\right\rfloor}$
$n > r$	$\left(1+\frac{r}{n}\right)\frac{\min(M,N)-n}{\min(M,N)-r}$	$\mathbf{1+\frac{r}{n}\cdot\frac{\min(M,N)-n}{\min(M,N)-r}}$	$\mathbf{\frac{\min(M,N)-r+1}{n-r+1}}$

Схожие с полученными здесь оценки снизу были получены при определенных условиях и в более ранних работах (как и ранее, опустим супремум и минимум при их формулировке):
$\frac{\|A-CC^+A\|_F^2}{\|A-A_r\|_F^2} \geqslant r+1$ [43]
$\frac{\|A-CC^+A\|_2^2}{\|A-A_r\|_2^2} \geqslant N/n$ при $M = N+1$ [68]
$\frac{\|A-CC^+A\|_F^2}{\|A-A_r\|_F^2} \geqslant \left(1+\frac{r}{n}\right)\frac{N-n}{N-r}$ при $M = N \vdots r$ [46]
$\frac{\|A-(CC^+A)_r\|_F^2}{\|A-A_r\|_F^2} \geqslant 1+\frac{r}{2n}$ при $M = 4nr$, $N = (4n+1)r$ [84]

Глава 3. Вероятностные оценки точности

Столбцовые и крестовые аппроксимации являются важным аналогом сингулярного разложения, поскольку позволяют достичь сколь угодно близкой к оптимальной аппроксимации по норме Фробениуса. Наилучшие известные оценки позволяют достичь аппроксимации точности $1+\varepsilon$ за линейное по $1/\varepsilon$ число столбцов. Например, в [13], теорема 5.1, доказана следующая оценка.

Теорема 3.1 ([13])**.** *Существует рандомизированный алгоритм с вычислительной сложностью* $O\left((M+N)^2 n + r^4 \ln r\right)$*, который с вероятностью как минимум* 20% *находит в матрице* $A \in \mathbb{R}^{M\times N}$ *столбцы* $C \in \mathbb{R}^{M\times n}$*, строки* $R \in \mathbb{R}^{n\times N}$ *и строит матрицу* $U \in \mathbb{R}^{n\times n}$ *такую, что*

$$\|A - CUR\|_F \leqslant \left(1 + 32400\frac{r}{n-4r}\right)\|A - A_r\|_F,$$

где A_r *– матрица наилучшего приближения ранга* r *по норме Фробениуса.*

Как видно, оценка требует существенного числа строк и столбцов n, чтобы быть близкой к оптимальной. В [13] представлены также алгоритмы, требующие меньшего числа столбцов n (с коэффициентом 40 вместо 32400) для достижения приемлемой точности, однако их сложность существенно выше. Кроме того, в алгоритме из [13] при построении приближения используются все элементы исходной матрицы. Эти недостатки, как будет видно, отсутствуют у алгоритмов, основанных на принципе максимального объема и проективного объема.

Таким образом, на практике часто применяются более быстрые эмпирические методы поиска крестовой аппроксимации, такие как Cross 2D [4], алгоритм maxvol [11] и его обобщения [42, 60]. Все они связаны с идеей поиска подматрицы большого объема или большого проективного объема. В разделе 2.3.1 показано, что если брать случайные столбцы с вероятностью, пропорциональной квадрату объема, то можно, в среднем, достичь коэффициента $\sqrt{1+\frac{r}{n-r+1}}$ для аппроксимации ранга n [46]. Однако, того же нельзя сказать о столбцах или подматрице большого объема, которую ищут вышеуказанные алгоритмы.

В данном разделе мы получим оценки на точность по норме Фробениуса с помощью семейства RANDSVD случайных матриц. Так как любая матрица принадлежит RANDSVD семейству, соответствующему её сингулярным числам (см. определение 3.1 далее), можно ожидать, что и в реальных задачах погрешность будет часто удовлетворять полученным здесь оценкам. Численные эксперименты (см. главу 6) показывают высокое совпадение наблюдаемых погрешностей с их теоретическим предсказанием.

Сразу отметим, что использование подматриц максимального объема не позволяет гарантировать высокой точности аппроксимации. Именно поэтому и необходимо построение вероятностных оценок.

Действительно, рассмотрим матрицы $A \in \mathbb{R}^{(r+1)\times N}$ вида

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & \ddots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{N-r+1}} & \frac{1}{\sqrt{N-r+1}} & \dots & \frac{1}{\sqrt{N-r+1}} \\ 0 & 0 & 0 & -\frac{\varepsilon\sqrt{N-r}}{\sqrt{N-r+1}} & \frac{\varepsilon}{\sqrt{N-r}\sqrt{N-r+1}} & \dots & \frac{\varepsilon}{\sqrt{N-r}\sqrt{N-r+1}} \end{bmatrix}. \tag{3.1}$$

Каким бы не было значение ε, матрица максимального объема находится в первых r столбцах любой из матриц A. Обозначим эти столбцы через C,

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{N-r+1}} \\ 0 & 0 & 0 & -\frac{\varepsilon\sqrt{N-r}}{\sqrt{N-r+1}} \end{bmatrix} \in \mathbb{R}^{(r+1)\times r}.$$

Оценим ошибку наилучшего столбцового приближения CW для A. Заметим, что такое приближение достигается на матрице $W = C^+A$. Действительно, обозначив через b произвольный столбец в A, а через w соответствующий ему столбец W, найдем

$$\arg\min_w \|b - Cw\|_2 = C^+b.$$

Объединив все столбцы в матрицу W, получим $W = C^+A$. Поскольку ранг проектора $I - CC^+$ равен 1, то ошибка приближения $A - CC^+A$ будет ранга 1, и

$$\|A - CC^+A\|_2 = \|A - CC^+A\|_F.$$

Прямыми вычислениями получаем

$$\left\|A - CC^+A\right\|_2 = \left\| A - \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{1+\varepsilon^2(N-r)} & \frac{\varepsilon\sqrt{N-r}}{1+\varepsilon^2(N-r)} \\ 0 & 0 & 0 & \frac{\varepsilon\sqrt{N-r}}{1+\varepsilon^2(N-r)} & \frac{\varepsilon^2(N-r)}{1+\varepsilon^2(N-r)} \end{bmatrix} A \right\|_2.$$

Учитывая то, что ошибка приближения в каждом из столбцов с номерами больше r одна и та же, и то, что для одного столбца верно

$$\frac{\|A - CC^+A\|_2}{\sqrt{N-r}} = \left\| \begin{bmatrix} 1 - \frac{1}{1+\varepsilon^2(N-r)} & \frac{\varepsilon\sqrt{N-r}}{1+\varepsilon^2(N-r)} \\ \frac{\varepsilon\sqrt{N-r}}{1+\varepsilon^2(N-r)} & 1 - \frac{\varepsilon^2(N-r)}{1+\varepsilon^2(N-r)} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{N-r+1}} \\ \frac{\varepsilon}{\sqrt{N-r}\sqrt{N-r+1}} \end{bmatrix} \right\|_2,$$

приходим к оценке

$$\left\|A - CC^{+}A\right\|_2 \geqslant \sqrt{N-r}\left\|\begin{bmatrix} 0 \\ \frac{\varepsilon\left(\sqrt{N-r}-1/\sqrt{N-r}\right)}{\left(1+\varepsilon^2(N-r)\right)\sqrt{N-r+1}} \end{bmatrix}\right\|_2 = \varepsilon\Omega\left(\sqrt{N-r}\right).$$

Осталось заметить, что крестовое приближение, являясь частным случаем столбцового ($CUR = CW$ для $W = UR$), не может давать оценку лучше.

Таким образом, при больших N для крестовых алгоритмов, основанных на приципе максимального объема, нельзя гарантировать высокую точность получаемых приближений, поскольку может оставаться множитель порядка $\sqrt{N}$. Тем не менее, наблюдаемая на практике высокая эффективность таких алгоритмов говорит в пользу того, что примеры, подобные рассмотренному выше, встречаются редко.

3.1. Вероятностная мера

Формализуем понятие «редкости», определив RANDSVD ансамбль на матрицах с фиксированными сингулярными числами.

Определение 3.1. RANDSVD ансамблем матриц $A \sim \mathrm{RANDSVD}\,(A_0)$ называется множество матриц вида

$$A = W_L A_0 W_R,$$

где $A_0 \in \mathbb{C}^{M\times N}$ фиксирована, а $W_L \in \mathbb{C}^{M\times M}$ и $W_R \in \mathbb{C}^{N\times N}$ – случайные унитарные матрицы; с индуцированной на этом множестве вероятностной мерой с помощью мер Хаара для матриц W_L и W_R.

В соответствии с определением, ансамбль RANDSVD(Σ) получает структуру вероятностного пространства и содержит все матрицы, имеющие одну и ту же матрицу сингулярных чисел Σ. Редкие события будут определяться множествами матриц, имеющих малую вероятностную меру.

Замечание 3.1. RANDSVD ансамбль является довольно известной конструкцией в вычислительной математике. Случайную RANDSVD матрицу для некоторых распределений сингулярных чисел можно получить в Matlab с помощью вызова функции $\mathrm{gallery}(\text{'randsvd'}, \ldots)$. Кроме того, RANDSVD ансамбли используются при тестировании программного пакета LAPACK. LAPACK функция dlarge позволяет получить матрицу из данного ансамбля.

3.2. Некоторые свойства связанных с матрицами случайных величин

Для получения вероятностных оценок нам понадобится несколько утверждений для того, чтобы оценить встречающиеся в RANDSVD вероятности. Для их доказательства мы часто будем использовать следующее следствие формулы Стирлинга.

Утверждение 3.1. *Для любого действительного $n \geqslant 0$ верно двойное неравенство*

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n < \Gamma(n+1).$$

Для $n \geqslant 1$,

$$\Gamma(n+1) < 1{,}1\sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

Кроме того, $\forall n \geqslant 1, m > n$ верно неравенство

$$\frac{\Gamma(m+1)}{\Gamma(n+1)} < \sqrt{\frac{m}{n}}\frac{m^m}{n^n}e^{n-m}.$$

Кроме того, нам понадобится аналог утверждения о μ-когерентности, который доказывается в утверждении 2 из [86].

Утверждение 3.2 ([86])**.** *Пусть случайный вектор v равномерно распределен на сфере в $\mathbb{R}^n$, $n > 1$. Тогда для его произвольной фиксированной компоненты v_i выполняется неравенство*

$$\mathcal{P}\left(|v_i|^2 > \frac{\mu}{n-1}\right) \leqslant \sqrt{\frac{2}{\pi\mu}}e^{-\frac{\mu}{2}}.$$

Доказательство. Равномерно распределенный на сфере вектор можно построить, как нормированный на 1 случайный гауссовый вектор с элементами x_i, а потому

$$\mathbb{P}(|v_i|^2 > t) = \mathbb{P}\left(\frac{|x_i|^2}{\sum\limits_{j=1}^{n}|x_j|^2} > t\right) = \mathbb{P}\left(\frac{|x_i|^2}{\sum\limits_{\substack{j=1\\ j\neq i}}^{n}|x_j|^2} > \frac{t}{1-t}\right),$$

$$\mathbb{P}(|v_i|^2 > t,\ i = \overline{1,n}) \leqslant n\mathbb{P}\left(\frac{|x_1|^2}{\sum\limits_{j=2}^{n}|x_j|^2} > \frac{t}{1-t}\right).$$

Случайная величина $\frac{|x_1|^2}{\sum\limits_{j=2}^{n}|x_j|^2}$ имеет распределение Фишера со степенями свободы 1 и $n-1$. Оценим

вероятность с помощью соответствующей плотности распределения:

$$\mathbb{P}(|v_i|^2 > t) \leqslant \int\limits_{(n-1)\frac{t}{1-t}}^{\infty} \frac{\sqrt{\frac{x(n-1)^{n-1}}{(x+n-1)^n}}}{x\mathrm{B}\left(\frac{1}{2}, \frac{n-1}{2}\right)} dx = \int\limits_{(n-1)\frac{t}{1-t}}^{\infty} \frac{\sqrt{\frac{(n-1)^{n-1}}{(x+n-1)^n}}}{\sqrt{x}\mathrm{B}\left(\frac{1}{2}, \frac{n-1}{2}\right)} dx \leqslant$$

$$\leqslant /x_0 = (n-1)\frac{t}{1-t}/ \leqslant \int\limits_{x_0}^{\infty} \frac{\sqrt{x_0+n-1}\sqrt{\frac{(n-1)^{n-1}}{(x+n-1)^n}}}{\sqrt{x_0(x+n-1)}\mathrm{B}\left(\frac{1}{2}, \frac{n-1}{2}\right)} dx =$$

$$= \frac{(n-1)^{\frac{n-1}{2}}\sqrt{x_0+n-1}}{\sqrt{x_0}\mathrm{B}\left(\frac{1}{2}, \frac{n-1}{2}\right)} \int\limits_{x_0}^{\infty} \frac{dx}{(x+n-1)^{\frac{n+1}{2}}} \leqslant$$

$$\leqslant \frac{(n-1)^{\frac{n-1}{2}}\sqrt{x_0+n-1}}{\sqrt{x_0}} \frac{\sqrt{\frac{n-1}{2}}}{\sqrt{\pi}} \frac{2}{n-1}(x_0+n-1)^{-\frac{n-1}{2}} =$$

$$= \sqrt{\frac{2}{\pi}} \frac{(n-1)^{\frac{n-2}{2}}}{\sqrt{x_0}}(x_0+n-1)^{-\frac{n-2}{2}} =$$

$$= \sqrt{\frac{2}{\pi}} \frac{(n-1)^{\frac{n-2}{2}}}{\sqrt{(n-1)\frac{t}{1-t}}}\left((n-1)\frac{t}{1-t}+n-1\right)^{-\frac{n-2}{2}} =$$

$$= \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{(n-1)t}}(1-t)^{\frac{n-1}{2}} \leqslant \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{(n-1)t}} e^{-t\frac{n-1}{2}} = \sqrt{\frac{2}{\pi\mu}} e^{-\frac{\mu}{2}}.$$

□

Для распределения χ^2 нам понадобится оценка вероятности того, что значение случайной величины существенно ниже среднего.

Утверждение 3.3. *Если случайная величина x обладает распределением $\chi^2(n)$, $n > 2$, то*

$$\mathcal{P}\left(x < n - \sqrt{2cn}\right) \leqslant \frac{e^{\frac{1}{2}+\sqrt{\frac{2}{n}}}}{\sqrt{2\pi c}} e^{-\frac{c}{2}}.$$

Доказательство. $\chi^2(n)$ с ростом n достаточно быстро сходится к нормальному распределению с матожиданием n и дисперсией $2n$. Найдем точку, где отношение их плотностей вероятности максимально. Приравнивая производную отношения плотностей вероятностей к нулю, получим

$$-\frac{1}{2n}(x-n) - \frac{\frac{n}{2}-1}{x} + \frac{1}{2} = 0,$$
$$x^2 - 2nx + n(n-2) = 0,$$
$$x = n \pm \sqrt{2n}.$$

Так как в нуле отношение плотностей вероятности равно нулю, то $n - \sqrt{2n}$ – это точка максимума.

Вычислим отношение плотностей вероятностей в этой точке

$$\begin{aligned}\frac{x^{\frac{n}{2}-1}e^{-\frac{x}{2}}\sqrt{4\pi n}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)e^{-\frac{(x-n)^2}{4n}}} &= \frac{\left(n-\sqrt{2n}\right)^{\frac{n}{2}-1}e^{-\frac{n}{2}+\sqrt{\frac{n}{2}}}\sqrt{4\pi n}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)e^{-\frac{1}{2}}} \\ &\leqslant \frac{n\left(n-\sqrt{2n}\right)^{\frac{n}{2}-1}e^{-\frac{n}{2}+\sqrt{\frac{n}{2}}}}{2^{\frac{n}{2}}\left(\frac{n}{2e}\right)^{\frac{n}{2}}e^{-\frac{1}{2}}} \\ &= \left(1-\sqrt{\frac{2}{n}}\right)^{\frac{n}{2}-1}e^{\frac{1}{2}+\sqrt{\frac{n}{2}}} \\ &\leqslant e^{\frac{1}{2}+\sqrt{\frac{2}{n}}}.\end{aligned}$$

Так как в любой другой точке $x < n$ отношение не больше, получаем

$$\begin{aligned}\mathcal{P}\left(x < n-\sqrt{2cn}\right) &\leqslant e^{\frac{1}{2}+\sqrt{\frac{2}{n}}}\int\limits_{\sqrt{2cn}}^{+\infty}\frac{1}{\sqrt{4\pi n}}e^{-\frac{x^2}{4n}}dx \\ &\leqslant e^{\frac{1}{2}+\sqrt{\frac{2}{n}}}\int\limits_{c}^{+\infty}\frac{1}{\sqrt{8\pi t}}e^{-\frac{t}{2}}dt \\ &\leqslant \frac{e^{\frac{1}{2}+\sqrt{\frac{2}{n}}}}{\sqrt{2\pi c}}e^{-\frac{c}{2}}.\end{aligned}$$

□

Кроме того, нам понадобится обобщение оценки на хвост распределения Хи-квадрат из [87].

Лемма 3.1 ([87])**.** *Пусть* $X = \sum\limits_{i=1}^{N} a_i\left(X_i - 1\right)$*, где* $X_i \sim \chi^2\left(1\right)$ *– независимые случайные величины,* $a_i \geqslant 0$*. Тогда*

$$\mathcal{P}\left(X > C\max_i a_i + \sqrt{C\sum_{i=1}^{N}a_i^2}\right) \leqslant e^{-C/2}.$$

Наконец, нам нужно будет сводить сумму взвешенных зависимых одинаково распределенных случайных величин к одной случайной величине, к которой можно было бы применить неравенство Маркова. Тот же результат в более общем виде доказан в [88].

Утверждение 3.4. *Пусть* $X = \sum\limits_{i=1}^{N} a_i x_i$*, где* $a_i = const > 0$*,* x_i *- неотрицательные одинаково распределенные случайные величины. Обозначим* $S = \sum\limits_{i=1}^{N} a_i$*. Тогда*

$$\mathcal{P}\left(X > cS\right) \leqslant \inf_{p\in\mathbb{N}}\frac{\mathbb{E}\left(x_1^p\right)}{c^p}.$$

Доказательство. Воспользуемся тем фактом, что матожидание произведения максимально, когда все распределения идеально коррелированны (что можно увидеть, например, применив $p-1$ раз неравенство Гельдера), что приводит к

$$\mathbb{E}\left(x_{i_1}\cdot\ldots\cdot x_{i_p}\right)\leqslant\left(\mathbb{E}\left(x_{i_1}^p\right)\right)^{1/p}\cdot\ldots\cdot\left(\mathbb{E}\left(x_{i_p}^p\right)\right)^{1/p}=\mathbb{E}\left(x_1^p\right). \tag{3.2}$$

Применим неравенство Маркова для произвольного p. Возведя X в степень p и взяв матожидание, заменим все произведения на степени x_1, используя (3.2):

$$\begin{aligned}\frac{\mathbb{E}\left(X^p\right)}{(cS)^p}&=\frac{\sum\limits_{i_1+\ldots+i_p=p}a_1^{i_1}\ldots a_p^{i_p}\binom{p}{i_1\ i_2\ \ldots\ i_p}\mathbb{E}\left(x_1^{i_1}\ldots x_p^{i_p}\right)}{(cS)^p}\\&\leqslant\frac{\sum\limits_{i_1+\ldots+i_p=p}a_1^{i_1}\ldots a_p^{i_p}\binom{p}{i_1\ i_2\ \ldots\ i_p}\mathbb{E}\left(x_1^p\right)}{(cS)^p}=\frac{\mathbb{E}\left(x_1^p\right)}{c^p}.\end{aligned}$$

Инфимум по p приведет нас к требуемому неравенству. □

3.3. Оценки для матожидания погрешности

Теорема 3.2. *Пусть матрица $A\in\mathbb{C}^{M\times N}$ принадлежит семейству матриц вида $A=Z_0W+F_0W=Z+F,\operatorname{rank}Z=r$, где $W\in\mathbb{C}^{N\times N}$ – случайная унитарная матрица. Пусть $U_Z\Sigma_ZV_Z$ – сокращенное сингулярное разложение Z. Пусть подматрица $\hat{V}_Z\in\mathbb{C}^{r\times n}$ обладает локально максимальным объемом в V_Z. Пусть ей соответствуют столбцы $C\in\mathbb{C}^{M\times n}$ матрицы A. Тогда*

$$\mathbb{E}_W\left\|A-CW\right\|_F^2\leqslant\frac{n+1}{n-r+1}\left\|F\right\|_F^2$$

Как и в рассмотренных ранее случаях, оценка будет близка к наилучшей возможной при $Z=A_r$.

Доказательство. Для начала введем матрицу $V\in\mathbb{C}^{M\times N}$ правых сингулярных векторов Z_0W. Благодаря унитарной инвариантности распределения Хаара, получаем, что V – также случайная унитарная матрица с тем же распределением, что и W, а VW^* – фиксированная матрица правых сингулярных чисел Z. Используя V в качестве новой случайной матрицы, переобозначим

$$Z_0:=Z_0WV^*,\quad F_0:=F_0WV^*,\quad A=Z_0V+F_0V.$$

Единственное отличие от изначального условия заключается в том, что V является матрицей правых сингулярных векторов Z.

Фиксируем строки $V_Z\in\mathbb{C}^{r\times N}$ матрицы V, соответствующие ненулевым сингулярным числам Z. Теперь они (в отличие от оставшихся строк V) не являются случайными. Следуя тем же

рассуждениям, что при доказательстве оценок по спектральной норме, теорема 2.10, получаем

$$A - CW = F - F_C \hat{V}_Z^+ V_Z = F - F P_C \hat{V}_Z^+ V_Z, \tag{3.3}$$

где $P_C \in \mathbb{C}^{N \times n}$ выбирает n столбцов F, соответствующие столбцам $\hat{V}_Z$ в V_Z, V_Z – правые сингулярные векторы Z.

Сведем задачу к случаю, когда $ZF^* = 0$. Для этого разделим матрицу F на две части: параллельную V_Z и ортогональную V_Z,

$$F = F V_Z^* V_Z + F \left(I - V_Z^* V_Z\right),$$

и подставим их в (3.3):

$$\begin{aligned} A - CW &= F - F P_C \hat{V}_Z^+ V_Z \\ &= F - F \left(I - V_Z^* V_Z\right) P_C \hat{V}_Z^+ V_Z - F V_Z^* V_Z P_C \hat{V}_Z^+ V_Z \\ &= F - F \left(I - V_Z^* V_Z\right) P_C \hat{V}_Z^+ V_Z - F V_Z^* \hat{V} \hat{V}_Z^+ V_Z \\ &= F - F \left(I - V_Z^* V_Z\right) P_C \hat{V}_Z^+ V_Z - F V_Z^* V_Z = F \left(I - V_Z^* V_Z\right) \left(I - P_C \hat{V}_Z^+ V_Z\right). \end{aligned}$$

Рассмотрим матрицу U_F ненулевых левых сингулярных векторов матрицы $F \left(I - V_Z^* V_Z\right)$. V на нее не влияет, поэтому эта матрица не является случайной. Благодаря унитарной инвариантности, умножение на нее слева никак не изменит норму Фробениуса $A - CW$. Зато теперь мы можем обозначить M как число строк в U_F и обозначить

$$F' = U_F^* F \left(I - V_Z^* V_Z\right),$$

что сделает F невырожденной и при этом удовлетворяющей $ZF^* = 0$. Кроме того, мы можем переобозначить и оставшиеся $N - r$ строк V через $V_F \in \mathbb{C}^{(N-r) \times N}$ и ввести матрицу сингулярных чисел F' как $\Sigma'_F \in \mathbb{C}^{M \times (N-r)}$, что приведет к оценке

$$\|A - CW\|_F = \left\|\Sigma'_F V_F \left(I - P_C \hat{V}_Z^+ V_Z\right)\right\|_F. \tag{3.4}$$

Так как проекция на U_F^* не увеличила сингулярных чисел, норма Фробениуса матрицы Σ'_F не возросла по сравнению с условием теоремы:

$$\left\|\Sigma'_F\right\|_F \leqslant \|F\|_F. \tag{3.5}$$

Используя матричную теорему Пифагора для разности в (3.4) (используя ортогональность строк), получаем

$$\|A - CW\|_F^2 = \left\|\Sigma'_F\right\|_F^2 + \left\|\Sigma'_F V_F P_C \hat{V}_Z^+ V_Z\right\|_F^2. \tag{3.6}$$

Введем случайную перестановку $\Pi \in \mathbb{C}^{(N-r) \times (N-r)}$ и заметим, что распределение $\Pi^{-1} V_F$ совпадает с распределением V_F, а потому можно переобозначить $V_F := \Pi^{-1} V_F$ и далее усреднить (3.6) только по перестановкам Π:

$$\mathbb{E}_W \|A - CW\|_F^2 = \mathbb{E}_V \|A - CW\|_F^2 = \left\|\Sigma'_F\right\|_F^2 + \mathbb{E}_\Pi \left\|\Sigma'_F \Pi V_F P_C \hat{V}_Z^+ V_Z\right\|_F^2.$$

Вычислим норму Фробениуса построчно: вероятность, что в k-й строке будет находиться произвольное j-е сингулярное число матрицы Σ'_F равна $1/(N-r)$. Таким образом, квадраты норм всех строк в среднем совпадают. Отсюда получаем

$$\mathbb{E}_\Pi \left\|\Sigma'_F \Pi V_F P_C \hat{V}_Z^+ V_Z\right\|_F^2 = \frac{1}{M} \sum_{k=1}^{N-r} \sigma_k^2\left(\Sigma'_F\right) \left\|V_F P_C \hat{V}_Z^+ V_Z\right\|_F^2 = \frac{1}{N-r} \left\|\Sigma'_F\right\|_F^2 \left\|\hat{V}_Z^+ V_Z\right\|_F^2 .$$

В итоге

$$\mathbb{E}_W \|A - CW\|_F^2 = \left\|\Sigma'_F\right\|_F^2 \left(1 + \frac{\left\|\hat{V}_Z^+ V_Z\right\|_F^2}{N-r}\right) \leqslant \|F\|_F^2 \left(1 + \frac{\left\|\hat{V}_Z^+ V_Z\right\|_F^2}{N-r}\right) \leqslant \frac{n+1}{n-r+1} \|F\|_F^2 , \tag{3.7}$$

где мы воспользовались сначала (3.5), а затем леммой 1.3. □

Замечание 3.2. Так как матрица Z в доказательстве выше была фиксирована (после фиксирования V_Z), та же оценка справедлива для $A = Z_0 + F_0 W$, где только матрица F является случайной.

В данной оценке подматрица локально максимального объема по сути ищется в матрице Z. Хотя локальная максимальность объема подматрицы в Z не гарантирует локальной максимальности объема в самой матрице A, в частном случае, когда $Z \to \infty$ (так что относительная погрешность сколь угодно мала), этого достаточно, чтобы с некоторого момента отношение объемов подматриц почти наверное определялось матрицей Z, а потому при достаточно малой погрешности подматрица локально максимального объема в A окажется и подматрицей локально максимального объема в Z. Если же погрешность не столь мала, то все равно можно часто рассчитывать на ρ-локальную максимальность объема в Z.

Рассуждения, аналогичные теореме 3.2, также позволяют показать, что полученные в разделе 2.3.2 оценки снизу по норме Фробениуса с помощью t_F-функции совпадают с оценками сверху «в среднем». А именно, при усреднении по правым сингулярным векторам получаем, что наилучшая погрешность в среднем не выше нижней оценки (2.84).

Теорема 3.3. *Пусть $A \in \mathbb{C}^{M\times N}$. Пусть матрица правых сингулярных векторов $V \in \mathbb{C}^{N\times N}$ матрицы A является случайной унитарной матрицей из распределения Хаара. Тогда*

$$\mathop{\mathbb{E}}_{\substack{V\in\mathbb{C}^{N\times N},\\ A=U\Sigma V}} \frac{\|A - (CC^+A)_r\|_F^2}{\|A - A_r\|_F^2} \leqslant 1 + \frac{t_F^2(r,n,N) - r}{N-r}.$$

Доказательство. Из полученного нами в теореме 3.2 уравнения (3.7) следует, что

$$\mathop{\mathbb{E}}_V \min_C \|A - CW\|_F^2 \leqslant \|A - A_r\|_F^2 \left(1 + \frac{\min\limits_{\hat{V}\in\mathbb{C}^{r\times n}} \left\|\hat{V}^+\right\|_F^2 - r}{N-r}\right)$$

для некоторой матрицы $W \in \mathbb{C}^{n\times N}$ ранга r и подматрицы $\hat{V}$, которая ранее соответствовала подматрице локально максимального объема в правых сингулярных векторах Z. Используя определение 2.4 t_F-функции, получаем

$$\mathop{\mathbb{E}}_{V}\min_{C}\left\|A-\left(CC^{+}A\right)_r\right\|_F^2 \leqslant \mathop{\mathbb{E}}_{V}\min_{C}\|A-CW\|_F^2 \leqslant \|A-A_r\|_F^2\left(1+\frac{t_F^2(r,n,N)-r}{N-r}\right).$$

□

Таким образом, на самом деле не обязательно выбирать подматрицу $\hat{V}$ из принципа (локально) максимального объема: достаточно использовать любой алгоритм, позволяющий ограничить $\left\|\hat{V}^{+}\right\|_F$. Другие методы, позволяющие это сделать, рассмотрены в подразделе 4.6. В частности, возможно, что на практике подобная минимизация величины $\left\|\hat{A}^{+}R\right\|_F$ может привести к точности выше, чем у подматриц локально максимального объема. Однако, стоит отметить, что в этом случае такие подматрицы тяжелее искать: при переходе в новые строки значение $\left\|\hat{A}^{+}R\right\|_F$ может возрасти, тогда как алгоритмы поиска локально максимального объема гарантируют рост объема на каждом шаге, а потому всегда останавливаются. Такие алгоритмы подробно рассматриваются в разделе 4.

Благодаря тому, что мы использовали в качестве основы произвольное приближение Z ранга r, используя теорему 3.2, легко получить оценку для крестовой аппроксимации.

Теорема 3.4. *Пусть $A = Z + F = U\left(Z_0 + F_0\right)V \in \mathbb{C}^{M\times N}$ принадлежит RANDSVD ансамблю (определение 3.1). Пусть столбцы $C \in \mathbb{C}^{M\times n}$ матрицы A соответствуют столбцам $Z_C \in \mathbb{C}^{M\times n}$ локально максимального r-проективного объема в матрице Z,* $\operatorname{rank} Z = r$. *Пусть подматрица $\hat{A}P \in \mathbb{C}^{m\times n}$ матрицы CP обладает локально максимальным проективным объемом, где $P = Z_C^{+}Z_C$. Пусть её строкам соответствуют строки $R \in \mathbb{C}^{m\times N}$ матрицы A. Тогда*

$$\mathbb{E}_A\left\|A - C\left(\hat{A}P\right)^{+}R\right\|_F^2 \leqslant \frac{m+1}{m-r+1}\cdot\frac{n+1}{n-r+1}\|F\|_F^2. \tag{3.8}$$

Видно, что реальный коэффициент погрешности при $m = n \gg r$ в среднем около $1 + r/n$, что существенно меньше упомянутой ранее оценки теоремы 3.1.

Доказательство. Прежде всего заметим, что столбцовое приближение теоремы 3.2 можно построить на основе Z_C вместо $\hat{V}_Z$. Действительно,

$$CZ_C^{+}Z = C\left(U_Z\Sigma_Z\hat{V}_Z\right)^{+}U_Z\Sigma_Z V_Z = C\hat{V}_Z^{+}\hat{V}_Z. \tag{3.9}$$

Поэтому, согласно теореме 3.2, получаем

$$\mathbb{E}_V\left\|A - CZ_C^{+}Z\right\|_F^2 \leqslant \frac{n+1}{n-r+1}\|F\|_F^2. \tag{3.10}$$

При этом $\mathcal{V}_r(Z_C) = \mathcal{V}(\Sigma_Z)\,\mathcal{V}\left(\hat{V}_Z\right)$, то есть максимальность проективного объема столбцов Z_C определяется максимальностью объема $\hat{V}_Z$, поскольку ему пропорциональна. Заметим также, что вместо столбцов локально максимального проективного объема мы могли бы искать подматрицу локально максимального проективного объема в произвольных строках ранга r матрицы Z, и получить равенство точно так же, как в (3.9).

Теперь обозначим $\Phi = CZ_C^+Z$ и $E = A - \Phi$. Если $\operatorname{rank}\Phi < r$, то можно далее строить аппроксимацию ранга $\operatorname{rank}\Phi$, и коэффициент в оценке (3.8) будет меньше (в этом случае итоговый проектор P будет иметь ранг меньше r). Поэтому далее считаем худший случай $\operatorname{rank}\Phi = r$. На основе Φ построим строковую аппроксимацию (что есть просто столбцовая аппроксимация для транспонированной матрицы, поэтому все оценки остаются верными). При этом будем искать подматрицу локально максимального проективного объема в столбцах, соответствующих столбцам C: легко проверить, что если $\operatorname{rank}\Phi = r$, то и ранг её столбцов, соответствующих C, также равен r. В этом случае получим строковое приближение вида

$$\Phi_C\hat{\Phi}^+R = \left(CZ_C^+Z_C\right)\left(\hat{A}Z_CZ_C^+\right)^+R = CP\left(\hat{A}P\right)^+R = C\left(\hat{A}P\right)^+R. \tag{3.11}$$

При этом матрица A представима в виде $A = U\Phi_0 + UE_0$, где

$$\Phi_0 = (U^*C)\,Z_C^+Z = (U^*C)\,(UZ_{VC})^+\,UZ_V = C_VZ_{VC}^+Z_V,$$

C_V – соответствующие C столбцы $A_V = Z_0V + F_0V = U^*A$, а Z_{VC} – соответствующие C столбцы $Z_V = Z_0V$. Так как $A = UA_V$, она удовлетворяет условию теоремы 3.2 (с точностью до транспонирования), а потому с учетом (3.11) получаем оценку

$$\begin{aligned}
\mathbb{E}_U\left\|A - C\left(\hat{A}P\right)^+R\right\|_F^2 &= \mathbb{E}_U\left\|A - \Phi_C\hat{\Phi}^+R\right\|_F^2 \\
&\leqslant \frac{m+1}{m-r+1}\,\|E\|_F^2 \\
&= \frac{m+1}{m-r+1}\,\|A-\Phi\|_F^2 \\
&= \frac{m+1}{m-r+1}\left\|A - CZ_C^+Z\right\|_F^2.
\end{aligned}$$

Взяв затем матожидание по V, с учетом (3.10) получаем (3.8). □

Заметим, что для получения вероятностных оценок важна независимость матриц левых и правых сингулярных векторов. Поэтому при использовании подматриц локально максимального объема (особенно главных, то есть лежащих на диагонали) для симметричных и кососимметричных матриц погрешность по норме Фробениуса может оказаться существенно выше.

3.4. Оценки для вероятности отличия погрешности от матожидания

Из теоремы 3.4 следует, что, например, для 90% матриц (в смысле RANDSVD распределения) принцип максимального (проективного) объема может быть использован для построения кре-

стовой аппроксимации с коэффициентом ошибки не более $\sqrt{10\left(1+\frac{r}{m-r+1}\right)\left(1+\frac{r}{n-r+1}\right)}$. Данная оценка существенно выше матожидания и не гарантирует точность аппроксимации, которая была бы сколь угодно близкой к сингулярному разложению. В то же время, как мы увидим далее в численных экспериментах главы 6, отличие от матожидания почти всегда мало.

Возникает гипотеза о том, что матрицы, для которых принцип максимального (проективного) объема не позволяет построить точных крестовых или столбцовых аппроксимаций крайне редки. В данном разделе мы формализуем эту гипотезу, доказав, что вероятность превысить коэффициент $1+\frac{cr}{n-r+1}$ в крестовой аппроксимации экспоненциально мала для матриц из RANDSVD ансамбля, вне зависимости от распределения их сингулярных чисел.

В дальнейшем в этом разделе мы остановимся на действительном случае. Хотя для матожиданий оценки в комплексном случае никак не изменятся, оценки на вероятности в комплексном случае будут использовать случайные величины с в два раза большим числом параметров (например, $\chi^2(2n)$ вместо $\chi^2(n)$). В итоге вероятность «успеха» в действительном случае окажется меньше (из-за большей относительной дисперсии), а потому мы рассмотрим здесь именно его.

Теорема 3.5. *Пусть матрица $A \in \mathbb{R}^{M\times N}$, представима в виде $A = Z_0 + F_0$, $\operatorname{rank} Z_0 = r$. Рассмотрим семейство матриц $A_V = Z + F = Z_0V + F_0V$, где V – случайная ортогональная матрица. Тогда для CW аппроксимация матрицы A_V с матрицей $W = \hat{Z}_{11}^{+}\begin{bmatrix} Z_{11} & Z_{12}\end{bmatrix}$, где $Z_{11} \in \mathbb{R}^{r\times n}$ - подматрица максимального объема в произвольных линейно независимых строках $\begin{bmatrix} Z_{11} & Z_{12}\end{bmatrix}$ из Z верно*

$$\mathcal{P}\left(\|A_V - CW\|_F^2 > \left(1+\frac{Cr}{n-r+1}\cdot\frac{N-n}{N-r-1}\right)\|F\|_F^2\right) \leqslant \frac{1{,}6C}{C-2}e^{-\frac{C}{2}}, \quad C > 2. \tag{3.12}$$

Доказательство. В теореме 3.2, уравнение (3.6), показано, что

$$\|A_V - CW\|_F^2 = \left\|\Sigma'_F\right\|_F^2 + \left\|\Sigma'_F V_F P_C \hat{V}_Z^{+} V_Z\right\|_F^2 = \left\|\Sigma'_F\right\|_F^2 + \left\|\Sigma'_F V_F P_C \hat{V}_Z^{+}\right\|_F^2, \tag{3.13}$$

где $\left\|\Sigma'_F\right\|_F \leqslant \|F\|_F$, а $V_F \in \mathbb{R}^{(N-r)\times N}$ содержит случайные ортонормированные строки, ортогональные строкам $V_Z \in \mathbb{R}^{r\times N}$. Их распределение не поменяется, если их домножить слева на случайную ортогональную матрицу $W_F \in \mathbb{R}^{(N-r)\times(N-r)}$, что мы и сделаем. При этом

$$\left\|V_F P_C \hat{V}_Z^{+}\right\|_F^2 = \left\|\hat{V}_Z^{+}\right\|_F^2 - \left\|\hat{V}_Z \hat{V}_Z^{+}\right\|_F^2 = \left\|\hat{V}_Z^{+}\right\|_F^2 - r \leqslant r\frac{N-n}{n-r+1},$$

согласно следствию 1.6.

Таким образом, если

$$V_F P_C \hat{V}_Z^{+} = U_{VZ}\Sigma_{VZ}V_{VZ},\ U_{VZ} \in \mathbb{R}^{(N-r)\times\min(r,N-r)},\ \Sigma_{VZ} \in \mathbb{R}^{\min(r,N-r)\times\min(r,N-r)},\ V_{VZ} \in \mathbb{R}^{\min(r,N-r)\times r}, \tag{3.14}$$

то

$$\|\Sigma_{VZ}\|_F^2 \leqslant r\frac{N-n}{n-r+1}.$$

Используя ортогональную инвариантность распределения Хаара, получаем, что матрица $U' = V_F U_{VZ} \in \mathbb{R}^{(N-r)\times\min(r,N-r)}$ является случайной матрицей с ортонормированными столбцами, также с распределением Хаара. Используя ортогональную инвариантность нормы Фробениуса, получим после подстановки (3.14) в (3.13)

$$\|A_V - CW\|_F^2 \leqslant \|F\|_F^2 + \left\|\Sigma'_F U' \Sigma_{VZ}\right\|_F^2. \tag{3.15}$$

В матрице $\Sigma'_F U' \Sigma_{VZ}$ всего не более $r(N-r)$ ненулевых элементов, каждый из которых является элементом U' с некоторым весом. Квадрат нормы Фробениуса, таким образом, является взвешенной суммой одинаково распределенных случайных величин, а потому мы можем применить утверждение 3.4. Пусть u – произвольный элемент матрицы U'. Пусть $X = \left\|\Sigma'_F U' \Sigma_{VZ}\right\|_F^2$, а $S = \sum\limits_{ij} \left|\Sigma'_F\right|_{ii}^2 |\Sigma_{VZ}|_{jj}^2 \leqslant \|F\|_F^2 \|\Sigma_{VZ}\|_F^2$. Тогда согласно утверждению 3.4

$$\mathcal{P}\left(X > cS\right) \leqslant \inf_{p\in\mathbb{N}} \frac{\mathbb{E}\left(|u|^{2p}\right)}{c^p}. \tag{3.16}$$

Теперь вычислим матожидание в (3.16), используя утверждение 3.2 для $v_i = u$.

$$\begin{aligned}
\mathcal{P}\left(X > \frac{cS}{N-r-1}\right) &\leqslant \inf_{p\in\mathbb{N}} (N-r-1)^p \frac{\mathbb{E}\left(|u|^{2p}\right)}{c^p} = p\,(N-r-1)^p \frac{\int\limits_0^{+\infty} x^{p-1}\mathcal{P}\left(|u|^2 > x\right)}{c^p} \\
&\leqslant \frac{p\sqrt{\frac{2}{\pi}}\frac{(N-r-1)^p}{\sqrt{N-r-1}} \int\limits_0^{+\infty} x^{p-1} \frac{e^{-\frac{(N-r-1)x}{2}}}{\sqrt{x}} dx}{c^p} = \frac{p2^p \int\limits_0^{+\infty} t^{p-1}\frac{e^{-t}}{\sqrt{t}} dt}{\sqrt{\pi}c^p} \\
&= \frac{p2^p\Gamma\left(p-1/2\right)}{\sqrt{\pi}c^p} \leqslant \frac{1{,}1p2^p\sqrt{2\pi(p-1)}}{\sqrt{\pi}c^p\sqrt{p-1}}\left(\frac{p-1}{e}\right)^{p-1} \\
&\leqslant 1{,}1\sqrt{2}e\frac{p}{p-1}\left(\frac{2(p-1)}{ec}\right)^p.
\end{aligned}$$

Между $c/2$ и $c/2+1$ есть целое значение p. Для него $\frac{p}{p-1} \leqslant \frac{c/2}{c/2-1} = \frac{c}{c-2}$, так как $p/(p-1)$ монотонно убывает. С другой стороны, для него же $\left(\frac{2(p-1)}{ec}\right)^p \leqslant \left(\frac{2(c/2)}{ec}\right)^{c/2+1}$, так как это выражение монотонно возрастает при $p > c/2 > 1$:

$$\left(\frac{2(p-1)}{ec}\right)^p = \left(\frac{p-1}{p}\right)^p \left(\frac{2p}{ec}\right)^p.$$

Левый множитель растет при $p > 1$, правый множитель имеет глобальный минимум при $p = c/2$ и растет при $p > c/2$.

В итоге, подставляя соответствующие p, получаем

$$\mathcal{P}\left(\left\|\Sigma'_F U' \Sigma_{VZ}\right\|_F^2 > \frac{cS}{N-r-1}\right) \leqslant \frac{1{,}6c}{c-2} e^{-\frac{c}{2}}. \tag{3.17}$$

В нашем случае

$$S \leqslant \|F\|_F^2 \|\Sigma_{VZ}\|_F^2 \leqslant \frac{r(N-n)}{n-r+1}\|F\|_F^2.$$

Подстановка S в (3.17) даст с учетом (3.15) требуемую оценку (3.12). □

Замечание 3.3. При $N = r + 1$ усреднение не требуется, чтобы оценка не превышала случая $C = 1$.

Следствие 3.1. Утверждение теоремы сохраняется, когда случайным является только шум: $A = Z + F_0 W_R$. Для этого случая

$$\|A - CW\|_F^2 = \|F\,(I - V^*V)\,\|_F^2 + \|F\,(I - V^*V)\,P_C\hat{V}^+\|_F^2, \tag{3.18}$$

где $V \in \mathbb{R}^{r\times N}$ – матрица правых сингулярных векторов матрицы Z.

Далее, рассмотрим полные сингулярные разложения

$$F_0 = U_F\Sigma_F V_F, \quad U_F \in \mathbb{R}^{M\times M},\ \Sigma_F \in \mathbb{R}^{M\times N}, \quad V_F \in \mathbb{R}^{N\times N}. \tag{3.19}$$

и

$$(I - V^*V)\,P_C\hat{V}^+ = U_V\Sigma_V V_V, \quad U_V \in \mathbb{R}^{N\times N},\ \Sigma_V \in \mathbb{R}^{N\times r}, \quad V_V \in \mathbb{R}^{r\times r}. \tag{3.20}$$

Для последнего справедлива оценка

$$\begin{aligned}\|\Sigma_V\|_F^2 &= \|\,(I - V^*V)\,P_C\hat{V}^+\|_F^2 = \|P_C\hat{V}^+\|_F^2 - \|VP_C\hat{V}^+\|_F^2 = \|\hat{V}^+\|_F^2 - \|\hat{V}\hat{V}^+\|_F^2 \\ &\leqslant r + r\frac{N-n}{n-r+1} - r = r\frac{N-n}{n-r+1}.\end{aligned}$$

Подставим сингулярные разложения (3.19) и (3.20) в (3.18) и учтем, что матрица $U' = V_F W_R U_V \in \mathbb{R}^{N\times r}$ является случайной матрицей с ортонормированными столбцами. Получим

$$\|A - CW\|_F^2 \leqslant \|F\|_F^2 + \|\Sigma_F U'\Sigma_V\|_F^2, \tag{3.21}$$

Отличие (3.21) от (3.15) заключается только в том, что размер случайной матрицы увеличился с $(N - r)\times r$ до $N \times r$. Так как больше ничего в доказательстве не меняется, в итоге получаем оценку

$$\mathcal{P}\left(\|A - CW\|_F^2 > \left(1 + \frac{Cr}{n-r+1}\cdot\frac{N-n}{N-1}\right)\|F\|_F^2\right) \leqslant \frac{1{,}6C}{C-2}e^{-\frac{C}{2}}.$$

Наконец, мы готовы доказать основной результат для крестовой аппроксимации. В целом, оно повторяет аналогичное доказательство для матожидания.

Сразу заметим, что если некоторая подматрица $Z_{11} \in \mathbb{R}^{m\times n}$ обладает максимальным r-проективным объемом в строках $[Z_{11}\ Z_{12}]$, то то же касается и любой её невырожденной подматрицы $Z'_{11} \in \mathbb{R}^{r\times n}$ в строках $\left[Z'_{11}\ Z'_{12}\right]$. Действительно, пусть $\left[Z'_{11}\ Z'_{12}\right] = P'\,[Z_{11}\ Z_{12}]$, где $P' \in \mathbb{R}^{r\times n}$ зануляет неиспользуемые строки и пусть $U \in \mathbb{R}^{r\times n}$ содержит r левых сингулярных векторов $[Z_{11}\ Z_{12}]$. Тогда поскольку $\mathcal{V}_r\,(Z_{11}) = \mathcal{V}\,(UZ_{11})$ максимально, то и $\mathcal{V}\,(P'Z_{11}) =$

$\mathcal{V}\left(P'U^TUZ_{11}\right) = \mathcal{V}\left(P'U^T\right)\mathcal{V}\left(UZ_{11}\right)$ максимально, так как первый множитель от выбранных столбцов не зависит. Поэтому далее мы можем использовать теорему 3.5 для подматриц в Z с числом строк $m \geqslant r$.

Теорема 3.6. *Пусть $A = Z + F = W_L\left(Z_0 + F_0\right)W_R$, $A \in \mathbb{C}^{M\times N}$, $\operatorname{rank} Z = r$. Пусть W_L и W_r – независимые друг от друга случайные ортогональные матрицы. Пусть столбцы $C \in \mathbb{R}^{M\times n}$ соответствуют подматрце $\hat{Z}_{11} \in \mathbb{R}^{m\times n}$ максимального r-проективного объема в Z, а строки R – подматрице максимального проективного объема в столбцах CP_Z, где P_Z – проектор на первые n столбцов Z. Тогда*

$$\mathcal{P}\left(\|A - C\left(\hat{A}P_Z\right)R\|_F^2 > \left(1 + \frac{Cr}{m-r+1}\cdot\frac{N-m}{N-r-1}\right)\left(1 + \frac{Cr}{n-r+1}\cdot\frac{N-n}{N-r-1}\right)\|F\|_F^2\right) \leqslant$$
$$\leqslant \frac{3{,}2C}{C-2}e^{-\frac{C}{2}}, \quad C > 2.$$

Доказательство. Как и в теореме 3.5, введем матрицу $\Phi = CW = CZ_{11}^{+}\left[Z_{11}\ Z_{12}\right]$. Введем матрицу $E = A - \Phi$. Тогда для каждого W_R, мы можем построить столбцовое приближение

$$R^TW_\Phi = R^T\hat{\Phi}_{11}^{T+}\left[\Phi_{11}^T\ \Phi_{21}^T\right]$$

матрицы A^T (строки R матрицы A соответствуют столбцам R^T матрицы A^T), которая умножается справа на случайную ортогональную матрицу W_L^T. Пусть $\Phi_{11}^T \in \mathbb{R}^{r\times m}$ соответствуют подматрице максимального объема в некоторых невырожденных строках $\left[\Phi_{11}^T\ \Phi_{21}^T\right] \in \mathbb{R}^{r\times M}$ матрицы Φ^T.

Тогда согласно теореме 3.5

$$\mathcal{P}\left(\|A^T - R^T\hat{\Phi}_{11}^{T+}\left[\Phi_{11}^T\Phi_{21}^T\right]\|_F^2 > \left(1 + \frac{Cr}{m-r+1}\cdot\frac{N-m}{N-r-1}\right)\|E\|_F^2\right) \leqslant \frac{1{,}6C}{C-2}e^{-\frac{C}{2}}, \quad C > 2. \tag{3.22}$$

Кроме того, согласно той же теореме можно оценить норму $E = A - \Phi = A - CW$:

$$\mathcal{P}\left(\|E\|_F^2 > \left(1 + \frac{Cr}{n-r+1}\cdot\frac{N-n}{N-r-1}\right)\|F\|_F^2\right) \leqslant \frac{1{,}6C}{C-2}e^{-\frac{C}{2}}, \quad C > 2.$$

Объединяя эти вероятности и транспонируя матрицы в (3.22), получаем

$$\mathcal{P}\left(\|A - \begin{bmatrix}\Phi_{11}\\ \Phi_{21}\end{bmatrix}\Phi_{11}^{+}R\|_F^2 > \left(1 + \frac{Cr}{m-r+1}\cdot\frac{N-m}{N-r-1}\right)\left(1 + \frac{Cr}{n-r+1}\cdot\frac{N-n}{N-r-1}\right)\|F\|_F^2\right) \tag{3.23}$$
$$\leqslant \frac{3{,}2C}{C-2}e^{-\frac{C}{2}}, \quad C > 2.$$

Теперь рассмотрим, что из себя представляет $\begin{bmatrix}\Phi_{11}\\ \Phi_{21}\end{bmatrix}\Phi_{11}^{+}R$. Запишем сокращенное сингулярное разложение

$$\Phi = U_\Phi\Sigma_\Phi V_\Phi, \quad U_\Phi \in \mathbb{R}^{M\times r}, \quad \Sigma_\Phi \in \mathbb{R}^{r\times r}, \quad V_\Phi \in \mathbb{R}^{r\times N}.$$

Тогда, если Φ_{11} соответствует подматрица $\hat{U}_\Phi \in \mathbb{R}^{m\times r}$ матрицы U_Φ и подматрица $\hat{V}_\Phi \in \mathbb{R}^{r\times r}$ матрицы V_Φ,

$$\begin{bmatrix} \Phi_{11} \\ \Phi_{21} \end{bmatrix} \Phi_{11}^+ = U_\Phi \Sigma_\Phi \hat{V}_\Phi \left(\hat{U}_\Phi \Sigma_\Phi \hat{V}_\Phi\right)^+ = U_\Phi \hat{U}_\Phi^+.$$

Аналогично,

$$Z_{11}^+ \left[Z_{11} Z_{12}\right] = \hat{V}^+ V,$$

где V взята из сокращенного сингулярного разложения

$$\Phi = U\Sigma V, \quad U \in \mathbb{R}^{M\times r}, \quad \Sigma \in \mathbb{R}^{r\times r}, \quad V \in \mathbb{R}^{r\times N}.$$

Значит, мы можем записать $\Phi = C\hat{V}^+ V$. При этом строки $[\Phi_{11}\ \Phi_{12}] \in \mathbb{R}^{m\times N}$ соответствуют подматрице $\hat{A}\hat{V}^+ V$, где $\hat{A} \in \mathbb{R}^{m\times n}$ – подматрица матрицы A.

В итоге,

$$\begin{aligned} \begin{bmatrix} \Phi_{11} \\ \Phi_{21} \end{bmatrix} \Phi_{11}^+ R &= U_\Phi \hat{U}_\Phi^+ R \\ &= U_\Phi \Sigma_\Phi V_\Phi \left(\hat{U}_\Phi \Sigma_\Phi V_\Phi\right)^+ R \\ &= \Phi \left([\Phi_{11}\ \Phi_{12}]\right)^+ R = C\hat{V}^+ V \left(\hat{A}\hat{V}^+ V\right)^+ R \\ &= C\hat{V}^+\hat{V} \left(\hat{A}\hat{V}^+\hat{V}\right)^+ R = C\left(\hat{A}P_Z\right)^+ R, \end{aligned}$$

где $P_Z = \hat{V}^+\hat{V}$ – ортопроектор на первые n столбцов матрицы Z. Подставив $C\left(\hat{A}P_Z\right)^+ R$ в (3.23), получаем утверждение теоремы. □

Замечание 3.4. Ранг итоговой аппроксимации может быть меньше r, так как нет гарантий, что $\operatorname{rank}\left(\hat{A}\hat{V}^+ V\right) \geqslant r$ или даже что ранг исходной матрицы хотя бы r. В случае, если $\operatorname{rank}\Phi < r$ при этом в теореме выбирается подматрица максимального $\operatorname{rank}\Phi$-проективного объема в столбцах CP_Z, что при $\operatorname{rank}\Phi < r$ только уменьшит коэффициент погрешности.

Следствие 3.2. При $Z = A_r$ получаем $F = A - A_r$ и крестовую аппроксимацию, сколь угодно близкую к точности сингулярного разложения при достаточно большом числе строк m и столбцов n.

3.4.1. Оценки для случая $\|F\|_2 \ll \|F\|_F$

Теоремы 3.5 и 3.6 дают завышенную вероятность ошибки, когда сингулярные числа F распределены более равномерно. В случае, когда $\|F\|_2 = \|F\|_F/\sqrt{N-r}$, легко показать, что вероятность превысить матожидание равна 0. Следующая теорема позволяет улучшить результат теоремы 3.5, когда сингулярные числа F убывают медленно.

Теорема 3.7. *В условиях теоремы 3.5*

$$\mathcal{P}\left(\|A_V - CW\|_F^2 > \left(1 + \frac{r\left(1 + C\frac{\|F\|_2^2}{\|F\|_F^2} + \sqrt{C}\frac{S}{\|F\|_F^2}\right)}{n-r+1} \cdot \frac{N-n}{N-r-\sqrt{2c(N-r)}}\right)\|F\|_F^2\right) \leqslant 1{,}1\sqrt{\pi C}\frac{C+2}{C}e^{-C} + \frac{e^{\frac{1}{2}+\sqrt{\frac{2}{N-r}}}}{\sqrt{4\pi c}}e^{-\frac{c}{2}+\ln r}, \quad S = \sqrt{\sum_{k=1}^{N-r}\sigma_k^4(F)}, \quad C \geqslant 2. \tag{3.24}$$

Доказательство. В доказательстве теоремы 3.5 мы получили формулу (3.15), норму Фробениуса в которой выпишем напрямую через сумму:

$$\|A_V - CW\|_F^2 \leqslant \|F\|_F^2 + \sum_{j=1}^{r}\sigma_j^2(\Sigma_{VZ})\sum_{i=1}^{N-r}\sigma_i^2(F)\left|U'_{ij}\right|^2,$$

где мы явно воспользовались тем фактом, что каждое сингулярное число Σ'_F не больше соответствующего сингулярного числа F.

Вычтем $1/(N-r-\sqrt{2c(N-r)})$ из каждого значения $\left|U'_{ij}\right|^2$, где c будет определено позднее:

$$\begin{aligned}\|A_V - CW\|_F^2 \leqslant \|F\|_F^2 &+ \frac{\|F\|_F^2\|\Sigma_{VZ}\|_F^2}{N-r-\sqrt{2c(N-r)}} \\ &+ \sum_{j=1}^{r}\sigma_j^2(\Sigma_{VZ})\sum_{i=1}^{N-r}\sigma_i^2(F)\left(\left|U'_{ij}\right|^2 - \frac{1}{N-r-\sqrt{2c(N-r)}}\right).\end{aligned} \tag{3.25}$$

Случайные величины $\left|U'_{ij}\right|^2$ можно ввести через

$$\left|U'_{ij}\right|^2 = \frac{Z_{ij}}{\sum\limits_{k=1}^{N-r}Z_{kj}}, \tag{3.26}$$

где случайные величины Z_{ij} независимы для разных k при одном и том же j и имеют Хи-квадрат распределение $Z_{ij} \sim \chi^2(1)$.

Среди всего пространства событий рассмотрим только те, при которых $\sum\limits_{k=1}^{N-r}Z_{kj} > N-r-\sqrt{2c(N-r)}$ для всех j. Согласно утверждению 3.3 его мощность в вероятностном пространстве не меньше

$$\mathcal{P}_c = 1 - r\frac{e^{\frac{1}{2}+\sqrt{\frac{2}{N-r}}}}{\sqrt{2\pi c}}e^{-c/2}. \tag{3.27}$$

Далее будем работать в соответствующем подмножестве вероятностного пространства. Заменим $\left|U'_{kj}\right|^2$ случайными величинами $Z_{ij} = \left|U'_{ij}\right|^2\sum\limits_{k=1}^{N-r}Z_{kj} \geqslant \left|U'_{ij}\right|^2\left(N-r-\sqrt{2c(N-r)}\right)$. Таким образом,

в (3.25) получаем

$$\|A_V - CW\|_F^2 \leqslant \|F\|_F^2 + \frac{\|F\|_F^2 \|\Sigma_{VZ}\|_F^2}{N - r - \sqrt{2c(N-r)}} + \frac{1}{N - r - \sqrt{2c(N-r)}} \sum_{j=1}^{r} \sigma_j^2 (\Sigma_{VZ}) \sum_{i=1}^{N} \sigma_i^2 (F) \left(Z_{ij} - 1\right). \tag{3.28}$$

Для каждого фиксированного j рассмотрим в правой части сумму

$$X_j = \sum_{i=1}^{N} \sigma_i^2 (F) \left(Z_{ij} - 1\right).$$

Обозначим $a_i = \sigma_i^2 (F)$, $S = \sqrt{\sum\limits_{i=1}^{N-r} \sigma_i^4 (F)}$ и применим лемму 3.1:

$$\mathcal{P}\left(X_j > C\|F\|_2^2 + \sqrt{C}S\right) \leqslant e^{-C/2}. \tag{3.29}$$

С помощью выражения (3.29) определим случайные величины $Y_j = C\|F\|_2^2 + \sqrt{C}S > X_j$ с функцией распределения

$$\mathcal{P}\left(Y_j < x\right) = 1 - e^{-x/2}.$$

Воспользовавшись тем фактом, что $Y_j > X_j$, подставим их в (3.28):

$$\|A_V - CW\|_F^2 \leqslant \|F\|_F^2 + \frac{\|F\|_F^2 \|\Sigma_{VZ}\|_F^2}{N - r - \sqrt{2c(N-r)}} + \frac{1}{N - r - \sqrt{2c(N-r)}} \sum_{j=1}^{r} \sigma_j^2 (\Sigma_{VZ}) Y_j. \tag{3.30}$$

В правой части (3.30) присутствует взвешенная сумма одинаково распределенных случайных величин Y_j

$$X = \sum_{j=1}^{r} \frac{\sigma_j^2 (\Sigma_{VZ})}{N - r - \sqrt{2c(N-r)}} Y_j, \tag{3.31}$$

для нее применимо утверждение 3.4, из которого следует, что

$$\mathcal{P}\left(X > \frac{\|\Sigma_{VZ}\|_F^2}{N - r - \sqrt{2c(N-r)}} \left(C\|F\|_2^2 + \sqrt{C}S\right)\right) \leqslant \inf_{p \in \mathbb{N}} \frac{\mathbb{E}\left(Y_1^p\right)}{C^p}. \tag{3.32}$$

Далее, оценим матожидание:

$$\frac{\mathbb{E}\left(Y_1^p\right)}{C^p} = \frac{p \int\limits_0^\infty x^{p-1} e^{-x/2} dx}{C^p} = \frac{p 2^p \Gamma(p)}{C^p} < \frac{2{,}2p\sqrt{2\pi(p-1)}}{C} \left(\frac{2(p-1)}{eC}\right)^{p-1}$$
$$= 1{,}1p\sqrt{2\pi/(p-1)} \left(\frac{p-1}{p}\right)^p \left(\frac{2p}{eC}\right)^p, \quad p \geqslant 2.$$

Так как коэффициенты перед последним множителем растут с ростом $p \geqslant 2$, оптимальное p будет меньше оптимального p для последнего множителя, которое, в свою очередь, равно $p = C/2$. Так как далее с ростом p производная остается положительной, то значение при оптимальном целом p будет не хуже, чем при $p = C/2 + 1$, что даст оценку

$$\mathcal{P}\left(X > \frac{\|\Sigma_{VZ}\|_F^2}{N - r - \sqrt{2c(N-r)}}\left(C\|F\|_2^2 + \sqrt{C}S\right)\right) \leqslant 1{,}1\sqrt{\pi C}\frac{C+2}{C}e^{-C/2}, \quad C \geqslant 2. \tag{3.33}$$

Объединив (3.33) с (3.30) с учетом определения X (3.31) и условия на $\mathcal{P}_c$ (3.27), получим утверждение теоремы. □

Таким образом, если $C\sum\limits_{k=1}^{N-r}\sigma_k^4(F) \ll \left(\sum\limits_{k=1}^{N-r}\sigma_k^2(F)\right)^2$, то коэффициент в числителе близок к единице, и столбцовая аппроксимация близка к оптимальной. Например, если $\sigma_k(F) \sim k^{-\alpha}$, $\alpha < 1/4$, то получаем коэффициент $1 + O\left(N^{-1}\right)$; если $1/4 < \alpha < 1/2$, то $1 + O\left(N^{4\alpha-2}\right)$.

Если же $\|F'\|_F = (N-r)\|F'\|_2$ (белый шум), то никакие вероятностные оценки не требуются: согласно (3.15),

$$\begin{aligned}\|A_V - CW\|_F^2 &\leqslant \|F\|_F^2 + \left\|\Sigma'_F U'\Sigma_{VZ}\right\|_F^2 \leqslant \|F\|_F^2 + \left\|\Sigma'_F\right\|_2^2\|U'\Sigma_{VZ}\|_F^2 \\ &= \|F\|_F^2\left(1 + \frac{\|\Sigma_{VZ}\|_F^2}{N-r}\right) \leqslant \left(1 + \frac{r}{n-r+1}\right)\|F\|_F^2.\end{aligned}$$

Заметим, однако, что полученный результат не обобщается на случай крестовой аппроксимации, поскольку в матрице погрешности столбцового приближения $A - \Phi = A - CW$ спектральная норма будет близка к норме Фробениуса, а потому при дальнейшем построении строковой аппроксимации на основе матрицы Φ не будет преимущества от использования теоремы 3.7 по сравнению с теоремой 3.5.

Теорема 3.6 показывает, что достаточно найти подматрицу, которой бы соответствовала подматрица большого проективного объема в матрице наилучшего приближения Z ранга r, чтобы с высокой вероятностью гарантировать аппроксимацию, близкую к оптимальной. В численных экспериментах главы 6 мы продемонстрируем, что для этого на практике достаточно находить подматрицу большого проективного объема в самой матрице A. Таким образом алгоритмы, основанные на максимизации объема искомой подматрицы, почти всегда будут давать погрешность аппроксимации, близкую к оптимальной от нормы Фробениуса, а матрицы, для которых это не так, редки с точки зрения RANDSVD ансамбля.

Глава 4. Поиск существенно невырожденных подматриц

Степень невырожденности подматриц обычно определяется на основе следующих величин [44], определения которых распространяется также на прямоугольные матрицы:

$$\left\|\hat{A}^{+}\right\|_2 / \|A\|_2, \quad \left\|\hat{A}^{+}\right\|_F / \|A\|_F, \quad \mathcal{V}\left(\hat{A}\right) / \max_{\tilde{A}\in\mathbb{C}^{r\times n}} \mathcal{V}\left(\tilde{A}\right).$$

Сначала мы рассмотрим случай $A \in \mathbb{C}^{r\times N}$, а затем перейдем к общему случаю поиска подматрицы локально максимального объема в матрице $A \in \mathbb{C}^{M\times N}$, так как для последнего оценки на число шагов были напрямую получены только для алгоритма поиска квадратных подматриц. Это связано не с проблемами обобщения оценок на число шагов (что делается тривиально), а в том, что для прямоугольных матриц не выведены формулы пересчета при одновременной замене одной строки и одного столбца.

4.1. Связь с нижними оценками столбцовых аппроксимаций

Нижние оценки для величин $\left\|\hat{A}^{+}\right\|_2 / \|A\|_2$ и $\left\|\hat{A}^{+}\right\|_F / \|A\|_F$ тесно связаны с величинами $t(r,n,N)$ и $t_F(r,n,N)$, а потому и с погрешностями столбцовых аппроксимаций матриц.

В частности, справедливо следующее утверждение.

Утверждение 4.1.

$$t(r,n,N) = \sup_{R\in\mathbb{C}^{r\times N}} \min_{\hat{R}\in\mathbb{C}^{r\times n}} \frac{\left\|\hat{R}^{+}\right\|_2^2}{\|R^{+}\|_2^2} \geqslant \sup_{R\in\mathbb{C}^{r\times N}} \min_{\hat{R}\in\mathbb{C}^{r\times n}} \frac{\left\|\hat{R}^{+}\right\|_F^2}{\|R^{+}\|_F^2} = \frac{N-r+1}{n-r+1}.$$

Доказательство. Равенство

$$t(r,n,N) = \sup_{R\in\mathbb{C}^{r\times N}} \min_{\hat{R}\in\mathbb{C}^{r\times n}} \frac{\left\|\hat{R}^{+}\right\|_2^2}{\|R^{+}\|_2^2}$$

уже было нами получено в лемме 2.2 раздела 2.2.2.

Оценка

$$\sup_{R\in\mathbb{C}^{r\times N}} \min_{\hat{R}\in\mathbb{C}^{r\times n}} \frac{\left\|\hat{R}^{+}\right\|_F^2}{\|R^{+}\|_F^2} \leqslant \frac{N-r+1}{n-r+1}$$

доказана в теореме 3.1 из [44]. Эффективный алгоритм её достижения будет нами рассмотрен в разделе 4.6.

Осталось лишь привести пример, доказывающий неравенства для нормы Фробениуса в противоположную сторону (неравенство для спектральной нормы нами уже было доказано в утверждении 2.3).

Рассмотрим матрицу $R \in \mathbb{R}^{r \times N}$:

$$R = \begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & \varepsilon & \cdots & \cdots & \varepsilon \end{bmatrix}$$

с $r-1$-й единицей и $N-r+1$ элементами, равными ε. Тогда при $\varepsilon < 1/\sqrt{N-r+1}$ получаем, что $\|R^+\|_2^2 = \varepsilon^{-2}/(N-r+1)$ и $\|R^+\|_F^2 = \varepsilon^{-2}/(N-r+1) + r - 1$.

Все подматрицы полного ранга $\hat{R} \in \mathbb{R}^{r \times n}$ матрицы R совпадают с точностью до перестановки столбцов, а их псевдообратная (с точностью до перестановки строк) есть

$$\hat{R}^+ = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & 0 & 1 & 0 \\ 0 & \ddots & 0 & \frac{\varepsilon^{-1}}{n-r+1} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \frac{\varepsilon^{-1}}{n-r+1} \end{bmatrix}.$$

Тогда

$$\frac{\left\|\hat{R}^+\right\|_F^2}{\|R^+\|_F^2} = \frac{\varepsilon^{-2}/(n-r+1) + r - 1}{\varepsilon^{-2}/(N-r+1) + r - 1} \xrightarrow[\varepsilon \to 0]{} \frac{N-r+1}{n-r+1},$$

а также

$$\frac{\left\|\hat{R}^+\right\|_2^2}{\|R^+\|_2^2} = \frac{\varepsilon^{-2}/(n-r+1)}{\varepsilon^{-2}/(N-r+1)} = \frac{N-r+1}{n-r+1},$$

что еще раз доказывает оценку на спектральную норму. □

4.2. Выбор стартовой подматрицы

Для дальнейших алгоритмов нам понадобится стартовать с подматрицы, не слишком сильно отличающейся от подматрицы максимального объема. Классическим методом поиска такой подматрицы является выбор ведущих столбцов, когда столбцы в подматрицу набираются по одному, причем каждый новый столбец выбирается с максимальной нормой в подпространстве, ортогональном уже выбранным столбцам. Другими словами, столбцы в матрицу добавляются жадно, так чтобы каждый новый столбец максимизировал её объем. Этот факт следует из следующей леммы.

Лемма 4.1. *Пусть $A \in \mathbb{C}^{r \times k_1}$ и $B \in \mathbb{C}^{r \times k_2}$, $k_1 + k_2 = k \leqslant r$. Тогда*

$$\mathcal{V}([A\ B]) = \mathcal{V}(A) \cdot \mathcal{V}\left(\left(I - AA^+\right) B\right) \leqslant \mathcal{V}(A) \cdot \mathcal{V}(B).$$

В частности, если $k_1 = k - 1$, $b \in \mathbb{C}^{r\times 1}$, то

$$\mathcal{V}([A\ b]) = \mathcal{V}(A) \cdot \left\|\left(I - AA^+\right) b\right\|_2 \leqslant \mathcal{V}(A) \cdot \|b\|_2 ,$$

а также (по индукции)

$$\mathcal{V}(A) \leqslant \prod_{i=1}^{k} \left\|A_{:,i}\right\|_2 .$$

В частности, при $k = r$ справедливо неравенство Адамара

$$|\det A| \leqslant \prod_{i=1}^{r} \left\|A_{:,i}\right\|_2 .$$

Доказательство. Рассмотрим QR разложение матрицы $[A\ B]$:

$$[A\ B] = QR = Q \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}.$$

Матрица Q не меняет сингулярных чисел, поэтому объем определяется объемом матрицы R. Для нее он равен модулю определителя, то есть модулю произведения диагональных элементов. Заметим сразу также, что

$$\mathcal{V}\left(\left(I - AA^+\right) B\right) = \mathcal{V}\left(Q \begin{bmatrix} 0 \\ R_{22} \end{bmatrix}\right) = |\det R_{22}| .$$

Таким образом,

$$\mathcal{V}([A\ B]) = |\det R_{11}|\,|\det R_{22}| = \mathcal{V}(A) \cdot \mathcal{V}\left(\left(I - AA^+\right) B\right) \leqslant \mathcal{V}(A) \cdot \mathcal{V}(B) ,$$

где последнее неравенство следует из того, что проекция не может увеличить сингулярные числа. Остальные оценки леммы следуют из полученной. □

Алгоритм жадного добавления столбцов можно записать следующим образом (см. Алгоритм 4.1).

На каждом шаге выбора ведущих столбцов к подматрице добавляется столбец, обладающий максимальной длиной в подпространстве, ортогональном уже выбранным столбцам. Данный подход обладает одним важным свойством: объем полученной подматрицы не может быть слишком далек от максимального.

Обозначим через $R_{\mathcal{I}}$ подматрицу, заданную набором столбцов из $\mathcal{I}$. В [29] было доказано следующее утверждение. Здесь мы приведем более простое доказательство, основанное на проективном объеме.

Алгоритм 4.1 Выбор ведущих столбцов

Вход: Матрица $R \in \mathbb{C}^{r\times N}$, требуемый ранг r.

Выход: Набор индексов столбцов $\mathcal{I}$ размера r, содержащий подматрицу большого объема.

```
1: I := ∅
2: for k := 1 to r do
3:     for i := 1 to N do
4:         γ_i := ‖R_{:,i}‖_2^2
5:     end for
6:     j = arg max_j γ_j
7:     I := I ∪ {j}
8:     R := R − (1/γ_i) R_{:,i} R*_{:,i} R
9: end for
```

Утверждение 4.2 ([29])**.** *Выбор ведущих столбцов после k шагов возвращает подматрицу $R_{\mathcal{I}} \in \mathbb{C}^{r\times k}$ с объемом, отличающимся от максимального не более, чем в $k!$ раз:*

$$\mathcal{V}\left(R_{\mathcal{I}}\right) \geqslant \frac{1}{k!}\mathcal{V}\left(R_{\mathcal{M}}\right),$$

где $R_{\mathcal{M}} \in \mathbb{C}^{r\times k}$ – подматрица максимального объема.

Оно следует по индукции из следующего утверждения.

Утверждение 4.3. *Пусть объем подматрицы $\hat{R} \in \mathbb{C}^{r\times k}$ в V_k раз меньше максимального в $R \in \mathbb{C}^{r\times N}$. Пусть к ней добавляется столбец $c \in \mathbb{C}^r$ на основе стратегии выбора ведущих столбцов. Тогда объем полученной подматрицы $\tilde{R} = \begin{bmatrix}\hat{R} & c\end{bmatrix} \in \mathbb{C}^{r\times(k+1)}$ отличается от максимального объема не более, чем в*

$$V_{k+1} \leqslant (k+1)V_k$$

раз.

Доказательство. Пусть $R_{\mathcal{M}} \in \mathbb{C}^{r\times(k+1)}$ – подматрица максимального объема. Тогда согласно лемме 1.1

$$\mathcal{V}_k^2\left(R_{\mathcal{M}}\right) \leqslant \sum_{\mathcal{J},|\mathcal{J}|=k} \mathcal{V}^2\left(R_{\mathcal{J}}\right) \leqslant (k+1)V_k^2\mathcal{V}^2\left(\hat{R}\right), \tag{4.1}$$

так как в сумме всего $k+1$ матриц размера $r\times k$, и объем каждой не больше максимального, который, в свою очередь, не превосходит $V_k\mathcal{V}\left(\hat{R}\right)$ по условию.

Далее оценим минимальное $k+1$-е сингулярное число $R_{\mathcal{M}}$. Его можно определить как минимальное значение нормы $R_{\mathcal{M}}$ в ортогональном дополнении к подпространствам размерности

k:

$$\begin{aligned}
\sigma^2_{\min}\left(R_{\mathcal{M}}\right) &= \min_{U\in\mathbb{C}^{r\times k}, U^*U=I} \left\|\left(I - UU^*\right) R_{\mathcal{M}}\right\|_2^2 \\
&\leqslant \left\|\left(I - \hat{R}\hat{R}^+\right) R_{\mathcal{M}}\right\|_2^2 \\
&\leqslant \left\|\left(I - \hat{R}\hat{R}^+\right) R_{\mathcal{M}}\right\|_F^2 \\
&= \sum_{i=1}^{k+1} \left\|\left(I - \hat{R}\hat{R}^+\right)\left(R_{\mathcal{M}}\right)_{:,i}\right\|_2^2 \\
&\leqslant (k+1)\left\|\left(I - \hat{R}\hat{R}^+\right) c\right\|_2^2,
\end{aligned} \tag{4.2}$$

поскольку c обладает наибольшей длиной в ортогональном дополнении к $\hat{R}$.

Из (4.1) и (4.2) окончательно получаем

$$V_{k+1}\mathcal{V}\left(\tilde{R}\right) = \mathcal{V}\left(R_{\mathcal{M}}\right) = \mathcal{V}_k\left(R_{\mathcal{M}}\right)\sigma_{\min}\left(R_{\mathcal{M}}\right) \leqslant (k+1)V_k\mathcal{V}\left(\hat{R}\right)\left\|\left(I - \hat{R}\hat{R}^+\right) c\right\|_2 = (k+1)V_k\mathcal{V}\left(\tilde{R}\right),$$

где последнее равенство следует из леммы 4.1. □

Данный индуктивный переход говорит о том, что если, например, при каждой замене мы увеличиваем объем не менее, чем в ρ раз, то не важно, когда именно происходит замена: после набора k столбцов или до этого: и то, и другое уменьшит итоговую оценку на отношение к максимальному объему в ρ раз. Такой подход (совершение замен до достижения требуемого ранга) применяется в [39] и [41] и позволяет определить требуемый ранг аппроксимации в процессе выполнения алгоритма поиска ρ-локально максимального объема.

Оценку на отношение к максимальному объему можно улучшить, если не использовать индуктивный аргумент.

Утверждение 4.4. *Выбор ведущих столбцов после k шагов возвращает подматрицу $R_{\mathcal{I}} \in \mathbb{C}^{r\times k}$ с объемом, отличающимся от максимального не более, чем в $k^{k/2}$ раз:*

$$\mathcal{V}\left(R_{\mathcal{I}}\right) \geqslant \frac{1}{k^{k/2}}\mathcal{V}\left(R_{\mathcal{M}}\right),$$

где $R_{\mathcal{M}} \in \mathbb{C}^{r\times k}$ – подматрица максимального объема.

Доказательство. Без ограничения общности считаем, что $\hat{R} = R_{\mathcal{I}}$ соответствует первым r столбцам матрицы R. Кроме того, вместо самой матрицы R будем сразу рассматривать матрицу R из её QR разложения, так чтобы $\hat{R} \in \mathbb{C}^{r\times k}$ считать верхней треугольной (с нулевыми строками после k-й). Мы можем так поступаить, поскольку Q не меняет сингулярных чисел, а значит и не влияет на объемы подматриц.

Согласно критерию выбора ведущих столбцов, справедливы оценки

$$\hat{R}_{ii} = \max_j \left\|R_{i:r,j}\right\|_2, \ \forall i. \tag{4.3}$$

Чтобы оценить отношение объемов, оценим напрямую сингулярные числа R_M. Их можно оценить через минимакс по подпространствам, где в качестве подпространств выберем ортогональные первым $i-1$ векторам стандартного базиса:

$$\sigma_i(U_M) = \min_{V \in \mathbb{C}^{(r-i+1)\times r},\ VV^*=I} \|VU_M\|_2 \leqslant \left\|(U_M)_{i:r,:}\right\|_2 \leqslant \sqrt{k} \max_j \left\|U_{i:r,j}\right\|_2 = \sqrt{k}\hat{R}_{ii},$$

где в конце мы воспользовались (4.3).

Взяв произведение по всем i, получаем

$$\mathcal{V}(R_M) = \prod_{i=1}^{k} \sigma_i(U_M) \leqslant k^{k/2} \prod_{i=1}^{k} \hat{R}_{ii} = k^{k/2}\mathcal{V}\left(\hat{R}\right).$$

□

Замечание 4.1. Оценка из [89] говорит о том, что если $r = r_1 + r_2$ и найдены $r_1 \times r_1$ подматрица локально максимального объема, и в ортогональном дополнении к крестовой аппроксимации на ее основе найдена подматрица $r_2 \times r_2$ локально максимального объема, то $r \times r$ подматрица обладает $r^{r/2}$-максимальным объемом. Полученный здесь результат можно интерпретировать как разбиение $r = r_1 + \ldots + r_r$, где все $r_i = 1$. Стоит отметить, что оценка на число шагов при поиске локально максимального объема при расширении размера от r_1 до r, полученная в [89], не является полной, поскольку там не было учтено время, требующееся на поиск $r_2 \times r_2$ подматрицы локально максимального объема. Здесь мы показали, что можно сразу набрать необходимое число столбцов (при последовательном наборе применяется оценка утверждения 4.3), и для этого достаточно стандартного алгоритма выбора ведущих столбцов. Если же, например, делать это иерархически (по типу $r_1 = r_2 = r/2$ и далее разбивая r_1 и r_2), такой подход приведет к асимптотически более высокой сложности алгоритма, а также потребует учета того факта, что текущие оценки не позволяют гарантировать достижения в точности локально максимального объема за конечное число шагов.

Оценка $k^{k/2}$ достигается с произвольной точностью на следующем примере, аналогичном матрице Кахана [90]

$$R = D \begin{bmatrix} I_k & \varphi F_k \end{bmatrix} \in \mathbb{C}^{k\times 2k}, \quad D = \operatorname{diag}\left(1, \zeta, \zeta^2, \ldots, \zeta^{r-1}\right).$$

где $\zeta^2 + \varphi^2 = 1$, а потому φ можно сделать сколь угодно близким к единице. В действительном случае для построения примера вместо матрицы Фурье $F_{k,ij} = e^{-2\pi\sqrt{-1}(i-1)(j-1)/k}$ можно использовать матрицу Адамара, однако матрицы Адамара существуют не для всех значений k. При $r > k$ оставшиеся $r-k$ строк R можно выбрать нулевыми.

Заметим, однако, что при $k = r$, если заранее выполнить LQ разложение матрицы R, и применять выбор ведущих столбцов в матрице Q (L не повлияет на отношение объемов), то данный

контрпример работать не будет, так как сингулярные числа Q равны (в отличие от примера, где они могут отличаться сколь угодно сильно). Тем не менее, можно создать аналогичный контрпример, добавив большое число столбцов, так что оценка $r^{r/2}$ достигается в пределе $\zeta \to 0$ и для ортонормированных строк. Однако, условие нормировки потребует как минимум порядка ζ^{1-r} дополнительных столбцов, чтобы приравнять нормы первой и последней строк. Таким образом, выбор ведущих столбцов в применении к Q может оказаться существенно более эффективным в случае наличия ограничения вида $N \leqslant \mathrm{poly}\,(r)$.

Аналогичным образом можно показать, что неполное разложение Гаусса также применимо для поиска подматриц большого объема в случае $k = r$.

Утверждение 4.5. *Пусть подматрица $R_{\mathcal{I}} \in \mathbb{C}^{r\times r}$ матрицы $R \in \mathbb{C}^{r\times N}$ содержит ведущие элементы, полученные с помощью разложения Гаусса с выбором ведущего элемента по строкам. Тогда её объем отличается от максимального не более чем в $r^{r/2}$ раз:*

$$\mathcal{V}\,(R_{\mathcal{I}}) \geqslant \frac{1}{r^{r/2}}\mathcal{V}\,(R_{\mathcal{M}})\,,$$

где $R_{\mathcal{M}} \in \mathbb{C}^{r\times r}$ – подматрица максимального объема.

Доказательство. Без ограничения общности считаем, что $\hat{R} = R_{\mathcal{I}}$ соответствует первым r столбцам матрицы R. Тогда можно рассмотреть LU разложение

$$R = LU, \quad L \in \mathbb{C}^{r\times r}, \quad U \in \mathbb{C}^{r\times N},$$

где $\hat{U} \in \mathbb{C}^{r\times r}$ – верхняя унитреугольная матрица, то есть $\mathcal{V}\left(\hat{U}\right) = 1$. По условию выбора ведущего элемента получаем

$$|U_{ik}| \leqslant |U_{ii}| = 1, \ \forall i, k.$$

То есть,

$$\|U\|_C \leqslant 1.$$

Умножение на L не влияет на отношение объемов, а потому

$$\mathcal{V}\,(R_M)\,/\mathcal{V}\left(\hat{R}\right) = \mathcal{V}\,(U_M)\,/\mathcal{V}\left(\hat{U}\right) = \mathcal{V}\,(U_M) = \prod_{k=1}^{r}\sigma_k\,(U_M) \leqslant \left(\frac{1}{r}\sum_{k=1}^{r}\sigma_k^2\,(U_M)\right)^{r/2}$$
$$= \left(\frac{\|U_M\|_F^2}{r}\right)^{r/2} \leqslant \left(r\,\|U_M\|_C^2\right)^{r/2} \leqslant r^{r/2}.$$

□

Пример, где оценка $r^{r/2}$ достигается, строится аналогично:

$$R = \begin{bmatrix} I_r & F_r \end{bmatrix} \in \mathbb{C}^{r\times 2r}.$$

В данном случае граница достигается. Этот же пример показывает, что оценка $r^{r/2}$ достигается для подматриц локально максимального объема [11].

В алгоритме поиска сильного выявляющего ранг QR разложения (SRRQR) [39] столбцы подматрицы $R_{\mathcal{I}} \in \mathbb{R}^{r\times k}$ переставляются по одному, пока не будет достигнут ρ-локально максимальный объем. Для $k = r$ он полностью совпадает с алгоритмом maxvol [11], который делает то же самое для квадратных подматриц. Используя утверждение 4.2 мы получаем, что SRRQR делает не более

$$\log_\rho k^{k/2} = \frac{k}{2}\log_\rho k$$

шагов. Так как стоимость одного шага составляет $O(Nr)$, получаем полную стоимость алгоритма $O\left(Nrk\log_\rho k\right)$, что ниже оценки $O\left(Nrk\log_\rho N\right)$, доказанной в [52].

Пусть теперь нам необходимо выбрать подматрицу размера $r \times n$, $n > r$. Мы уже знаем, как выбрать первые r столбцов, однако мы не можем использовать алгоритм 4.1 для выбора оставшихся $n-r$ столбцов. Для этого подходит алгоритм из [42], однако можно построить более быстрый его аналог, что мы и сделаем далее.

Пусть нам дана невырожденная подматрица $\hat{V} \in \mathbb{C}^{r\times n}$ некоторой матрицы $V \in \mathbb{C}^{r\times N}$, и мы хотим максимально увеличить объем $\hat{V}$, добавляя в нее по одному новые столбцы.

Согласно лемме 1.2, необходимо искать столбец с максимальным значением $\left\|\hat{V}^+ v\right\|_2$ или, в терминах матрицы $C = X^{-1}V$, где $\hat{V} = XQ$, $QQ^* = I$ (см. (1.9)), столбец c матрицы C с максимальной длиной. При этом, после его добавления, мы получим

$$\left[\hat{V}\ v\right]^+ V = \left(X\left[Q\ c\right]\right)^+ XC = \left[Q\ c\right]^+ C.$$

Рассмотрим матрицу $A = A^* = I - \left(1 - \frac{1}{\sqrt{1+\|c\|_2^2}}\right)\frac{cc^*}{\|c\|_2^2} \in \mathbb{C}^{r\times r}$. Тогда

$$\begin{aligned} A^2 &= I - 2\left(1 - \frac{1}{\sqrt{1+\|c\|_2^2}}\right)\frac{cc^*}{\|c\|_2^2} + \left(1 - \frac{2}{\sqrt{1+\|c\|_2^2}} + \frac{1}{1+\|c\|_2^2}\right)\frac{cc^*}{\|c\|_2^2} \\ &= I - \frac{cc^*}{\|c\|_2^2} + \frac{1}{1+\|c\|_2^2}\cdot\frac{cc^*}{\|c\|_2^2} \\ &= I - \frac{cc^*}{1+\|c\|_2^2}. \end{aligned}$$

Обозначим $Q' = A\,[Q\ c]$ и заметим, что

$$\begin{aligned}
Q'Q'^* &= A\,[Q\ c]\,(A\,[Q\ c]\,A)^* \\
&= AQQ^*A + Ac\,(Ac)^* \\
&= A^2 + c\left(1 - \left(1 - \frac{1}{\sqrt{1 + \|c\|_2^2}}\right)\right)\left(c\left(1 - \left(1 - \frac{1}{\sqrt{1 + \|c\|_2^2}}\right)\right)\right)^* \\
&= I - \frac{cc^*}{1 + \|c\|_2^2} + \frac{cc^*}{1 + \|c\|_2^2} = I.
\end{aligned}$$

Таким образом,

$$[Q\ c]^+ C = [Q\ c]^+ A^{-1}AC = (A\,[Q\ c])^+ AC = Q'^* AC,$$

и мы можем обозначить $C' = AC$ и продолжить добавление столбцов на основе

$$C' = AC = C - \left(1 - \frac{1}{\sqrt{1 + \|c\|_2^2}}\right)\frac{cc^*}{\|c\|_2^2}C. \tag{4.4}$$

Обозначим через $l_j = \left\|C_j\right\|_2^2$, $j = \overline{1, N}$ квадраты норм столбцов $C_j \in \mathbb{C}^{r\times 1}$ матрицы C. Аналогично через $l'_j = \left\|C'_j\right\|_2^2$ обозначим квадраты норм столбцов C'. Пусть в подматрицу добавляется столбец $c = C_i$. Тогда

$$\begin{aligned}
l'_j &= \left(C_j - \left(1 - \frac{1}{\sqrt{1 + \|C_i\|_2^2}}\right)\frac{C_iC_i^*}{\|C_i\|_2^2}C_j\right)^*\left(C_j - \left(1 - \frac{1}{\sqrt{1 + \|C_i\|_2^2}}\right)\frac{C_iC_i^*}{\|C_i\|_2^2}C_j\right) \\
&= C_j^*C_j + \left(1 - \frac{1}{\sqrt{1 + \|C_i\|_2^2}}\right)^2\frac{\left|C_i^*C_j\right|^2}{\|C_i\|_2^2} - 2\left(1 - \frac{1}{\sqrt{1 + \|C_i\|_2^2}}\right)\frac{\left|C_i^*C_j\right|^2}{\|C_i\|_2^2}.
\end{aligned}$$

В итоге, используя определения l_i и l_j, получаем

$$l'_j = l_j - \frac{\left|C_i^*C_j\right|^2}{1 + l_i}. \tag{4.5}$$

Данное выражение можно пересчитывать проще, обозначив $C_0 = \hat{V}_0^{-1}V$ как начальное значение C, а далее определить $C = XC_0$ и пересчитывать только X (в начале $X = I$), поддерживая $\hat{C}\hat{C}^* = QQ^* = I$. Еще проще обозначить $Y = X^*X = \left(\hat{C}_0\hat{C}_0^*\right)^{-1}$, потому что тогда

$$C_i^*C_j = C_{0i}^*YC_{0j}.$$

Саму матрицу Y можно пересчитать с помощью формулы Шермана-Вудбери-Моррисона:

$$Y' = \left(\hat{C}'_0\hat{C}'^*_0\right)^{-1} = \left(\hat{C}_0\hat{C}_0^* + C_{0i}C_{0i}^*\right)^{-1} = \left(Y^{-1} + C_{0i}C_{0i}^*\right)^{-1} = Y - \frac{YC_{0i}C_{0i}^*Y}{1 + C_{0i}^*YC_{0i}} = Y - \frac{1}{1 + l_i}\left(C_{0i}^*Y\right)^*C_{0i}^*Y. \tag{4.6}$$

Так как rect-maxvol применяется уже после того, как найдена подматрица $r \times r$ локально максимального объема (или как минимум после выбора ведущих столбцов), число обусловленности Y будет ограничено:

$$\operatorname{cond}(Y) = \left\|\hat{C}_0\right\|_2^2 \left\|\hat{C}_0^+\right\|_2^2 = \left\|\hat{V}_0^{-1}\hat{V}\right\|_2^2 \left\|\hat{V}^+\hat{V}_0\right\|_2^2 \leqslant \left\|\hat{V}_0^{-1}V\right\|_2^2 \left\|\hat{V}^+\hat{V}\right\|_2^2 = \left\|\hat{V}_0^{-1}\right\|_2^2 \leqslant 1 + r(N-r)$$

для подматрицы локально максимального объема согласно следствию 1.6. Это позволяет не опасаться погрешности, возникающей из-за ограниченности машинной точности. Запишем полученные выражения для быстрого пересчета (4.5) и (4.6) в виде алгоритма 4.2.

Алгоритм 4.2 rect-maxvol [42], быстрая версия

Вход: Строки $R \in \mathbb{C}^{r\times N}$, стартовый набор индексов столбцов $\mathcal{I}$ размера r, итоговое число столбцов n.

Выход: В $\mathcal{I}$ добавляются $n-r$ индексов столбцов, жадно максимизирующих объем подматрицы $\hat{A} = R_{:,\mathcal{I}}$.

$Y := I$
$C := R_{:,\mathcal{I}}^{-1} R$
for $j := 1$ **to** N **do**
    $l_j := \|C_{:,j}\|_2^2$
end for
for $k := r+1$ **to** n **do**
    $i := \arg\max_{i, i\notin\mathcal{I}} l_i$
    Индекс i добавляется в множество $\mathcal{I}$
    $Y' := C_{:,i}^* Y$
    $C' := Y'C$
    $Y := Y - \frac{1}{1+l_i} Y'^* Y'$
    for $j := 1$ **to** N **do**
        $l_j := l_j - \frac{1}{1+l_i}\left|C'_j\right|^2$
    end for
end for

4.3. Поиск подматриц локально максимального объема в фиксированных строках или столбцах

Идея максимизации объема подматрицы впервые появилась в работах по дизайну экспериментов, где она соответствует D-оптимальности [91, 92]. Жадный алгоритм набора столбцов был позже переоткрыт в [42] и затем назван rect-maxvol. Мы воспользуемся теми же идеями в качестве базы для алгоритма поиска подматрицы ρ-локально максимального объема. Здесь нами

будет также доказана оценка на число шагов такого алгоритма, которая, в частности, улучшает оценку для случая квадратных подматриц ($n = r$).

Начнем с того, что напомним читателю алгоритм maxvol [11], который позволяет находить квадратные подматрицы локально максимального объема в фиксированных столбцах C (или строках R). Алгоритм стартует с произвольной невырожденной подматрицы $\hat{A} \in \mathbb{C}^{r\times r}$ в столбцах $C \in \mathbb{C}^{M\times r}$ (или строках $R \in \mathbb{C}^{r\times N}$) матрицы $A \in \mathbb{C}^{M\times N}$. Он поочередно заменяет одну из строк (а когда это невозможно, то переключается на столбцы) подматрицы $\hat{A}$ на новую строку (или столбец) из C (или R).

Замена выбирается так, чтобы максимально увеличить объем $\hat{A}$, и использует следующий критерий. Здесь и далее будем работать в строках R и, соответственно, менять столбцы $\hat{A}$ (через C мы далее будем обозначать определенную ниже матрицу коэффициентов).

Утверждение 4.6 ([11])**.** *Пусть $\hat{A} \in \mathbb{C}^{r\times r}$ – подматрица в первых r столбцах матрицы $R \in \mathbb{C}^{r\times N}$. Тогда замена i-го столбца подматрицы $\hat{A}$ на j-й столбец R ($j > r$) меняет объем $\hat{A}$, $V_{old} = \mathcal{V}(\hat{A})$, в*

$$V_{new}/V_{old} = \left|\left(\hat{A}^{-1}R\right)_{ij}\right|$$

раз.

Алгоритм 4.3 maxvol [11]

Вход: Матрица $R \in \mathbb{C}^{r\times N}$, стартовый набор столбцов $\mathcal{I}$ размера r. Параметр ρ.

Выход: Индексы столбцов подматрицы ρ-локально максимального объема записываются в $\mathcal{I}$.

1: $C := R_{\mathcal{I}}^{-1}R$
2: $\{i, j\} \arg\max_{i,j} \left|C_{i,j}\right|$
3: **while** $\left|C_{i,j}\right| > \rho$ **do**
4: Меняем i на j в $\mathcal{I}$
5: Обновляем C
6: $\{i, j\} \arg\max_{i,j} \left|C_{i,j}\right|$
7: **end while**

Основная идея алгоритма состоит в том, что поиск локально максимального объема в R можно заменить на поиск в $C = R_{\mathcal{I}}^{-1}R$, поскольку умножение на $R_{\mathcal{I}}^{-1}$ меняет объем всех подматриц

на один и тот же фактор $\mathcal{V}\left(R_{\mathcal{I}}^{-1}\right)$. Действительно, для произвольной подматрицы $\tilde{R} \in \mathbb{C}^{r\times n}$, $n \geqslant r$,

$$\begin{aligned}\mathcal{V}^2\left(R_{\mathcal{I}}^{-1}\tilde{R}\right) &= \left|\det\left(R_{\mathcal{I}}^{-1}\tilde{R}\tilde{R}^* R_{\mathcal{I}}^{-*}\right)\right| \\ &= \left|\det R_{\mathcal{I}}^{-1}\right|\left|\det\left(\tilde{R}\tilde{R}^*\right)\right|\left|\det R_{\mathcal{I}}^{-*}\right| \\ &= \left|\det R_{\mathcal{I}}^{-1}\right|^2\left|\det\left(\tilde{R}\tilde{R}^*\right)\right| \\ &= \mathcal{V}^2\left(R_{\mathcal{I}}^{-1}\right)\mathcal{V}^2\left(\tilde{R}\right).\end{aligned}$$

Кроме того, текущая подматрица в C всегда единичная: $C_{\mathcal{I}} = R_{\mathcal{I}}^{-1}R_{\mathcal{I}} = I$, а потому её определитель после замены одного из столбцов на столбец $C_{:,j}$ легко оценить:

$$\begin{vmatrix} 1 & 0 & C_{1j} & 0 & 0 \\ 0 & 1 & \vdots & \vdots & \vdots \\ \vdots & 0 & C_{ij} & 0 & \vdots \\ \vdots & \vdots & \vdots & 1 & 0 \\ 0 & 0 & C_{rj} & 0 & 1 \end{vmatrix} = C_{ij}.$$

Таким образом, на каждом шаге алгоритма объем растет в $\left|C_{i,j}\right| > \rho$ раз. Следовательно, если стартовать с подматрицы, полученной после выбора ведущих столбцов, то, исходя из утверждения 4.4, число шагов s ограничено величиной

$$s \leqslant \log_\rho r^{r/2} = \frac{r}{2}\log_\rho r. \tag{4.7}$$

Схожая оценка была получена в [89], где предполагался последовательный набор сначала r, а потом r_ρ столбцов. Стоит отметить, что при $r > 0$ полученная там оценка является менее строгой, поскольку в доказательстве предполагает, что $r \times r$ подматрица является доминантной (то есть $\rho = 1$), для чего не существует полиномиальной оценки числа шагов. Тем не менее, при $r = 0$ набор r_ρ столбцов из [89] приводит к той же оценке, что и выбор ведущих столбцов или исключение Гаусса.

Обновление матрицы C является одноранговым и занимает $O(Nr)$ операций, что приводит к общей сложности алгоритма maxvol $O\left(Nr^2\log_\rho r\right)$. Та же оценка на число шагов применима и к другим методам поиска локально максимального объема, в частности, для сильного выявляющего ранг разложения Холецкого [41].

Если требуется найти прямоугольную подматрицу большого объема, то можно сначала найти $r \times r$ подматрицу с помощью алгоритма maxvol, а затем увеличить её размер с помощью $\mathrm{rect\text{-}maxvol}$ [42]. К сожалению, не существует оценок на то, насколько $\mathrm{rect\text{-}maxvol}$ может гарантированно увеличить объем подматрицы, и насколько новая подматрица окажется близкой к подматрице локально максимального объема. Более того, вполне может оказаться, что подматрицы $r\times n$ локально максимального объема не содержат подматриц $r\times r$ локально максимального

объема, а потому простое увеличение размера без замены столбцов может оказаться бесполезным.

Чтобы решить данную проблему, мы далее построим алгоритм $\mathrm{dominant}$, который, аналогично maxvol, позволяет находить $r \times n$ подматрицы ρ-локально максимального объема для $n \geqslant r$. Благодаря выведенным далее формулам пересчета, алгоритм $\mathrm{dominant}$ позволяет производить замены за $O(Nn)$ операций. Более того, для него также можно ограничить число шагов, что приводит к итоговой оценке сложности $O\left(Nnr \log_\rho n\right)$. Наконец, нами будет позже показано, что при $n \gg r$ объем подматриц локально максимального объема близок к максимальному объему, что не верно в случае $n = r$ [29, 50].

В [78], по аналогии с [8] были доказаны оценки крестовой аппроксимации по спектральной норме, основанные на подматрицах локально максимального объема. При этом было предложено использовать алгоритм $\mathrm{rect\text{-}maxvol}$ для их достижения. Однако, как уже упоминалось выше, $\mathrm{rect\text{-}maxvol}$ сам по себе не гарантирует, что итоговая подматрица будет обладать ρ-локально максимальным объемом или что её объем будет хоть как-то оцениваться через максимальный. С другой стороны, возможность достижения подматриц ρ-локально максимального объема позволяет гарантировать достижимость на практике оценок по спектральной норме. Алгоритм $\mathrm{dominant}$ нами будет позже использован и для поиска подматриц большого проективного объема, что, как показано в главе 3, часто приводит к точности аппроксимации, сколь угодно близкой к сокращенному сингулярному разложению.

Чтобы обобщить поиск локально максимального объема на случай $n \geqslant r$ столбцов, воспользуемся уже известными результатами, доказанными для алгоритма $\mathrm{rect\text{-}maxvol}$. Данный алгоритм набора столбцов основан на следующих леммах, которыми мы воспользуемся для построения алгоритма $\mathrm{dominant}$. Первая описывает то, насколько объем меняется при добавлении нового столбца к $\hat{R} = \hat{A}$, и уже была нами получена как лемма 1.2. Здесь мы сформулируем её же для матрицы $C = \hat{R}^+ R$.

Лемма 4.2 ([42])**.** *Пусть $\hat{R} \in \mathbb{C}^{r\times n}$, $n \geqslant r$ – подматрица матрицы $R \in \mathbb{C}^{r\times N}$. Тогда для подматрицы $\widehat{R} \in \mathbb{C}^{r\times(n+1)}$, полученной из $\hat{R}$ добавлением к ней j-го столбца R,*

$$\frac{\mathcal{V}^2\left(\widehat{R}\right)}{\mathcal{V}^2\left(\hat{R}\right)} = 1 + \left\|C_{:,j}\right\|_2^2,$$

где $C = \hat{R}^+ R$.

Следующая лемма описывает обновление матрицы $C = \hat{R}^+ R$, когда подматрица $\hat{R}$ расширяется до $\widehat{R}$ новым столбцом.

Лемма 4.3 ([42])**.** *Пусть $\hat{R} \in \mathbb{C}^{r\times n}$, $n \geqslant r$ – подматрица матрицы $R \in \mathbb{C}^{r\times N}$. Пусть $C = \hat{R}^+ R \in \mathbb{C}^{n\times N}$. Тогда для подматрицы $\widehat{R} \in \mathbb{C}^{r\times(n+1)}$ полученной из $\hat{R}$ добавлением к ней j-го столбца R и*

$\tilde{C} = \hat{R}^{+}R \in \mathbb{C}^{(n+1)\times N}$,

$$\tilde{C} = \begin{bmatrix} C - \frac{1}{1+\left\|C_{:,j}\right\|_2^2} C_{:,j} C_{:,j}^{*} C \\ \frac{1}{1+\left\|C_{:,j}\right\|_2^2} C_{:,j}^{*} C \end{bmatrix}. \tag{4.8}$$

Более того, для любого k новый квадрат длины k-го столбца $\tilde{C}_{:,k}$ становится равным

$$\left\|\tilde{C}_{:,k}\right\|_2^2 = \left\|C_{:,k}\right\|_2^2 - \frac{1}{1+\left\|C_{:,j}\right\|_2^2}\left|C_{:,k}^{*} C_{:,j}\right|^2. \tag{4.9}$$

Последнее выражение (4.9) соответствует полученному нами ранее выражению (4.5).

Заметим, что, в сравнении с выбором ведущих столбцов, для $\mathrm{rect\text{-}maxvol}$ не существует аналога утверждения 4.2. Не смотря на это, утверждения 4.2 уже достаточно, чтобы гарантировать, что объем стартовой подматрицы не слишком мал. Поэтому нам будет достаточно добавить произвольные $n - r$ столбцов к уже выбранным r ведущим столбцам. После этого мы будем менять столбцы по одному. Леммы 4.2 и 4.3 используются для доказательства следующего критерия обновления.

Лемма 4.4. *Пусть $\hat{R} \in \mathbb{R}^{r\times n}$ – подматрица в первых n столбцах матрицы $R \in \mathbb{R}^{r\times N}$. Тогда, замена i-го столбца $\hat{R}$ на j-й столбец R (для $i > n$) меняет квадрат объема $\mathcal{V}^2\left(\hat{R}\right)$ в*

$$B_{ij} = |C_{ij}|^2 + \left(1 + \left\|C_{:,j}\right\|_2^2\right)\left(1 - \left\|C_{:,i}\right\|_2^2\right)$$

раз, где $C = \hat{R}^{+}R$.

Поэтому когда $\max\limits_{ij} B_{ij} > \rho^2$, мы можем увеличить объем как минимум в ρ раз, а когда $\max\limits_{ij} B_{ij} \leqslant \rho^2$, согласно определению 1.6 мы получаем подматрицу ρ-локально максимального объема.

Доказательство. Для доказательства леммы нам потребутся вывести формулы быстрого обновления матрицы C.

Обновление будем осуществлять в два шага. На первом мы добавляем j-й столбец, а на втором удаляем i-й столбец. Мы обозначим соответствующие подматрицы как $\hat{R} \in \mathbb{R}^{r\times n}$ (начальная), $\widehat{R} \in \mathbb{R}^{r\times(n+1)}$ (после добавления j-го столбца) и $\hat{R}' \in \mathbb{R}^{r\times n}$ (после добавления j-го и удаления i-го столбца). С их помощью получаем матрицы $C = \hat{R}^{+}R \in \mathbb{R}^{n\times N}$, $\tilde{C} = \widehat{R}^{+}R \in \mathbb{R}^{(n+1)\times N}$ и $C' = \hat{R}'^{+}R \in \mathbb{R}^{n\times N}$.

Из уравнения (4.9) следует, что квадрат длины j-го столбца меняется при его добавлении как

$$\left\|\tilde{C}_{:,j}\right\|_2^2 = \left\|C_{:,j}\right\|_2^2 - \frac{\left\|C_{:,j}\right\|_2^4}{1+\left\|C_{:,j}\right\|_2^2} = \frac{\left\|C_{:,j}\right\|_2^2}{1+\left\|C_{:,j}\right\|_2^2}.$$

Аналогичная формула верна при удалении i-го столбца. Чтобы её получить, заменим i на j, а C на C' (таким образом, мы рассматриваем $\tilde{C}$ как обновление C' при добавлении i-го столбца):

$$\left\|\tilde{C}_{:,i}\right\|_2^2 = \frac{\left\|C'_{:,i}\right\|_2^2}{1 + \left\|C'_{:,i}\right\|_2^2}.$$

Данное равенство можно переписать в следующем виде:

$$\frac{1}{1 + \left\|C'_{:,i}\right\|_2^2} = 1 - \left\|\tilde{C}_{:,i}\right\|_2. \tag{4.10}$$

Мы также можем выписать изменение длины i-го столбца при переходе отC к $\tilde{C}$, используя (4.9):

$$\left\|\tilde{C}_{:,i}\right\|_2^2 = \left\|C_{:,i}\right\|_2^2 - \frac{1}{1 + \left\|C_{:,j}\right\|_2^2} \left|C_{:,i}^T C_{:,j}\right|^2. \tag{4.11}$$

Теперь заметим, что $C_{:,i}^T C_{:,j} = C_{ij}$:

$$C_{:,i}^T C_{:,j} = R_{:,i}^T \left(\hat{R}^+\right)^T \hat{R}^+ R_{:,j} = R_{:,i}^T \left(\hat{R}\hat{R}^T\right)^{-1} R_{:,j} = \hat{R}_{:,i}^T \left(\hat{R}\hat{R}^T\right)^{-1} R_{:,j} = \hat{R}_{:,i}^+ R_{:,j} = C_{ij},$$

где мы воспользовались тем фактом, что i-й столбец R также является i-м столбцом $\hat{R}$. Затем мы подставляем $C_{:,i}^T C_{:,j} = C_{ij}$ в (4.11):

$$\left\|\tilde{C}_{:,i}\right\|_2^2 = \left\|C_{:,i}\right\|_2^2 - \frac{1}{1 + \left\|C_{:,j}\right\|_2^2} \left|C_{ij}\right|^2. \tag{4.12}$$

Дважды применяя лемму 4.2 (для добавления и для удаления), мы получаем итоговое изменение квадрата объема подматрицы:

$$B_{ij} = \frac{\mathcal{V}^2\left(\hat{R}'\right)}{\mathcal{V}^2\left(\hat{R}\right)} \cdot \frac{\mathcal{V}^2\left(\hat{R}\right)}{\mathcal{V}^2\left(\hat{R}\right)} = \left(1 + \left\|C_{:j}\right\|_2^2\right) / \left(1 + \left\|C'_{:i}\right\|_2^2\right) = \left(1 + \left\|C_{:j}\right\|_2^2\right)\left(1 - \left\|\tilde{C}_{:i}\right\|_2^2\right),$$

где последнее равенство следует из (4.10). Осталось лишь подставить $\left\|\tilde{C}_{:,i}\right\|_2^2$ из уравнения (4.12), чтобы получить равенство

$$B_{ij} = \left(1 + \left\|C_{:,j}\right\|_2^2\right)\left(1 - \left\|C_{:,i}\right\|_2^2\right) + |C_{ij}|^2.$$

□

Использование матрицы B_{ij} для принятия решения о замене столбцов приводит нас к следующему алгоритму. Мы назовем его dominant, поскольку он позволяет находить доминантную прямоугольную подматрицу. Заметим, что хотя сама идея набора или замены строк/столбцов в этом и предыдущем алгоритмах не нова, и применялась ранее в задаче поиска D-оптимального

дизайна [91], предложенный здесь алгоритм позволяет эффективней производить обновления благодаря использованию матрицы C. И, что особенно важно, будет показано, что такой подход позволяет быстро достичь ρ-локально максимального объема, что гарантирует высокую точность по норме Чебышева (секция 2.1.1) и в большинстве случаев по норме Фробениуса (секция 3).

Алгоритм 4.4 dominant

Вход: Матрица $R \in \mathbb{R}^{r \times N}$, стартовый набор индексов столбцов $\mathcal{I}$ размера n. Например, $\mathcal{I} = \{1, ..., n\}$. Параметр ρ.

Выход: Обновленные набор индексов $\mathcal{I}$, соответствующий подматрице ρ-локально максимального объема.

1: $C := R_{\mathcal{I}}^{+} R$
2: **for** $i := 1$ to n **do**
3: **for** $j := n + 1$ to N **do**
4: $B_{i,j} := \left(1 + \|C_{:,j}\|_2^2\right)\left(1 - \|C_{:,i}\|_2^2\right) + |C_{i,j}|^2$
5: **end for**
6: **end for**
7: $\{i, j\} := \arg\max_{i,j} B_{i,j}$
8: **while** $B_{i,j} > \rho^2$ **do**
9: Замена j на i в $\mathcal{I}$
10: Обновление C и B
11: $\{i, j\} := \arg\max_{i,j} B_{i,j}$
12: **end while**

Построение стартовой матрицы C требует $O(Nnr)$ операций. Далее мы покажем, как обновлять матрицу C за $O(Nn)$ операций. Тогда s шагов алгоритма займут в общей сложности $O(Nnr + Nns)$ операций.

Быстрое обновление от C до $\tilde{C}$ при добавлении нового столбца уже известно (лемма 4.3). Таким образом, осталось показать, как перейти от $\tilde{C}$ до C' при удалении столбца подматрицы.

Обозначим $\bar{C} \in \mathbb{R}^{n \times N}$ первые n строк $\tilde{C}$ (без n+1-й строки). Эта подматрица задается верхним блоком в уравнении (4.8). Её j-й столбец это

$$\bar{C}_{:,j} = C_{:,j} - \frac{1}{1 + \left\|C_{:,j}\right\|_2^2} C_{:,j} C_{:,j}^T C_{:,j} = \frac{1}{1 + \left\|C_{:,j}\right\|_2^2} C_{:,j}. \tag{4.13}$$

i-й столбец $\bar{C}_{:,i}$ найдем из уравнения (4.13), заменив i на j, а C на C' (снова интерпретируя $\tilde{C}$

как полученную из C' при добавлении i-го столбца) и подставив (4.10):

$$\bar{C}_{:,i} = \frac{1}{1 + \left\| C'_{:,i} \right\|_2^2} C'_{:,i} = C'_{:,i} \left(1 - \left\| \tilde{C}_{:,i} \right\|_2^2 \right). \tag{4.14}$$

Опять же, интерпретируя $\tilde{C}$ как полученную из C' при добавлении i-го столбца, получаем из уравнения (4.8):

$$\bar{C} = C' - \frac{1}{1 + \left\| C'_{:,i} \right\|_2^2} C'_{:,i} C'^T_{:,i} C'. \tag{4.15}$$

Теперь мы можем подставить $C'_{:,i}$ из уравнения (4.14), а $1 + \left\| \tilde{C}_{:,i} \right\|_2^2$ из уравнения (4.10), чтобы выразить C' в терминах $\bar{C}$ в уравнении (4.15):

$$C' = \bar{C} + \left(\frac{1}{1 - \left\| \tilde{C}_{:,i} \right\|_2^2} \bar{C}_{:,i} \right) \left(\frac{1}{1 + \left\| C'_{:,i} \right\|_2^2} C'^T_{:,i} C' \right). \tag{4.16}$$

Согласно лемме 4.3, выражение внутри второй пары скобок $\frac{1}{1+\left\| C'_{:,i} \right\|_2^2} C'^T_{:,i} C'$ есть последняя строка (нижний блок) матрицы $\tilde{C}$, что дает нам достаточно информации для вычисления C' из $\tilde{C}$, причем выражение (4.16) требует $O(Nn)$ операций. Подробный псевдокод для алгоритма есть в приложении А.

Далее мы докажем оценку на число шагов алгоритма $\mathrm{dominant}$.

Утверждение 4.7. *Пусть стартовый набор n столбцов $\mathcal{I}$ содержит r столбцов $\mathcal{I}'$, полученных с помощью выбора ведущих столбцов. Тогда цикл while в алгоритме 4.4 выполняется*

$$s \leqslant r \log_\rho n$$

раз.

Доказательство. Из следствия 1.1,

$$\mathcal{V}\left(R_{\mathcal{I}}\right) = \sqrt{\sum_{\mathcal{J}' \subset \mathcal{I}, |\mathcal{J}'| = r} \left| \det R_{\mathcal{J}'} \right|^2},$$

а поскольку одна из подматриц – это $R_{\mathcal{I}'}$,

$$\mathcal{V}\left(R_{\mathcal{I}}\right) \geqslant \sqrt{\left| \det R_{\mathcal{I}'} \right|^2} = \mathcal{V}\left(R_{\mathcal{I}'}\right). \tag{4.17}$$

С другой стороны, применяя следствие 1.1 к произвольной подматрице $R_{\mathcal{J}} \in \mathbb{C}^{r \times n}$,

$$\mathcal{V}\left(R_{\mathcal{J}}\right) = \sqrt{\sum_{\mathcal{J}' \subset \mathcal{J}, |\mathcal{J}'| = r} \left| \det R_{\mathcal{J}'} \right|^2}.$$

Поскольку всего в ней C_n^r подматриц размера $r \times r$, то

$$\max_{\mathcal{J},|\mathcal{J}|=n} \mathcal{V}\left(R_{\mathcal{J}}\right) \leqslant \sqrt{\max_{\mathcal{J}',|\mathcal{J}'|=r} C_n^r \left|\det R_{\mathcal{J}'}\right|^2} \leqslant \sqrt{C_n^r} \max_{\mathcal{J}',|\mathcal{J}'|=r} \mathcal{V}\left(R_{\mathcal{J}'}\right). \tag{4.18}$$

Используя утверждение 4.4 вместе с уравнениями (4.17) и (4.18), получаем

$$\mathcal{V}\left(R_{\mathcal{I}}\right) \geqslant \mathcal{V}\left(R_{\mathcal{I}'}\right) \geqslant \frac{1}{\sqrt{C_n^r} r^{r/2}} \max_{\mathcal{J},|\mathcal{J}|=n} \mathcal{V}\left(R_{\mathcal{J}}\right).$$

Поскольку $\mathrm{dominant}$ увеличивает объем хотя бы в ρ раз на каждой итерации цикла, число шагов s ограничено величиной

$$s \leqslant \log_\rho \left(\sqrt{C_n^r} r^{r/2}\right) \leqslant \log_\rho \left(\sqrt{en}\right)^r = r \log_\rho (en).$$

□

Так что если, например, $\rho \leqslant \sqrt{n}$, ограничение на число шагов доминирует в оценке вычислительной сложности, которая в этом случае составит $O\left(Nnr \log_\rho n\right)$. После остановки $B_{ij} \leqslant \rho^2$ по определению 1.6 мы получаем подматрицу ρ-локально максимального объема.

Наконец, мы также можем ограничить нормы $R_{\mathcal{I}}^+ R$ благодаря следующей теореме.

Теорема 4.1. *Для любой подматрицы ρ-локально максимального объема $\hat{R} \in \mathbb{R}^{r\times n}$ матрицы $R \in \mathbb{R}^{r\times N}$*

$$\left\|\hat{R}^+ R\right\|_F^2 \leqslant r + \frac{r + (\rho^2 - 1)n}{n - r + 1} (N - n).$$

Доказательство. Выберем произвольный столбец j вне $\hat{R}$. Поскольку для подматрицы ρ-локально максимального объема $B_{ij} \leqslant \rho^2$ для любого i,

$$\begin{aligned} nc^2 \geqslant \sum_{i=1}^n B_{ij} &= \left\|C_{:,j}\right\|_2^2 + \left(1 + \left\|C_{:,j}\right\|_2^2\right)\left(n - \sum_{i=1}^n \left\|C_{:,i}\right\|_2^2\right) \\ &= \left\|C_{:,j}\right\|_2^2 + \left(1 + \left\|C_{:,j}\right\|_2^2\right)\left(n - \left\|\hat{R}^+\hat{R}\right\|_F^2\right) \\ &= (n - r + 1)\left\|C_{:,j}\right\|_2^2 + (n - r), \end{aligned}$$

поэтому, преобразуя и используя произвольность j,

$$\max_{j>n} \left\|C_{:,j}\right\|_2^2 \leqslant \frac{r + (c^2 - 1)n}{n - r + 1}.$$

Теперь, выражая $\left\|\hat{R}^+ R\right\|_F^2$ как сумму квадратов длин столбцов, получаем

$$\left\|\hat{R}^+ R\right\|_F^2 \leqslant \left\|\hat{R}^+\hat{R}\right\|_F^2 + \sum_{j=n+1}^N \left\|\hat{R}^+ R_{:,j}\right\|_2^2 = r + \sum_{j=n+1}^N \left\|C_{:,j}\right\|_2^2 \leqslant r + \frac{r + (c^2 - 1)n}{n - r + 1} (N - n).$$

□

Следствие 4.1. Подматрица $\hat{R} \in \mathbb{R}^{r\times n}$ на выходе алгоритма 4.4 обладает следующими свойствами:

$$\begin{aligned}\frac{\left\|\hat{R}^{+}\right\|_F^2}{\|R^{+}\|_F^2} &\leqslant \left(1+\frac{1+(\rho^2-1)n/r}{n-r+1}\,(N-n)\right)\frac{r\,\|R^{+}\|_2^2}{\|R^{+}\|_F^2},\\ \frac{\left\|\hat{R}^{+}\right\|_2^2}{\|R^{+}\|_2^2} &\leqslant 1+\frac{r+(\rho^2-1)n}{n-r+1}\,(N-n)\,.\end{aligned} \tag{4.19}$$

Доказательство.

$$\left\|\hat{R}^{+}R\right\|_F^2 \geqslant \left\|\hat{R}^{+}\right\|_F^2 \sigma_r^2\,(R) = \left\|\hat{R}^{+}\right\|_F^2 / \left\|R^{+}\right\|_2^2, \tag{4.20}$$

где $\sigma_r\,(R)$ – это r-е (минимальное) сингулярное число R. Первое неравенство в выражении (4.19) следует из уравнения (4.20) и того, что $\hat{R}$ обладает ρ-локально максимальным объемом, так что согласно теореме 4.1,

$$\left\|\hat{R}^{+}R\right\|_F^2 \leqslant r+\frac{r+(\rho^2-1)n}{n-r+1}\,(N-n)\,.$$

Поскольку сингулярные числа $\hat{R}^{+}R$ не меньше сингулярных чисел $\hat{R}^{+}\hat{R}$, то первые r сингулярных чисел не меньше 1. А сумма квадратов r сингулярных чисел даст нам норму Фробениуса:

$$\left\|\hat{R}^{+}R\right\|_F^2 = \sum_{k=1}^{r}\sigma_k^2\left(\hat{R}^{+}R\right) \geqslant \left\|\hat{R}^{+}R\right\|_2^2 + r - 1.$$

Следовательно,

$$\left\|\hat{R}^{+}\right\|_2^2 / \left\|R^{+}\right\|_2^2 = \left\|\hat{R}^{+}\right\|_2^2 \sigma_r^2\,(R) \leqslant \left\|\hat{R}^{+}R\right\|_2^2 \leqslant 1+\frac{r+(c-1)n}{n-r+1}\,(N-n)\,,$$

что доказывает второе неравенство. □

Имеет смысл отметить, что норма Чебышева матрицы $C = \hat{R}^{+}R$ также ограничена.

Утверждение 4.8. *Для подматрицы $\hat{R} \in \mathbb{C}^{r\times n}$ ρ-локально максимального объема в матрице $R \in \mathbb{C}^{r\times N}$*

$$\left\|\hat{R}^{+}R\right\|_C^2 \leqslant \rho^2 - \frac{\rho^2+1}{2}\left(1-\frac{r}{n}\right).$$

Если $\rho = 1$, то

$$\left\|\hat{R}^{+}R\right\|_C^2 \leqslant \frac{r}{n}.$$

Доказательство. Пусть максимальный коэффициент $\rho' \leqslant \rho$ в произвольном столбце i достигается на строке k матрицы C:

$$|C_{ik}|^2 + (1+l_i)\,(1-l_k) = \rho'^2.$$

С учетом $|C_{ik}|^2 \leqslant l_i$ и $l_k \geqslant 0$ получаем

$$\begin{aligned} l_i + 1 + l_i &\leqslant \rho'^2,\\ l_i &\geqslant \frac{\rho'^2-1}{2}.\end{aligned} \tag{4.21}$$

Далее, с учетом (4.21) и $\max\limits_j l_j \geqslant \frac{1}{n}\sum\limits_{j=1}^{n} l_j = r/n$ получаем для любого j, что

$$\left|C_{ij}\right|^2 + \left(1 + \frac{\rho'^2 - 1}{2}\right)\left(1 - \frac{r}{n}\right) \leqslant \rho'^2,$$

$$\left|C_{ij}\right|^2 \leqslant \rho'^2 - \frac{\rho'^2 - 1}{2} \leqslant \rho^2 - \frac{\rho^2 - 1}{2}.$$

Так как i и j произвольные, то

$$\left\|\hat{R}^+ R\right\|_C^2 = \|C\|_C^2 \leqslant \rho - \frac{\rho - 1}{2}\left(1 - \frac{r}{n}\right).$$

□

Чтобы построить алгоритм крестовой аппроксимации, нам также потребуется уметь менять строки $r \times n$ подматриц и столбцы $m \times r$ подматриц соответственно. Такой алгоритм уже был разработан в [39] и соответствует сильному выявляющему ранг QR разложению (SRRQR). Хотя он был сформулирован для действительных матриц, его легко обобщить на комплексный случай (см. приложение А). Он использует следующий критерий обновления, доказательство которого для полноты изложения также приведено.

Лемма 4.5 ([39], лемма 3.1)**.** *Пусть*

$$A = Q\begin{bmatrix} R & B \\ & C \end{bmatrix} \in \mathbb{C}^{m\times N},$$

где $\hat{A} = QR$ есть QR разложение подматрицы $\hat{A} \in \mathbb{C}^{m\times r}$ матрицы A.

Тогда замена i-го столбца подматрицы $\hat{A}$ j-м столбцом матрицы A меняет объем $\mathcal{V}\left(\hat{A}\right) = V_{old}$ в

$$V_{new}^2 / V_{old}^2 = \left|\left(R^{-1}B\right)_{i,j-r}\right|^2 + \left\|R_{i,:}^{-1}\right\|_2^2 \left\|C_{:,j-r}\right\|_2^2 = \left|\left(\hat{A}^+ A\right)_{ij}\right|^2 + \left\|\hat{A}_{i,:}^{+}\right\|_2^2 \left\|(A - QQ^* A)_{:,j}\right\|_2^2 \qquad (4.22)$$

раз, где $A_{i,:}^{-1}$ есть i-я строка $\hat{A}^{-1}$, а $C_{:,j-r}$ – это $(j-r)$-й столбец C.

Доказательство. Без ограничения общности считаем, что новый столбец имеет индекс $r + 1$, а удаляемый столбец имеет индекс r (иначе переставим столбцы). Рассмотрим полное QR разложение расширенной матрицы $\tilde{A} \in \mathbb{C}^{m\times(r+1)}$:

$$\tilde{A} = \tilde{Q}\begin{bmatrix} \hat{R} & b \\ 0 & c \\ 0 & 0 \end{bmatrix} = \tilde{Q}\begin{bmatrix} * & \ddots & \ddots & * \\ 0 & \ddots & \ddots & \vdots \\ 0 & 0 & \hat{R}_{rr} & b_r \\ 0 & \cdots & 0 & c \\ 0 & \cdots & 0 & 0 \end{bmatrix}, \quad b \in \mathbb{C}^r, c \in \mathbb{C}.$$

Матрица $\tilde{Q} \in \mathbb{C}^{m\times m}$ никак не влияет на объемы подматриц из столбцов, так как не меняет сингулярных чисел.

Таким образом, объем матрицы $\hat{A}$ есть модуль произведения диагональных элементов $\hat{R}$. Новая матрица $\hat{A}^{\text{new}}$ будет выглядеть следующим образом:

$$\hat{A}^{\text{new}} = \begin{bmatrix} * & \ddots & \ddots & * \\ 0 & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & b_r \\ 0 & \cdots & 0 & c \\ 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Последние две строки всегда можно повернуть с помощью матрицы $U \in \mathbb{C}^{m\times m}$, содержащей 2×2 матрицу поворота в правом нижнем блоке, не затронув при этом остальные строки или сингулярные числа. Поворот выполним так, чтобы занулить строки после r-й, а элемент на диагонали был положительным:

$$UQ^*\hat{A}^{\text{new}} = \begin{bmatrix} * & \ddots & \ddots \\ 0 & \ddots & \ddots \\ 0 & 0 & \sqrt{|b_r|^2 + |c|^2} \\ 0 & \cdots & 0 \\ 0 & \cdots & 0 \end{bmatrix}.$$

Теперь видно, что раз диагональ $UQ^*\hat{A}^{\text{new}}$ отличается от $Q^*\hat{A}$ только в последнем элементе, а объем выражается как модуль произведения диагональных элементов, то

$$\mathcal{V}^2\left(\hat{A}^{\text{new}}\right)/\mathcal{V}^2\left(\hat{A}\right) = V_{new}^2/V_{old}^2 = \frac{|b_r|^2 + |c|^2}{\left|\hat{R}_{rr}\right|^2}. \tag{4.23}$$

Во-первых, заметим, что обращение оставляет матрицу $\hat{R}$ верхней треугольной, а потому

$$|b_r|^2 = \left|\hat{R}_{rr}\right|^2 \left|\left(\hat{R}^{-1}b\right)_r\right|^2. \tag{4.24}$$

Кроме того, так как $\hat{A}^+ = \hat{R}^{-1}Q^*$ и Q^* не меняет 2-нормы строк псевдообратной матрицы, то 2-норма r-й строки есть

$$\left\|\hat{A}^+_{r,:}\right\|_2^2 \left\|\hat{R}^{-1}_{r,:}\right\|_2^2 = 1/\left|\hat{R}_{rr}\right|^2. \tag{4.25}$$

Подстановка (4.24) и (4.25) в (4.23) и замена индексов r и $r+1$ на произвольные дает в итоге (4.22). □

Стоимость одного шага SRRQR с учетом возможности быстрого обновления составляет $O(Nm)$ для $m \times r$ подматриц.

4.3.1. Построение выявляющего спектр LU разложения

Покажем, как алгоритмы SRRQR и maxvol могут быть использованы для построения крестовых аппроксимаций матриц. Как мы уже упоминали в самом начале диссертации, выявляющее ранг LU разложение можно рассматривать как частный случай крестового разложения, называемого скелетным. Напомним основную теорему из [7] и простейший алгоритм достижения оценок из нее, основанный на сильном RRQR [39].

Теорема 4.2 ([7]). *Пусть подматрица $A_1 \in \mathbb{C}^{M\times r}$ матрицы $A \in \mathbb{C}^{M\times N}$ обладает ρ-локальным максимальным объемом. Пусть подматрица $\hat{A} \in \mathbb{C}^{r\times r}$ обладает ρ-локальным максимальным объемом в A_1.*

Тогда на основе столбцов A_1 и строк, соответствующих $\hat{A}$, можно построить выявляющее спектр LU разложение (определение 1.4) с

$$p_1(r,N) = p_2(r,N) = \sqrt{1+\rho^2 r(N-r)}\sqrt{1+\rho^2 r(M-r)}.$$

Кроме того,

$$\|L_{21}L_{11}^{-1}\|_C \leqslant \rho$$

и

$$\|A_1^+ A\|_C \leqslant \rho.$$

Причем найти подматрицы можно за $O\left((nnz(A)r + (M+N)r)\log_\rho r\right)$ операций, где $nnz(A)$ – число ненулевых элементов матрицы A.

Здесь мы также в явном виде указали эффективность для разреженных матриц, более точные оценки времени работы (в [7] логарифмический фактор был опущен, так как редко встречается на практике), полученные после (4.7), а также то, что разложение является выявляющим спектр.

Так как в [7] не было доказано, что разложение является выявляющим спектр, докажем это здесь. Для этого дважды применим соответствующие оценки для RRQR.

Теорема 4.3 ([39]). *Пусть подматрица $A_1 \in \mathbb{C}^{M\times r}$ матрицы $A \in \mathbb{C}^{M\times N}$ обладает ρ-локальным максимальным объемом.*

Тогда на основе столбцов A_1 можно построить сильное RRQR разложение (определение 1.3) с

$$p_1(r,N) = p_2(r,N) = \sqrt{1+\rho^2 r(N-r)}$$

и

$$\left\|R^{-1}B\right\|_C \leqslant \rho.$$

Примененяя свойство выявления спектра для QR разложения на основе столбцов A_1 ρ-локально максимального объема (теорема 4.3), получаем неравенство

$$\sigma_i(A_1) \geqslant \sigma_i(A) / \sqrt{1+\rho^2 r(N-r)}.$$

Далее, внутри столбцов A_1 подматрица $\hat{A}$ задает строки, на основе которых можно построить полное QR разложение матрицы A_1^T. Так как оно также основано на подматрице ρ-локально максимального объема (внутри A_1, но не во всей A), получаем неравенство:

$$\begin{aligned}\sigma_i(L_{11}U_{11}) &\geqslant \sigma_i\left(\begin{bmatrix} L_{11} \\ L_{21} \end{bmatrix} U_{11}\right) / \sqrt{1+\rho^2 r(M-r)} \\ &= \sigma_i\left(A_1\right) / \sqrt{1+\rho^2 r(M-r)}) \\ &\geqslant \sigma_i\left(A\right) / \left(\sqrt{1+\rho^2 r(M-r)}\sqrt{1+\rho^2 r(N-r)}\right), \end{aligned} \tag{4.26}$$

что доказывает свойство выявления спектра полученного RRLU разложения.

Численно столбцы A_1 можно найти с помощью SRRQR алгоритма [39], а подматрицу $\hat{A} \in \mathbb{C}^{r\times r}$ в уже фиксированных столбцах $A_1 \in \mathbb{C}^{M\times r}$ можно найти с помощью алгоритма maxvol [11].

Прежде, чем переходить далее, отметим, что оценки для RRQR и RRLU достижимы с $f = 1$ за $O(NM^3 r \log M)$ операций [43], однако высокая сложность и быстрое достижение $f = 1$ на практике даже со стандартным алгоритмом делает их малоинтересными. В качестве альтернативы имеет смысл использовать метод, описанный в теоремах 2.15 и 2.16.

Заметим далее, что можно построить рандомизированный алгоритм, который позволяет избавиться от лишнего $\sqrt{M}$, возникающего в оценке погрешности RRLU по сравнению с RRQR. Для этого воспользуемся «отбором по объему» (volume sampling), предложенному в [43].

Определение 4.1 ([43])**.** Отбор по объему – выбор подматрицы $\hat{A} \in \mathbb{C}^{r\times r}$ матрицы $A \in \mathbb{C}^{M\times r}$ с вероятностью, пропорциональной квадрату её объема.

В [46] предложен алгоритм, позволяющий осуществлять отбор по объему за $O(Mr^3)$.

В [83] было доказано, что отбор по объему в матрице Z, где $A = Z + F$ приводит к

$$\mathbb{E}\|A - CW\|_F \leqslant \sqrt{r+1}\|F\|_F, \tag{4.27}$$

где $W = Z_{11}^{-1}\begin{bmatrix} Z_{11} & Z_{12} \end{bmatrix}$. Если Z является RRQR разложением, то таким образом можно построить RRLU. Используя отбор по объему так же, как в [44], мы получаем следующую теорему.

Теорема 4.4. *Пусть подматрица $A_1 \in \mathbb{C}^{M\times r}$ матрицы $A \in \mathbb{C}^{M\times N}$ обладает ρ-локальным максимальным объемом. Пусть подматрица $\hat{A} \in \mathbb{C}^{rtimesr}$ выбирается $\log_\eta \delta^{-1}$ раз в матрице A_1 с помощью Volume Sampling.*

Тогда на основе столбцов A_1 и строк, соответствующих одной из выбранных подматриц $\hat{A}$ можно построить выявляющее спектр LU разложение с

$$p_1(r,N) = \sqrt{1+\rho^2 r(N-r)}\sqrt{1+\eta^2 r(M-r)}$$

и

$$p_2(r, N) = \rho\eta(r+1)\sqrt{N-r}.$$

Причем с вероятностью $1 - 2\delta$ *потребуется* $O\left((nnz(A)r + (M+N)r)\log_\rho r + Mr^3 \log_\eta \delta^{-1}\right)$ *операций, где* $nnz(A)$ *– число ненулевых элементов матрицы* A.

Среднее значение (матожидание) множителя $\eta = 1$.

Оценка для p_1 следует из того, что при отборе по объему матожидание квадрата любого элемента из $L_{21}L_{11}^{-1}$ не превосходит 1, что дает в среднем такой же множитель, как и худший для локально максимального объема.

Оценка для p_2 следует из выражений (4.28) и (4.27) и распространяется на норму Фробениуса ошибки. Следует также отметить, что на практике отбор по объему обычно не требуется, так как оценка (4.27) обычно достигается для подматриц локально максимального объема (что доказано в главе 3).

Наконец заметим, что в p_1 один из $\sqrt{r}$ тоже можно убрать.

Утверждение 4.9. *В условиях теоремы 4.4 условие выявления спектра можно заменить на выявление ранга с*

$$p_1(r, M, N) = \rho\eta\sqrt{r(M-r+1)(N-r+1)}$$

и распространить тот же коэффициент на $\left\|(L_{11}U_{11})^{-1}\right\|_F$.

Доказательство. Следует из доказанного в [44]

$$\mathbb{E}\| (L_{11}U_{11})^{-1} \|_F \leqslant \sqrt{M-r+1}\|A_1^+\|_F$$

и утверждения 4.10. □

В [39] было доказано, что множитель $\sqrt{1+\rho r(N-r)}$ справедлив также для нормы Фробениуса. Однако нам далее понадобится оценка на норму Фробениуса другого вида

Утверждение 4.10. *В условиях теоремы 4.3 верно неравенство*

$$\|R^{-1}\|_F = \|A_1^+\|_F \leqslant \sqrt{r(N-r+1)}\sigma_r^{-1}(A).$$

Доказательство. Пусть $P_r \in \mathbb{C}^{N\times N}$ – проектор на подпространство первых r правых сингулярных векторов A, то есть $AP_r = A_r$. Тогда

$$\begin{bmatrix} R^{-1} & 0 \\ 0 & \left\|R^{-1}\right\|_F I_{M-r} \end{bmatrix} AP_r = \begin{bmatrix} I_r & R^{-1}B \\ 0 & \left\|R^{-1}\right\|_F C \end{bmatrix} P_r = WP_r$$

Умножим все справа на $(AP_r)^+$. Тогда

$$\left\| \begin{bmatrix} R^{-1} & 0 \\ 0 & \left\|R^{-1}\right\|_F I_{M-r} \end{bmatrix} AP_r(AP_r)^+ \right\|_F \leqslant \|W\|_F \left\|(AP_r)^+\right\|_2$$

$$= \|W\|_F \sigma_r^{-1}(A)$$
$$\leqslant \sqrt{r(N-r+1)}\sigma_r^{-1}(A),$$

где последнее неравенство доказано в [39].

Осталось лишь заметить, что $(AP_r)^+$ – ортогональный проектор на подпространство размерности r, которое в худшем случае совпадает с первыми r строками, что дает слева $\|R^{-1}\|_F$.

Тем же методом в [39] доказываются и оценки на сингулярные числа, из которых и следует сильное RRQR. □

Можно заметить также, что

$$\|A - QR\|_F \leqslant \sqrt{f^2(r+1)(N-r)}\sigma_{r+1}(A), \tag{4.28}$$

оценив норму Фробениуса через сумму квадратов ошибок в каждом столбце, где ошибка ранга 1.

Напомним, что на практике множитель вида $O(\sqrt{MN})$ не наблюдается, поэтому имеет смысл просто применять алгоритм maxvol для попеременного поиска в строках и столбцах (описанного в следующем разделе), а оценки здесь по большей части имеют лишь теоретическую ценность.

RRLU из данного раздела, хотя и имеет гарантии точности аппроксимации по спектральной норме, тем не менее, не дает никаких гарантий по норме Чебышева, не гарантирует локальной максимальности объема найденной подматрицы во всей матрице, и не позволяет строить сильное RRLU разложение. Для построения сильного RRLU разложения нам необходим будет алгоритм поиска подматрицы локально максимального объема во всей матрице.

Можно также доказать, что RRLU из теоремы 4.2 достигает гарантий сильного RRLU, но с коэффициентами примерно в $\rho^2 r^2 M$ раз выше (что можно получить, оценив величину ρ для ρ-локально максимального объема найденной подматрицы). Тем не менее, хотя такая оценка будет полиномиальной по r, M и N, формально разложение все же не является сильным, поскольку для этого по определению требуется $\left\|\hat{A}^{-1}R\right\|_C \leqslant \mathrm{const}$, тогда как можно гарантировать лишь $\left\|\hat{A}^{-1}R\right\|_C \leqslant \rho^2 r + \left(1+\rho^2 r\right)\sqrt{1+\rho^2 r(M-r)}$.

4.4. Поиск подматриц локально максимального объема во всей матрице

4.4.1. Поиск квадратных подматриц

В [93] было показано, что неполное разложение Гаусса позволяет найти лишь подматрицу 2^{r-1}-локально максимального объема в своих строках и столбцах (данная оценка является точной с широко известным контрпримером), чего недостаточно, чтобы гарантировать высокой точности аппроксимации. Данную проблему можно решить с помощью алгоритма maxvol [11]. Здесь мы покажем, как алгоритм 4.3 из предыдущего раздела может быть использован для поиска подматрицы локально максимального объема во всей матрице. Как и ранее (см. раздел 4.3),

алгоритм начинает с некоторой подматрицы $\hat{A}$ размера $r \times r$ на пересечении столбцов $C \in \mathbb{C}^{M \times r}$ и строк $R \in \mathbb{C}^{r \times N}$ матрицы $A \in \mathbb{C}^{M \times N}$. В выбранных строках R или столбцах C применяется алгоритм 4.3.

Перед началом замен происходит вычисление $C\hat{A}^{-1}$ или $\hat{A}^{-1}R$. После этого каждая замена является обновлением ранга 1, а потому может быть вычислена быстрее, чем умножение на новое значение $\hat{A}^{-1}$. В результате, алгоритм maxvol сначала производит как можно больше замен в строках, потом в столбцах, и так далее, в результате чего умножения на $\hat{A}^{-1}$ для вычисления $C\hat{A}^{-1}$ или $\hat{A}^{-1}R$ заново довольно редки. На практике алгоритму требуется 3-4 переходы между строками и столбцами, чтобы найти подматрицу локально максимального объема (для ранга 1 такое наблюдение можно объяснить теоремой 5.1).

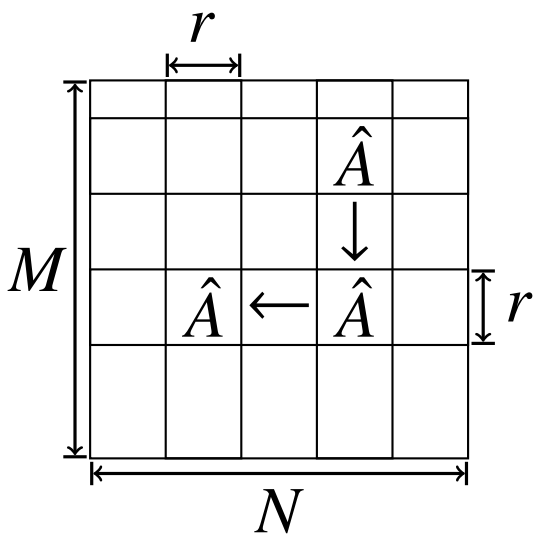


Рис. 4.1: Изменение текущей подматрицы $\hat{A} = A_{I,\mathcal{J}}$ в процессе алгоритма maxvol.

Как только никакие замены более не приводят к росту объема (согласно утверждению 4.6 все элементы $C\hat{A}^{-1}$ и $\hat{A}^{-1}R$ в этом случае не превосходят по модулю 1), формируется крестовая аппроксимация $C\hat{A}^{-1}R$.

Метод крестовой аппроксимации на основе алгоритма maxvol кратко записан как алгоритм 4.5. Стартовые строки и столбцы обычно выбираются случайно.

Несмотря на его эффективность на практике, из-за того, что maxvol «не видит» все элементы матрицы, невозможно теоретически гарантировать его скорость сходимости или точность полученной аппроксимации. В частности, для старта требуется r столбцов C полного ранга (чтобы в них нашлась невырожденная подматрица). И если таких наборов в матрице мало, то шанс выбрать подходящие столбцы на старте, не рассматривая почти всю матрицу, также мал, и алгоритм работать не будет. Если же объем стартовых столбцов (а потому и объем стартовой подматрицы) существенно ниже максимального, то это может привести к росту необходимого числа замен.

Очевидно, следует также избегать блочных матриц вида $A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$, поскольку алгоритм в таком случае «застрянет» в одном из блоков и не сможет его покинуть, поскольку не может сделать одновременную замену строки и столбца. Напомним, что оценки из [9, 10] (как и оценки теоремы 2.1) требуют того, чтобы подматрица обладала локально максимальным объемом, в том числе и с точки зрения одновременной замены одной из её строк и одного из её столбцов.

Алгоритм 4.5 maxvol [11]

Вход: Матрица $A \in \mathbb{C}^{M \times N}$, стартовые наборы индексов строк $\mathcal{I}$ и столбцов $\mathcal{J}$ размера r. Например, $\mathcal{I} = \mathcal{J} = \{1, ..., r\}$.

Выход: Факторы крестовой аппроксимация $C\hat{A}^{-1}R$ ранга r.

while были перестановки строк или столбцов **do**
  $\hat{A} := A_{\mathcal{I},\mathcal{J}}$
  $C := A_{:,\mathcal{J}}$
  while $\max\limits_{i,j} \left| \left(C\hat{A}^{-1} \right)_{ij} \right| > 1$ **do**
    $\{i, j\} = \arg\max\limits_{i,j} \left| \left(C\hat{A}^{-1} \right)_{ij} \right|$
    Замена индекса j на i в $\mathcal{I}$
    Обновление $C\hat{A}^{-1}$
  end while
  $\hat{A} := A_{\mathcal{I},\mathcal{J}}$
  $R := A_{\mathcal{I},:}$
  while $\max\limits_{i,j} \left| \left(\hat{A}^{-1}R \right)_{ij} \right| > 1$ **do**
    $\{i, j\} = \arg\max\limits_{i,j} \left| \left(\hat{A}^{-1}R \right)_{ij} \right|$
    Замена индекса i на j в $\mathcal{J}$
    Обновление $\hat{A}^{-1}R$
  end while
end while

Однако, поиск оптимальной одновременной замены уже потребует рассмотрения всей матрицы, что может быть слишком дорого с вычислительной точки зрения.

Таким образом, алгоритм 4.5 позволяет гарантировать локальную максимальность объема найденной подматрицы лишь в текущих строках и столбцах. В следующем подразделе мы рассмотрим, как можно найти подматрицу локально максимального объема во всей матрице, то есть гарантировать локальную максимальность при одновременной замене строки и столбца (см. определение 1.6), что, в свою очередь, позволит гарантировать высокую точность по норме Чебышева, в частности, согласно следствию 2.1.

4.4.2. Гарантированное достижение ρ-локально максимального объема

Принцип максимального объема широко используется для построения крестовых и столбцовых аппроксимаций матриц. Простейшим примером является адаптивный крестовый метод [4, 5], основанный на неполном разложении Гаусса. Его идея заключается в поочередном добавлении к подматрице $\hat{A} \in \mathbb{C}^{k \times k}$, $k < r$, новой строки и столбца, соответствующим максимальному

модулю ошибки в своей строке или столбце, что, тем самым, максимизирует объем расширенной $(k+1)\times(k+1)$ подматрицы. Его модификация на основе хода ладьи (rook pivoting) [6] поочередно переходит между строками и столбцами, пока не найдет элемент, являющийся максимальным одновременно в строке и в столбце. Добавление происходит до тех пор, пока не будет достигнут требуемый ранг или пока максимальная оцененная поэлементная погрешность (следующий максимальный элемент ошибки) не окажется ниже требуемой границы. Граница часто выбирается на основе максимального по модулю элемента погрешности, поскольку известно, что скелетная аппроксимация на основе подматриц локально максимального или ρ-локально максимального объема гарантирует высокую точность по норме Чебышева $\|A\|_C = \max\limits_{i,j} \left|A_{ij}\right|$ [9, 10]. Наконец, $C\hat{A}^{-1}R$ аппроксимации также можно строить на основе описанного выше алгоритма maxvol [11], который также используется при построении аппроксимаций тензоров на основе тензорных поездов [37]. Однако, существующие алгоритмы, в том числе упомянутые выше, не гарантируют локальной максимальности объема найденной подматрицы. Более того, было неясно, существует ли алгоритм, позволяющий найти подматрицу почти локально максимального объема за полиномиальное время. Здесь представлен такой алгоритм (основанный на работе [59]), доказаны оценки на число шагов в нем и точность полученной аппроксимации.

Здесь будет доказано, что подматрицу 3ρ-локально максимального объема в матрице $A \in \mathbb{C}^{M\times N}$ можно найти за $O\left(MNr\left(\log r + \log_\rho r\right)\right)$ операций, а подматрицу ρ-локально максимального объема ($\rho \leqslant 3$) за $O\left(MNr^3 \log_\rho r\right)$ операций. На основе данных подматриц возможно построение сильного RRLU разложения с гарантиями для точности аппроксимации по спектральной норме и норме Чебышева.

Впервые искать подматрицу локально максимального объема во всей матрице было предложено в [59]. При этом был доказан следующий критерий.

Лемма 4.6 ([59])**.** *Подматрица $\hat{A} = A_{11} \in \mathbb{C}^{r\times r}$ обладает ρ-локально максимальным объемом в матрице*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \in \mathbb{C}^{M\times N},$$

$$C = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} \in \mathbb{C}^{M\times r},$$

$$R = \begin{bmatrix} A_{11} & A_{12} \end{bmatrix} \in \mathbb{C}^{r\times N},$$

тогда и только тогда, когда верны следующие неравенства:

$$\|\hat{A}^{-1}R\|_C \leqslant \rho, \tag{4.29}$$

$$\|C\hat{A}^{-1}\|_C \leqslant \rho, \tag{4.30}$$

$$\max_{i,j,k,l}\left|\left(\hat{A}^{-1}R\right)_{jl}\left(C\hat{A}^{-1}\right)_{ki}+\hat{A}^{-1}_{ji}\left(A-C\hat{A}^{-1}R\right)_{kl}\right|\leqslant\rho. \tag{4.31}$$

При этом при замене i-й строки подматрицы $\hat{A}$ на k-ю строку матрицы A объем подматрицы растет в $\left|\left(C\hat{A}^{-1}\right)_{ki}\right|$ раз; при замене j-го столбца подматрицы $\hat{A}$ на l-й столбец матрицы A объем растет в $\left|\left(\hat{A}^{-1}R\right)_{jl}\right|$ раз; при одновременной замене строки и столбца объем растет в $\left|\left(\hat{A}^{-1}R\right)_{jl}\left(C\hat{A}^{-1}\right)_{ki}+\hat{A}^{-1}_{ji}\left(A-C\hat{A}^{-1}R\right)_{kl}\right|$ раз.

Неравенство (4.29) указывает, что подматрица $\hat{A}$ обладает ρ-локально максимальным объемом в строках R, при этом $\|\hat{A}^{-1}R\|_C$ есть максимально возможное увеличение объема подматрицы при замене одного из её столбцов. Аналогично, (4.30) указывает, что подматрица $\hat{A}$ обладает ρ-локально максимальным объемом в столбцах C. Наконец, (4.31) определяет, что максимальный рост объема при одновременной замене строки и столбца (столбца i на столбец k, строки j на строку l) не превосходит ρ.

Кроме того, была доказана следующая теорема.

Теорема 4.5 ([59])**.** *На основе строк $R\in\mathbb{C}^{r\times N}$ и столбцов $C\in\mathbb{C}^{M\times r}$, соответствующих подматрице $\hat{A}\in\mathbb{C}^{r\times r}$ матрицы $A\in\mathbb{C}^{M\times N}$, такой, что*

$$\begin{gathered}\|\hat{A}^{-1}R\|_C\leqslant\sqrt{\rho},\\ \|C\hat{A}^{-1}\|_C\leqslant\sqrt{\rho},\\ \|\hat{A}^{-1}\|_C\|A-C\hat{A}^{-1}R\|_C\leqslant 2\rho\end{gathered} \tag{4.32}$$

(например, достаточно того, чтобы она обладала $\sqrt{\rho}$-локально максимальным объемом во всей матрице) можно построить сильное RRLU разложение с

$$\begin{gathered}f=\sqrt{\rho},\\ p_1(r,M,N)=p_2(r,M,N)=\sqrt{(1+3\rho r(M-r))(1+3\rho r(N-r))}.\end{gathered}$$

В [59] результат был сформулирован менее строго, поэтому повторим его доказательство здесь.

Доказательство. Сначала докажем, что

$$\sigma_i\left(U_{22}\right)\leqslant\sqrt{\left(1+3\rho r\left(N-r\right)\right)\left(1+3\rho r\left(N-r\right)\right)}\sigma_{r+i}\left(A\right)=\sigma_{r+i}\left(A\right)p_2\left(r,M,N\right).$$

Как и в лемме 4.6, достаточно рассмотреть случай, когда $\hat{A}$ находится на пересечении первых r строк и столбцов, $\hat{A}=A_{11}$, причем матрица A может быть записана в блочном виде как

$$A=\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}.$$

Тогда дополнение Шура $U_{22} = A_{22} - A_{21}A_{11}^{-1}A_{12}$, а подматрица $\hat{A} = A_{11}$ соответствует строкам $R = \begin{bmatrix} A_{11} & A_{12} \end{bmatrix}$ и столбцам $C = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix}$. Погрешность соответствующей крестовой аппроксимации равна

$$A - C\hat{A}^{-1}R = \begin{bmatrix} 0 & 0 \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & U_{22} \end{bmatrix}.$$

Рассмотрим матрицу

$$D = \begin{bmatrix} abA_{11} & 0 \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}, \tag{4.33}$$

где

$$a = \sqrt{2\rho r\,(M-r)},$$
$$b = \sqrt{2\rho r\,(N-r)}.$$

Заметим, что

$$\begin{aligned} \left\|A_{11}^{-1}\right\|_2 \|A_{22} - A_{21}A_{11}^{-1}A_{12}\|_2 &= \left\|\hat{A}^{-1}\right\|_2 \|A - C\hat{A}^{-1}R\|_2 \\ &\leqslant r\left\|\hat{A}^{-1}\right\|_C \cdot \sqrt{(M-r)\,(N-r)}\|A - C\hat{A}^{-1}R\|_C \\ &\leqslant 2\rho r\sqrt{(M-r)\,(N-r)} \\ &= ab, \end{aligned} \tag{4.34}$$

где мы воспользовались условием (4.32).

Затем, используя (4.34), мы можем выписать следующие неравенства для сингулярных чисел, соответствующих подматрицам матрицы D:

$$\sigma_{\min}(abA_{11}) = ab\sigma_{\min}(A_{11}) = ab/\left\|A_{11}^{-1}\right\|_2 \geqslant \|A_{22} - A_{21}A_{11}^{-1}A_{12}\|_2.$$

Поскольку $\operatorname{rank} A_{11} = r$, первые r наибольших сингулярных чисел D (4.33) соответствуют подматрице abA_{11}, а потому для оставшихся сингулярных чисел

$$\sigma_{r+i}(D) = \sigma_i\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right).$$

Теперь представим матрицу D в виде следующего произведения, соответствующего блочному исключению Гаусса:

$$D = \begin{bmatrix} abA_{11} & 0 \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix} = \begin{bmatrix} aI & 0 \\ -A_{21}A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} bI & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix}.$$

Тогда

$$\begin{aligned} \sigma_i\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right) = \sigma_{r+i}(D) \leqslant \\ \leqslant \left\|\begin{bmatrix} aI & 0 \\ -A_{21}A_{11}^{-1} & I \end{bmatrix}\right\|_2 \sigma_{r+i}\left(\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\right) \left\|\begin{bmatrix} bI & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix}\right\|_2. \end{aligned} \tag{4.35}$$

Спектральная норма выписанных блочных матриц оценивается как

$$\begin{aligned}\left\|\begin{bmatrix} aI & 0 \\ -A_{21}A_{11}^{-1} & I \end{bmatrix}\right\|_2^2 &\leqslant \|aI\|_2^2 + \left\|-A_{21}A_{11}^{-1}\right\|_2^2 + \|I\|_2^2 \\ &\leqslant 1 + a^2 + r(M-r)\left\|A_{21}A_{11}^{-1}\right\|_C^2 \\ &\leqslant 1 + 2\rho r\,(M-r) + r\,(M-r)\,\rho \\ &= 1 + 3\rho r\,(M-r)\,.\end{aligned} \tag{4.36}$$

Аналогично,

$$\left\|\begin{bmatrix} bI & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix}\right\|_2^2 \leqslant 1 + 3\rho r\,(N-r)\,. \tag{4.37}$$

Подстановка (4.36) и (4.37) в (4.35) приводит к искомому неравенству:

$$\begin{aligned}\sigma_i\,(U_{22}) &= \sigma_i\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right) \\ &\leqslant \sqrt{(1 + 3\rho r\,(N-r))\,(1 + 3\rho r\,(N-r))}\,\sigma_{r+1}\,(A) \\ &= p_2\,(r, M, N)\,\sigma_{r+i}\,(A)\,.\end{aligned}$$

Теперь докажем свойство выявления спектра для $p_1\,(r, M, N) = p_2\,(r, M, N)$:

$$\sigma_i\,(L_{11}U_{11}) = \sigma_i\,(A_{11}) \geqslant \sigma_i\,(A)\,/\sqrt{(1 + 3\rho r\,(N-r))\,(1 + 3\rho r\,(N-r))}. \tag{4.38}$$

Для этого запишем матрицу A в виде следующего произведения:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & aI \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ 0 & \frac{1}{ab}\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right) \end{bmatrix} \begin{bmatrix} I & A_{11}^{-1}A_{12} \\ 0 & bI \end{bmatrix}.$$

Тогда

$$\begin{aligned}&\sigma_i\,(A) \leqslant \\ &\leqslant \left\|\begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & aI \end{bmatrix}\right\|_2 \sigma_i\left(\begin{bmatrix} A_{11} & 0 \\ 0 & \frac{1}{ab}\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right) \end{bmatrix}\right) \left\|\begin{bmatrix} I & A_{11}^{-1}A_{12} \\ 0 & bI \end{bmatrix}\right\|_2.\end{aligned} \tag{4.39}$$

Снова оценим спектральные нормы факторов по аналогии с уравнениями (4.36) и (4.37):

$$\left\|\begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & aI \end{bmatrix}\right\|_2^2 \leqslant 1 + 3\rho r\,(M-r)\,, \tag{4.40}$$

$$\left\|\begin{bmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & bI \end{bmatrix}\right\|_2^2 \leqslant 1 + 3\rho r\,(N-r)\,. \tag{4.41}$$

Кроме того,

$$\sigma_i\left(\begin{bmatrix} A_{11} & 0 \\ 0 & \frac{1}{ab}\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right) \end{bmatrix}\right) = \sigma_i\left(\frac{1}{ab}D\right) = \sigma_i\,(A_{11}) \tag{4.42}$$

поскольку теперь $i \leqslant r$, а ранее мы уже показали, что первые r сингулярных чисел соответствуют подматрице A_{11}. Объединяя вместе уравнения (4.39)-(4.42), получаем

$$\sigma_i(A) \leqslant \sqrt{(1+3\rho r(N-r))(1+3\rho r(N-r))}\sigma_i(A_{11}). \tag{4.43}$$

Оценка (4.38) следует прямо из (4.43). □

Однако, все равно не ясно, как найти подматрицу ρ-локально максимального объема достаточно быстро. Сами авторы в работе [14] говорит о том, что все еще не существует эффективных алгоритмов её поиска. Покажем, что оценок теоремы 4.5 возможно достичь довольно быстро.

Для начала нам понадобится результат о близости объема подматрицы локально максимального объема к объему подматрицы максимального объема.

Утверждение 4.11. *Пусть $\hat{A} \in \mathbb{C}^{r\times r}$ является подматрицей ρ-локально максимального объема матрицы $A \in \mathbb{C}^{M\times N}$, причем в своих строках и столбцах $\hat{A}$ является подматрицей $\sqrt{\rho}$-локально максимального объема. Пусть $A_M \in \mathbb{C}^{r\times r}$ является подматрицей максимального объема в A. Тогда*

$$\mathcal{V}(A_M) \leqslant \mathcal{V}(\hat{A})\left(\rho\left(2r^2+r\right)\right)^r \leqslant \mathcal{V}(\hat{A})\left(3\rho r^2\right)^r. \tag{4.44}$$

Доказательство. Начнем с наблюдения, что существует подматрица $\tilde{A} \in \mathbb{C}^{m\times n}$ с $m,n \leqslant 2r$, содержащая подматрицы $\hat{A}$ и A_M. Добавим в нее копии общих строк $\hat{A}$ и A_M, так что

$$\tilde{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \in \mathbb{C}^{2r\times 2r},$$

где $A_{11} = \hat{A}$ является подматрицей ρ-локально максимального объема, а $A_{22} = A_M$ является подматрицей максимального объема. Эти свойства сохраняются, поскольку повторное использование тех же строк и столбцов приводит к нулевому определителю.

Аналогично теореме 4.5, построим разложение

$$\tilde{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & aI \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ 0 & \frac{1}{ab}\left(A_{22}-A_{21}A_{11}^{-1}A_{12}\right) \end{bmatrix} \begin{bmatrix} I & A_{11}^{-1}A_{12} \\ 0 & bI \end{bmatrix}$$

с теми же a и b, что и ранее. В нашем случае размерам матрицы $\tilde{A}$ соответствуют $a = b = \sqrt{2\rho}r$. В результате ограничения на сингулярные числа (4.42), получаем

$$\begin{aligned} \mathcal{V}_r\left(\begin{bmatrix} A_{11} & 0 \\ 0 & \frac{1}{ab}\left(A_{22}-A_{21}A_{11}^{-1}A_{12}\right) \end{bmatrix}\right) &= \prod_{i=1}^{r}\left(\begin{bmatrix} A_{11} & 0 \\ 0 & \frac{1}{ab}\left(A_{22}-A_{21}A_{11}^{-1}A_{12}\right) \end{bmatrix}\right) \\ &\leqslant \prod_{i=1}^{r}(A_{11}) = \mathcal{V}_r(A_{11}) = \mathcal{V}_r\left(\hat{A}\right). \end{aligned} \tag{4.45}$$

Оценим также проективный объем последних строк $\tilde{A}$

$$\begin{aligned}\mathcal{V}_r\left(\begin{bmatrix} A_{21} & A_{22} \end{bmatrix}\right) &= \mathcal{V}\left(\begin{bmatrix} A_{21} & A_{22} \end{bmatrix}\right) \\ &\leqslant \left(\left\|\begin{bmatrix} A_{21} & A_{22} \end{bmatrix}\right\|_F^2 / r\right)^{r/2} \\ &\leqslant \left(\rho r + a^2\right)^{r/2} \\ &= \left(\rho\left(2r^2 + r\right)\right)^{r/2}.\end{aligned} \tag{4.46}$$

и последних столбцов $\tilde{A}$

$$\begin{aligned}\mathcal{V}_r\left(\begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix}\right) &= \mathcal{V}\left(\begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix}\right) \\ &\leqslant \left(\left\|\begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix}\right\|_F^2 / r\right)^{r/2} \\ &\leqslant \left(\rho r + b^2\right)^{r/2} \\ &= \left(\rho\left(2r^2 + r\right)\right)^{r/2}.\end{aligned} \tag{4.47}$$

Используя утверждение 1.3, получаем

$$\begin{aligned}\mathcal{V}(A_M) = \mathcal{V}_r(A_M) &= \mathcal{V}_r\left(\begin{bmatrix} A_{21}A_{11}^{-1} & aI \end{bmatrix}\begin{bmatrix} A_{11} & 0 \\ 0 & \frac{1}{ab}\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right) \end{bmatrix}\begin{bmatrix} A_{11}^{-1}A_{12} \\ bI \end{bmatrix}\right) \\ &\leqslant \mathcal{V}_r\left(\begin{bmatrix} A_{21}A_{11}^{-1} & aI \end{bmatrix}\right)\mathcal{V}_r\left(\begin{bmatrix} A_{11} & 0 \\ 0 & \frac{1}{ab}\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right) \end{bmatrix}\right)\mathcal{V}_r\left(\begin{bmatrix} A_{11}^{-1}A_{12} \\ bI \end{bmatrix}\right).\end{aligned}$$

Подставляя (4.45), (4.46) и (4.47), получаем требуемое неравенство (4.44). □

В алгоритме далее мы будем начинать поиск с $r \times r$ подматрицы, полученной с помощью исключения Гаусса с полным выбором ведущего элемента. Нам понадобится следующий результат.

Лемма 4.7 ([93])**.** *Пусть строки подматрицы $\hat{A} \in \mathbb{R}^{r\times r}$ получены с помощью исключения Гаусса с частичным выбором ведущего элемента в столбцах $C \in \mathbb{R}^{M\times r}$. Тогда*

$$\left\|C\hat{A}^{-1}\right\|_C \leqslant 2^{r-1}.$$

Очевидно, тот же результат справедлив и для полного выбора. В этом случае также справедливо неравенство $\left\|\hat{A}^{-1}R\right\|_C \leqslant 2^{r-1}$. Более того, в [94] был доказан следующий результат.

Лемма 4.8 ([94]). *Пусть строки $R \in \mathbb{R}^{M\times r}$ и столбцы $C \in \mathbb{R}^{r\times N}$ соответствуют подматрице $\hat{A} \in \mathbb{R}^{r\times r}$ в первых r строках и столбцах, содержащей первые r ведущих элементов, выбранных в ходе исключения Гаусса с полным выбором ведущих элементов в матрице $A \in \mathbb{R}^{M\times N}$. Тогда*

$$\left\|A - C\hat{A}^{-1}R\right\|_C \leqslant 4^r \gamma_r \sigma_{r+1}(A),$$

где $\gamma_r \leqslant 2\sqrt{r+1}(r+1)^{\frac{\ln(r+1)}{4}}$ – фактор роста [95].

Следствие 4.2. В условиях леммы 4.8,

$$\left\|A - C\hat{A}^{-1}R\right\|_C \leqslant 4^{r-1/2} r^{\frac{2+\ln r}{4}} \left\|\hat{A}^{-1}\right\|_2^{-1}.$$

Доказательство. Рассмотрим подматрицу $\bar{A} \in \mathbb{R}^{(r-1)\times(r-1)}$, полученную после $r-1$ шага исключения Гаусса. Применив лемму 4.8 к матрице $\hat{A}$ (вместо A), после $r-1$ шага (вместо r) получим

$$\left\|\hat{A} - \hat{C}\bar{A}^{-1}\hat{R}\right\|_C \leqslant 4^{r-1}\gamma_{r-1}\sigma_r\left(\hat{A}\right),$$

где $\hat{R} \in \mathbb{R}^{(r-1)\times r}$ и $\hat{C} \in \mathbb{R}^{r\times(r-1)}$ – это строки и столбцы $\hat{A}$, соответствующие $\bar{A}$.

Тогда, по определению фактора роста, мы можем оценить погрешность на r-м шаге исключения Гаусса:

$$\begin{aligned}
\left\|A - C\hat{A}^{-1}R\right\|_C &\leqslant \gamma_1 \left\|\hat{A} - \hat{C}\bar{A}^{-1}\hat{R}\right\|_C \\
&\leqslant 2\left\|\hat{A} - \hat{C}\bar{A}^{-1}\hat{R}\right\|_C \\
&\leqslant 4^{r-1/2}\gamma_{r-1}\sigma_r\left(\hat{A}\right) \\
&\leqslant 4^{r-1/2} r^{\frac{2+\ln r}{4}} \sigma_r\left(\hat{A}\right) \\
&= 4^{r-1/2} r^{\frac{2+\ln r}{4}} \left\|\hat{A}^{-1}\right\|_2^{-1}.
\end{aligned}$$

□

Вместе лемма 4.7 и следствие 4.2 говорят о том, что исключение Гаусса с полным выбором ведущего элемента позволяет найти подматрицу ρ-локально максимального объема, хоть и с большим ρ.

Следствие 4.3. Пусть строки $R \in \mathbb{R}^{M\times r}$ и столбцы $C \in \mathbb{R}^{r\times N}$ соответствуют подматрице $\hat{A} \in \mathbb{R}^{r\times r}$ в первых r строках и столбцах, содержащей первые r ведущих элементов, выбранных в ходе исключения Гаусса с полным выбором ведущих элементов в матрице $A \in \mathbb{R}^{M\times N}$. Тогда $\hat{A}$ – подматрица ρ-локально максимального объема с

$$\rho \leqslant 3 \cdot 4^{r-1} r^{\frac{2+\ln r}{4}}.$$

Доказательство. Сравнивая лемму 4.7 с условиями (4.29) и (4.30) леммы 4.6, мы видим, что они верны для $\rho = 2^{r-1}$. Теперь проверим последнее условие, куда мы подставим неравенства,

полученные в лемме 4.7 и следствии 4.2:

$$\begin{aligned}
&\max_{i,j,k,l}\left|\left(\hat{A}^{-1}R\right)_{ik}\left(C\hat{A}^{-1}\right)_{jl}+\hat{A}^{-1}_{ij}\left(A-C\hat{A}^{-1}R\right)_{lk}\right| \leqslant \\
&\leqslant \|C\hat{A}^{-1}\|_C\|\hat{A}^{-1}R\|_C+\|\hat{A}^{-1}\|_C\|A-C\hat{A}^{-1}R\|_C \\
&\leqslant 4^{r-1}+4^{r-1/2}r^{\frac{2+\ln r}{4}} \\
&\leqslant 3\cdot 4^{r-1}r^{\frac{2+\ln r}{4}}.
\end{aligned}$$

□

Наконец, мы готовы перейти к поиску подматрицы локально максимального объема. Для достижения 3ρ-локально максимального объема мы воспользуемся алгоритмом 4.6.

Поиск начинается после исключения Гаусса: алгоритм заменяет строки и столбцы на основе оценок роста объема ρ_1, ρ_2 и ρ_3. Эти оценки соответствуют росту объема при замене строки, столбца и одновременной замене строки и столбца соответственно. Заметим, что после того, как алгоритм остановился, верны неравенства

$$\begin{aligned}
\left\|\hat{A}^{-1}R\right\|_C &\leqslant \sqrt{\rho}, \\
\left\|C\hat{A}^{-1}\right\|_C &\leqslant \sqrt{\rho}, \\
\left\|\hat{A}^{-1}\right\|_C\left\|A-C\hat{A}^{-1}R\right\|_C &\leqslant 2\rho.
\end{aligned} \tag{4.48}$$

Первые два неравенства следуют из $\max(\rho_1,\rho_2)\leqslant\sqrt{\rho}$ (неравенство в строке 15 алгоритма не выполнено, если произошел «break»), а третье следует из

$$\begin{aligned}
\rho_3 &= \left|\left(\hat{A}^{-1}R\right)_{j_3l_3}\left(C\hat{A}^{-1}\right)_{k_3i_3}+\hat{A}^{-1}_{j_3i_3}\left(A-C\hat{A}^{-1}R\right)_{k_3l_3}\right| \\
&\geqslant \left|\hat{A}^{-1}_{j_3i_3}\left(A-C\hat{A}^{-1}R\right)_{k_3l_3}\right|-\left|\left(\hat{A}^{-1}R\right)_{j_3l_3}\left(C\hat{A}^{-1}\right)_{k_3i_3}\right| \\
&= \max_{i,j}\left|\hat{A}^{-1}_{ji}\right|\max_{k,l}\left|\left(A-C\hat{A}^{-1}R\right)_{kl}\right|-\left|\left(\hat{A}^{-1}R\right)_{j_3l_3}\left(C\hat{A}^{-1}\right)_{k_3i_3}\right| \\
&\geqslant \max_{i,j}\left|\hat{A}^{-1}_{ji}\right|\max_{k,l}\left|\left(A-C\hat{A}^{-1}R\right)_{kl}\right|-\max_{j_2l_2}\left|\left(\hat{A}^{-1}R\right)_{j_2l_2}\right|\max_{i_1k_1}\left|\left(C\hat{A}^{-1}\right)_{i_1k_1}\right| \\
&= \max_{i,j}\left|\hat{A}^{-1}_{ji}\right|\max_{k,l}\left|\left(A-C\hat{A}^{-1}R\right)_{kl}\right|-\rho_2\rho_1 \\
&= \left\|\hat{A}^{-1}\right\|_C\left\|A-C\hat{A}^{-1}R\right\|_C-\rho_2\rho_1,
\end{aligned}$$

а значит

$$\left\|\hat{A}^{-1}\right\|_C\left\|A-C\hat{A}^{-1}R\right\|_C\leqslant\rho_2\rho_1+\rho_3\leqslant\rho+\sqrt{\rho}\leqslant 2\rho.$$

Из условий (4.48) также следует, что найденная подматрица обладает 3ρ-локально максимальным объемом. Для доказательства этого факта достаточно проверить условия (4.29)-(4.31) леммы 4.6. Первые два следуют напрямую из (4.48), а последнее (с коэффициентом 3ρ вместо ρ)

Алгоритм 4.6 Поиск подматрицы 3ρ-локально максимального объема

Вход: Матрица $A \in \mathbb{R}^{M\times N}$, требуемый ранг r, параметр $\rho \geqslant 1$.

Выход: Подматрица $\hat{A} \in \mathbb{R}^{r\times r}$ 3ρ-локально максимального объема во всей матрице A.

1: Начальная подматрица $\hat{A} \in \mathbb{R}^{r\times r}$ выбирается с помощью r шагов исключения Гаусса с полным выбором ведущего элемента
2: Пусть $\hat{A}$ находится на пересечении строк $R \in \mathbb{R}^{r\times N}$ и столбцов $C \in \mathbb{R}^{M\times r}$
3: **loop**
4: Оценка роста объема при замене строки:
5: $i_1, k_1 = \arg\max_{i,k} \left|\left(C\hat{A}^{-1}\right)_{ki}\right|$, $j_1 = l_1 = 1$
6: $\rho_1 = \max_{i,k} \left|\left(C\hat{A}^{-1}\right)_{ki}\right|$
7: Оценка роста объема при замене столбца:
8: $j_2, l_2 = \arg\max_{j,l} \left|\left(\hat{A}^{-1}R\right)_{jl}\right|$, $i_2 = k_2 = 1$
9: $\rho_2 = \max_{j,l} \left|\left(\hat{A}^{-1}R\right)_{jl}\right|$
10: Оценка роста объема при одновременной замене строки и столбца:
11: $i_3, j_3, k_3, l_3 = \arg\max_{i,j,k,l} \left|\hat{A}^{-1}_{ji}\right| \left|\left(A - C\hat{A}^{-1}R\right)_{kl}\right|$
12: $\rho_3 = \left|\left(\hat{A}^{-1}R\right)_{j_3l_3}\left(C\hat{A}^{-1}\right)_{k_3i_3} + \hat{A}^{-1}_{j_3i_3}\left(A - C\hat{A}^{-1}R\right)_{k_3l_3}\right|$
13: Выбор максимального роста:
14: $\rho_n = \max(\rho_1, \rho_2, \rho_3)$, $n \in \{1, 2, 3\}$.
15: **if** $\rho_n > \sqrt{\rho}$ **then**
16: Перестановка строк i_n и k_n
17: Перестановка столбцов j_n и l_n
18: Обновление $\hat{A}^{-1}$, $C\hat{A}^{-1}$, $\hat{A}^{-1}R$ и $A - C\hat{A}^{-1}R$
19: **else**
20: **break**
21: **end if**
22: **end loop**

следует из

$$\begin{aligned}
&\max_{i,j,k,l} \left|\left(\hat{A}^{-1}R\right)_{jl}\left(C\hat{A}^{-1}\right)_{ki} + \hat{A}^{-1}_{ji}\left(A - C\hat{A}^{-1}R\right)_{kl}\right| \leqslant \\
&\leqslant \max_{i,j,k,l} \left|\left(\hat{A}^{-1}R\right)_{jl}\left(C\hat{A}^{-1}\right)_{ki}\right| + \max_{i,j,k,l} \left|\hat{A}^{-1}_{ji}\left(A - C\hat{A}^{-1}R\right)_{kl}\right| \\
&\leqslant \left\|\hat{A}^{-1}R\right\|_C \left\|C\hat{A}^{-1}\right\|_C + \left\|\hat{A}^{-1}\right\|_C \left\|A - C\hat{A}^{-1}R\right\|_C \leqslant \rho + 2\rho \leqslant 3\rho.
\end{aligned}$$

Алгоритм 4.6 также непременно останавливается, поскольку на каждом шаге объем растет на

коэффициент $\max(\rho_1, \rho_2, \rho_3) > \sqrt{\rho} \geqslant 1$, а потому не может рассмотреть одну и ту же подматрицу дважды.

Если мы хотим достичь ρ-локально максимального объема, достаточно производить замены напрямую согласно условиям леммы 4.6. Из нее следует, что представленный далее алгоритм 4.7 останавливается, когда достигает подматрицы ρ-локально максимального объема.

Алгоритм 4.7 Поиск подматрицы ρ-локально максимального объема

Вход: Матрица $A \in \mathbb{R}^{M \times N}$, требуемый ранг r, параметр $\rho \geqslant 1$.

Выход: Подматрица $\hat{A} \in \mathbb{R}^{r \times r}$ ρ-локально максимального объема во всей матрице A.

1: Начальная подматрица $\hat{A} \in \mathbb{R}^{r \times r}$ выбирается с помощью r шагов исключения Гаусса с полным выбором ведущего элемента
2: Пусть $\hat{A}$ находится на пересечении строк $R \in \mathbb{R}^{r \times N}$ и столбцов $C \in \mathbb{R}^{M \times r}$
3: **loop**
4: Оценка роста объема (разрешено использовать $i = k$ и $j = l$):
5: $i, j, k, l = \underset{i,j,k,l}{\arg\max} \left| \left(\hat{A}^{-1}R\right)_{jl} \left(C\hat{A}^{-1}\right)_{ki} + \hat{A}^{-1}_{ji} \left(A - C\hat{A}^{-1}R\right)_{kl} \right|$
6: $\rho_{\max} = \left| \left(\hat{A}^{-1}R\right)_{jl} \left(C\hat{A}^{-1}\right)_{ki} + \hat{A}^{-1}_{ji} \left(A - C\hat{A}^{-1}R\right)_{kl} \right|$
7: **if** $\rho_{\max} > \rho$ **then**
8: Перестановка строк i_n и k_n
9: Перестановка столбцов j_n и l_n
10: Обновление $\hat{A}^{-1}$, $C\hat{A}^{-1}$, $\hat{A}^{-1}R$ и $A - C\hat{A}^{-1}R$
11: **else**
12: **break**
13: **end if**
14: **end loop**

Теперь оценим число шагов и вычислительную сложность достижения 3ρ-локально максимального объема и ρ-локально максимального объема.

Утверждение 4.12. *Алгоритм 4.6 останавливается после* $O\left(r\left(1 + \frac{1}{\log \rho}\right)\log r\right)$ *итераций и имеет вычислительную сложность* $O\left(MNr\left(1 + \frac{1}{\log \rho}\right)\log r\right)$.

Алгоритм 4.7 останавливается после $O\left(r \log_\rho r\right)$ *итераций и имеет вычислительную сложность* $O\left(MNr^3\left(1 + \frac{1}{\log \rho}\right)\log r\right)$.

Доказательство. Выберем стартовую подматрицу $A_{(0)}$ с помощью метода Гаусса с полным выбором ведущего элемента. Согласно следствию 4.3, $A_{(0)}$ обладает ρ-локально максимальным объемом с

$$\rho \leqslant 3 \cdot 4^{r-1} r^{\frac{2+\ln r}{4}}, \tag{4.49}$$

причем в своих строках и столбцах согласно лемме 4.7 имеем $2^{r-1} \leqslant \sqrt{\rho}$-локально максимальный объем.

Рассмотрим последовательность подматриц $\hat{A} = A_{(s)}$, полученных после s замен. Согласно лемме 4.11,

$$\begin{aligned} \mathcal{V}(A_M) &\leqslant \mathcal{V}(A_{(s)}) \left(3\rho_{(s)} r^2\right)^r, \\ \rho_{(s)} &\geqslant \frac{1}{3r^2} \left(\frac{\mathcal{V}(A_M)}{\mathcal{V}(A_{(s)})}\right)^{\frac{1}{r}}, \end{aligned} \tag{4.50}$$

где $A_M \in \mathbb{R}^{r\times r}$ – подматрица максимального объема.

На каждом шаге алгоритма 4.6 мы совершаем замену согласно следующему правилу. Мы выбираем новый столбец k и/или новую строку l, которые соответствуют

$$\begin{aligned} \rho' = \max &\left(\left|\left(\hat{A}^{-1}R\right)_{j_2 l_2}\right|, \left|\left(C\hat{A}^{-1}\right)_{k_1 i_1}\right|,\right. \\ &\left.\left|\hat{A}^{-1}_{j_3 i_3}\left(A - C\hat{A}^{-1}R\right)_{k_3 l_3} + \left(\hat{A}^{-1}R\right)_{j_3 l_3}\left(C\hat{A}^{-1}\right)_{k_3 i_3}\right|\right) \\ \geqslant \max &\left(\max_{j,l}\left|\left(\hat{A}^{-1}R\right)_{jl}\right|, \max_{i,k}\left|\left(C\hat{A}^{-1}\right)_{ki}\right|,\right. \\ &\left.\max_{i,j}\left|\hat{A}^{-1}_{ji}\right| \max_{k,l}\left|\left(A - C\hat{A}^{-1}R\right)_{kl}\right| - \left|\left(\hat{A}^{-1}R\right)_{j_3 l_3}\right|\left|\left(\hat{C}A^{-1}\right)_{k_3 i_3}\right|\right) \\ \geqslant \max &\left(\max_{j,l}\left|\left(\hat{A}^{-1}R\right)_{jl}\right|, \max_{i,k}\left|\left(C\hat{A}^{-1}\right)_{ki}\right|,\right. \\ &\left.\max_{i,j}\left|\hat{A}^{-1}_{ji}\right| \max_{k,l}\left|\left(A - C\hat{A}^{-1}R\right)_{kl}\right| - \max_{j_2,l_2}\left|\left(\hat{A}^{-1}R\right)_{j_2 l_2}\right| \max_{i_1,k_1}\left|\left(\hat{C}A^{-1}\right)_{k_1 i_1}\right|\right). \end{aligned} \tag{4.51}$$

куда мы подставили определения индексов $i_1, k_1, j_2, k_2, i_3, j_3, k_3, l_3$ из алгоритма 4.6.

Такой выбор увеличит объем минимум в $\rho' = \max(\rho_1, \rho_2, \rho_3)$ раз.

Кроме того, если на s-м шаге мы имеем подматрицу в точности $\rho_{(s)}$-локально максимального объема с $\rho_{(s)} \geqslant 3$, то $\rho' \geqslant \sqrt{\rho_{(s)}/3}$. Действительно, если $\max\limits_{j,l}\left|\hat{A}^{-1}R\right|_{jl} < \sqrt{\rho_{(s)}/3}$ и $\max\limits_{i,k}\left|C\hat{A}^{-1}\right|_{ki} < \sqrt{\rho_{(s)}/3}$, то

$$\begin{aligned} &\max_{i,j}\left|\hat{A}^{-1}\right|_{ji} \max_{k,l}\left|A - C\hat{A}^{-1}R\right|_{kl} - \max_{j_2,l_2}\left|\left(\hat{A}^{-1}R\right)_{j_2 l_2}\right| \max_{i_1,k_1}\left|\left(C\hat{A}^{-1}\right)_{k_1 i_1}\right| \geqslant \\ &\geqslant \max_{i,j,k,l}\left|\hat{A}^{-1}_{ji}\left(A - C\hat{A}^{-1}R\right)_{kl} + \left(\hat{A}^{-1}R\right)_{jl}\left(C\hat{A}^{-1}\right)_{ki}\right| \\ &- 2\max_{j_2,l_2}\left|\left(\hat{A}^{-1}R\right)_{j_2 l_2}\right| \max_{i_1,k_1}\left|\left(C\hat{A}^{-1}\right)_{k_1 i_1}\right| \\ &\geqslant \rho_{(s)} - 2\frac{\rho_{(s)}}{3} = \rho_{(s)}/3 \geqslant \sqrt{\rho_{(s)}/3}. \end{aligned}$$

С учетом (4.50) получаем, что отношение к максимальному объему меняется как

$$\frac{\mathcal{V}\left(A_M\right)}{\mathcal{V}\left(A_{(s+1)}\right)} \leqslant \frac{\mathcal{V}\left(A_M\right)}{\mathcal{V}\left(A_{(s)}\right)} \cdot \sqrt{\frac{3}{\rho_{(s)}}} \leqslant 3r\left(\frac{\mathcal{V}(A_M)}{\mathcal{V}(A_{(s)})}\right)^{1-\frac{1}{2r}}. \tag{4.52}$$

Обозначим

$$\alpha_{(s)} = \frac{\left(\frac{\mathcal{V}(A_M)}{\mathcal{V}\left(A_{(s)}\right)}\right)^{\frac{1}{2r}}}{3r}. \tag{4.53}$$

Из (4.49) и леммы 4.11

$$\alpha_{(0)} \leqslant \sqrt{\rho_{(0)}/3} \leqslant 2^{r-1} r^{\frac{2+\ln r}{8}} \leqslant 2^r r^{\frac{2+\ln r}{8}}. \tag{4.54}$$

Подставив $\alpha_{(s)}$ (4.53) в уравнение (4.52), оценим как $\alpha_{(s)}$ меняется на каждом шаге:

$$\begin{aligned} \left(9r^2\right)^r \alpha_{(s+1)}^{2r} &\leqslant \left(9r^2\right)^r \alpha_{(s)}^{2r-1}, \\ \alpha_{(s+1)} &\leqslant \alpha_{(s)}^{1-\frac{1}{2r}}, \\ \alpha_{(s)} &\leqslant \alpha_{(0)}^{\left(1-\frac{1}{2r}\right)^s}, \\ \ln\ln\alpha_{(s)} - \ln\ln\alpha_{(0)} &\leqslant s\ln\left(1-\frac{1}{2r}\right) \leqslant -\frac{s}{2r}. \end{aligned} \tag{4.55}$$

Таким образом, чтобы достичь, к примеру, $\alpha_{(s)} \leqslant 3^{1/2} < 2^r r^{\frac{2+\ln r}{8}}$, понадобится

$$\begin{aligned} s_1 &\leqslant 2r\ln\ln\left(2^r r^{\frac{2+\ln r}{8}}\right) - 2r\ln\ln 3^{1/2} + 1 \\ &= O\left(r\log r\right) \end{aligned} \tag{4.56}$$

шагов. После этого получаем

$$\frac{\mathcal{V}\left(A_M\right)}{\mathcal{V}\left(A_{(s_1)}\right)} \leqslant \left(3r\alpha_{(s_1)}\right)^{2r} \leqslant \left(27r^2\right)^r.$$

Затем, пока $\rho' \geqslant \sqrt{\rho}$, алгоритм 4.6 увеличивает объем хотя бы в $\sqrt{\rho}$ раз на каждом шаге. Так как мы не можем превзойти максимальный объем, алгоритму потребуется еще не больше

$$s_2 \leqslant \log_{\sqrt{\rho}}\left(27r^2\right)^r = O\left(r\frac{\log r}{\log\rho}\right) \tag{4.57}$$

шагов до остановки.

Каждый шаг алгоритма – это обновления ранга не больше 2 для подматриц $\hat{A}$, C и R, так как идет замена не более одной строки и не более одного столбца. Отсюда следует, что и пересчет $C\hat{A}^{-1}$, $\hat{A}^{-1}R$ и $C\hat{A}^{-1}R$ также малоранговый и требует в сумме не более $O(MN)$ операций, после чего выбор индексов i, j, k, l, соответствующих ρ' (4.51), также займет не более $O\left(MN\right)$ операций. С учетом ограничений (4.56) и (4.57), общее число шагов $s_1+s_2 = O\left(r\left(1+\frac{1}{\log\rho}\right)\log r\right)$, а общая вычислительная сложность алгоритма $O\left(MNr\left(1+\frac{1}{\log\rho}\right)\log r\right)$. При этом исключение Гаусса с полным выбором ведущего элемента занимает $O\left(MNr\right)$ операций, что не влияет на общую асимптотику.

Если требуется найти подматрицу ρ-локально максимального объема с $\rho \leqslant 3$, можно вместо ρ' использовать условие

$$\max_{i,j,k,l} \left| \left(\hat{A}^{-1}R\right)_{jl} \left(C\hat{A}^{-1}\right)_{ki} + \hat{A}^{-1}_{ji} \left(A - C\hat{A}^{-1}R\right)_{kl} \right| > \rho, \tag{4.58}$$

что и делает алгоритм 4.7.

Так как в этом случае объем растет не менее, чем в ρ' раз (уравнение (4.58) также включает индексы, соответствующие ρ'), общее число шагов не вырастет. Но теперь на каждом шаге необходимо проверять все MNr^2 вариантов четверки i, j, k, l, что увеличивает сложность каждого шага в r^2 раз. Таким образом, итоговая сложность поиска ρ-локально максимального объема составляет $O\left(MNr^3\left(1 + \frac{1}{\log \rho}\right)\log r\right)$. При $\rho \leqslant 3$ оценка упрощается до $O\left(MNr^3 \log_\rho r\right)$. □

Заметим также, что, как показано в [10], крестовая аппроксимация с помощью подматрицы ρ-локально максимального объема позволяет достичь высокой точности аппроксимации по норме Чебышева. Сформулируем данный факт в терминах LU разложения.

Теорема 4.6 ([10])**.** *Пусть неполное LU разложение* (1.1) *построено на основе подматрицы* $\hat{A} = L_{11}U_{11} \in \mathbb{C}^{r\times r}$ *ρ-локально максимального объема. Тогда*

$$\|A - LU\|_C \leqslant \rho\,(r+1)^2\,\|E\|_C\,,$$

где E – матрица погрешности наилучшего приближения по норме Чебышева.

В нашем случае для достижения оценки теоремы 4.6 не требуется достижения ρ-локально максимального объема, хотя может потребоваться достижение 3ρ-локально максимального объема. Данный факт следует из утверждения ниже, основанного на алгоритме 4.8, где мы более точно оцениваем значение ρ_3.

Утверждение 4.13. *Алгоритм 4.8 требует не более $O\left(MNr\left(1 + \frac{1}{\log\rho}\right)\log r\right)$ операций, гарантирует те же оценки, что и алгоритм 4.6 и дополнительно позволяет строить неполное LU разложение на основе найденной подматрицы, для которого*

$$\|A - LU\|_C \leqslant \rho(r+1)^2\|E\|_C. \tag{4.59}$$

Доказательство. В алгоритме 4.8, вычисление ρ_3 позволяет дополнительно проверять условие

$$\left| \left(\hat{A}^{-1}R\right)_{jl_3} \left(C\hat{A}^{-1}\right)_{k_3 i} + \hat{A}^{-1}_{ji} \left(A - C\hat{A}^{-1}R\right)_{k_3 l_3} \right| \leqslant \sqrt{\rho} \tag{4.60}$$

для всех $i, j \in \overline{1, r}$, но при этом только для

$$k_3, l_3 = \arg\max_{k,l} \left| \left(A - C\hat{A}^{-1}R\right)_{kl} \right|.$$

Алгоритм 4.8 Поиск подматрицы 3ρ-локально максимального объема с лучшими гарантиями по норме Чебышева

Вход: Матрица $A \in \mathbb{R}^{M\times N}$, требуемый ранг r, параметр $\rho \geqslant 1$.

Выход: Подматрица $\hat{A} \in \mathbb{R}^{r\times r}$ 3ρ-локально максимального объема во всей матрице A.

Начальная подматрица $\hat{A} \in \mathbb{R}^{r\times r}$ выбирается с помощью r шагов исключения Гаусса с полным выбором ведущего элемента
Пусть $\hat{A}$ находится на пересечении строк $R \in \mathbb{R}^{r\times N}$ и столбцов $C \in \mathbb{R}^{M\times r}$
loop
    Оценка роста объема при замене строки:
    $i_1, k_1 = \arg\max_{i,k} \left|\left(C\hat{A}^{-1}\right)_{ki}\right|$, $j_1 = l_1 = 1$
    $\rho_1 = \max_{i,k} \left|\left(C\hat{A}^{-1}\right)_{ki}\right|$
    Оценка роста объема при замене столбца:
    $j_2, l_2 = \arg\max_{j,l} \left|\left(\hat{A}^{-1}R\right)_{jl}\right|$, $i_2 = k_2 = 1$
    $\rho_2 = \max_{j,l} \left|\left(\hat{A}^{-1}R\right)_{jl}\right|$
    Оценка роста объема при одновременной замене строки и столбца:
    $k_3, l_3 = \arg\max_{k,l} \left|\left(A - C\hat{A}^{-1}R\right)_{kl}\right|$
    $i_3, j_3 = \arg\max_{i,j} \left|\left(\hat{A}^{-1}R\right)_{jl_3}\left(C\hat{A}^{-1}\right)_{k_3i} + \hat{A}^{-1}_{ji}\left(A - C\hat{A}^{-1}R\right)_{k_3l_3}\right|$
    $\rho_3 = \left|\left(\hat{A}^{-1}R\right)_{j_3l_3}\left(C\hat{A}^{-1}\right)_{k_3i_3} + \hat{A}^{-1}_{j_3i_3}\left(A - C\hat{A}^{-1}R\right)_{k_3l_3}\right|$
    Выбор максимального роста:
    $\rho_n = \max(\rho_1, \rho_2, \rho_3)$, $n \in \{1, 2, 3\}$.
    if $\rho_n > \sqrt{\rho}$ **then**
        Перестановка строк i_n и k_n
        Перестановка столбцов j_n и l_n
        Обновление $\hat{A}^{-1}$, $C\hat{A}^{-1}$, $\hat{A}^{-1}R$ и $A - C\hat{A}^{-1}R$
    else
        break
    end if
end loop

Если при данных k_3 и l_3 можно найти i и j такие, что условие (4.60) нарушено, рост объема при такой замене составит хотя бы $\sqrt{\rho}$, согласно лемме 4.6. Только если этот рост выше всех прочих рассматриваемых замен, соответствующих ρ' (4.51), мы производим данную замену. Сразу отметим, что стоимость алгоритма от этого не возрастет, так как все неравенства для

количества шагов остаются верными, а стоимость проверки всех i и j для данных k_3 и l_3 составляет дополнительно $O(MN)$ на поиск k_3 и l_3, и еще $O(r^2)$ на поиск соответствующих им оптимальных i и j, что в сумме не увеличивает стоимость одного шага алгоритма.

Теперь покажем, что, когда алгоритм остановился, выполнено неравенство (4.59). Будем рассуждать от противного: если оно не выполнено, то оно не выполнено для тех k и l, где C-норма погрешности достигает максимума. Именно эту строку k_3 и столбец l_3 мы и проверяем. Рассмотрим подматрицу $\tilde{A} \in \mathbb{C}^{(r+1)\times(r+1)}$, которая является расширением подматрицы $\hat{A}$ данным строкой и столбцом. Если (4.59) неверно, то

$$\left\|\tilde{A} - \tilde{L}\tilde{U}\right\|_C > \rho(r+1)^2\|E\|_C \geqslant \rho(r+1)^2\|\tilde{E}\|_C,$$

где $\|\tilde{E}\|_C$ – погрешность наилучшего приближение $\tilde{A}$ по C-норме, подматрицы $\tilde{L} \in \mathbb{C}^{(r+1)\times r}$ и $\tilde{U} \in \mathbb{C}^{r\times(r+1)}$ матриц L и U соответствуют столбцам и столбцам подматрицы $\tilde{A}$. Тогда из теоремы 4.6 получаем, что подматрица $\hat{A}$ не обладает ρ-локально максимальным объемом внутри $\tilde{A}$, а значит существует замена, включающая только строку k_3 или только столбец l_3, или их обоих, которая увеличивает объем более, чем в ρ раз, что противоречит дополнительной проверке (4.60), которую мы ввели в алгоритм. Данное противоречие показывает, что неравенство (4.59) на самом деле выполнено после остановки алгоритма. □

Замечание 4.2. Чуть более точная оценка с учетом $\left\|C\hat{A}^{-1}\right\|_C \leqslant \sqrt{\rho}$ и $\left\|\hat{A}^{-1}R\right\|_C \leqslant \sqrt{\rho}$ (для ρ_3 можно оставить сравнение с ρ вместо $\sqrt{\rho}$) приводит к

$$\|A - LU\|_C \leqslant \left(\sqrt{\rho}r + 1\right)^2 \|E\|_C.$$

Аналогичным образом можно искать прямоугольные подматрицы и подматрицы локально максимального проективного объема, однако такой поиск будет существенно дороже. Начинать при этом можно, например, с подматрицы, содержащей $r \times r$ подматрицу $3^{1/4}$-локально максимального объема. Как показано выше, её можно достичь за $O(MNr\log r)$ операций, а её отличие от максимального объема не более $\left(1 + 3\sqrt{3}r^2\right)^r$ раз. По следствию 1.1 из теоремы Бине-Коши, максимальный объем $r \times n$ подматрицы будет не более, чем в $\sqrt{C_n^r}$ раз больше максимального объема среди $r \times r$ подматриц (ровно столько слагаемых в сумме, каждое из которых не больше максимума). А из леммы 1.1 получаем, что максимальный r-проективный объем $m \times n$ подматрицы не превосходит максимального объема $r \times r$ подматриц более, чем в $\sqrt{C_m^r C_n^r}$ раз. Отсюда получаем, что если далее увеличивать объем (или проективный объем) в ρ раз на каждом шаге, то потребуется не более $\log_\rho\left(\left(1 + 3\sqrt{3}r^2\right)^r \sqrt{C_n^r}\right) = O\left(r\log_\rho n\right)$ и $\log_\rho\left(\left(1 + 3\sqrt{3}r^2\right)^r \sqrt{C_m^r}\sqrt{C_n^r}\right) = O\left(r\log_\rho(mn)\right)$ шагов соответственно.

Для поиска прямоугольных подматриц на каждом шаге требуется проверить $MNnr$ возможных замен, каждую можно выполнить с помощью быстрого пересчета за $O(nr)$, откуда получаем

полную сложность $O\left(MNr\log r + MNn^2r^3\log_\rho n\right)$. Для проективного объема всего требуется проверить $MNmn$ замен на каждом шаге, каждая требует сингулярного разложения $m \times n$ матрицы, откуда получаем полную сложность $O\left(MNr\log r + MNm^2n^2\min(m,n)\, r\log_\rho(mn)\right)$.

Если требуется только достичь соответствующих оценок на норму Чебышева, достаточно проверять элемент с наибольшей погрешностью, что сокращает число проверяемых замен до $(n+1)(r+1)$ и $(m+1)(n+1)$ соответственно. Пересчет погрешности при этом требует $O(MN)$ и $O(MNr)$ операций соответственно, откуда получаем оценки

$$O\left(MNr\log r + MNr\log_\rho r + n^3r^4\log_\rho n\right)$$

и

$$O\left(MNr\log r + MNr^2\log_\rho r + m^3n^3\min(m,n)\, r\log_\rho(mn)\right)$$

для вычислительной сложности. Сформулируем данные результаты в виде теоремы.

Теорема 4.7. *Подматрица ρ-локально максимального объема размера $r \times n$ во всей матрице может быть найдена за $O\left(MNr\log r + MNn^2r^3\log_\rho n\right)$ операций, а соответствующие ей оценки по норме Чебышева можно гарантировать за $O\left(MNr\log r + MNr\log_\rho r + n^2r^3\log_\rho n\right)$ операций.*

Подматрица ρ-локально максимального r-проективного объема размера $m \times n$ во всей матрице может быть найдена за $O\left(MNr\log r + MNm^2n^2\min(m,n)\, r\log_\rho(mn)\right)$ операций, а за $O\left(MNr\log r + MNr^2\log_\rho r + m^2n^2\min(m,n)\, r\log_\rho(mn)\right)$ операций можно гарантировать соответствующие ей оценки по норме Чебышева.

Полученные оценки показывают, что, в отличие от подматриц почти максимального объема, поиск которых при $n < r$ является NP-сложной задачей [29] (насколько известно автору, вопрос о NP-сложности приближенного поиска при $n \geqslant r$ все же остается открытым), подматрицы, близкие к локально максимальному объему можно найти за полиномиальное время, что позволяет гарантировать выполнение различных оценок, которые часто ассоциируются с подматрицами максимального объема. В частности, возможно быстрое построение крестовых аппроксимаций с гарантиями точности по норме Чебышева и спектральной норме.

В случае симметричных неотрицательно определенных матриц матрицу локально максимального объема искать гораздо проще. Так как $A - C\hat{A}^{-1}R$ также является симметричной неотрицательно определенной, когда $\hat{A}^{-1}$ – подматрица на диагонали, то замена строки и столбца с одинаковыми номерами всегда выгоднее, чем разных: после удаления добавить всегда выгоднее максимальный по модулю элемент, а он в $A - C\hat{A}^{-1}R$ с $\hat{A} \in \mathbb{C}^{(r-1)\times(r-1)}$ находится на диагонали. Таким образом, при поиске подматрицы локально максимального объема можно ограничиться симметричными заменами.

Любая симметричная неотрицательно определенная матрица $A \in \mathbb{C}^{N\times N}$ представима в виде $A = B^*B$, а потому такой поиск эквивалентен поиску столбцов максимального объема в B. При этом исключение Гаусса с полным поиском ведущего элемента в A будет эквивалентно QR с поиском ведущего столбца в B, что гарантирует отличие объема найденной матрицы от максимального не более чем в r^r раз, согласно утверждению 4.4. Схожий факт был впервые замечен в [96] (с коэффициентом $(r!)^2$ вместо r^r) исходя из рассуждений и оценок, полученных в [60].

При этом критерий для одновременной замены i-й строки и столбца матрицы $\hat{A}$ на j-ю строку и столбец матрицы A определяется отношением объемов, равным

$$\frac{V_{new}}{V_{old}} = |(\hat{A}^{-1}R)_{ij}|^2 + \hat{A}^{-1}_{ii}\left(A - C\hat{A}^{-1}R\right)_{jj}$$

и может быть пересчитан за $O(Nr)$. Такие замены в A эквивалентны сильному RRQR (SRRQR) в B, а потому гарантируют нахождение ρ-локально максимального объема не более чем за $\frac{r}{2}\log_\rho r$ шагов (до этого наилучшей известной оценкой было $r\log_\rho\sqrt{rN}$ шагов [52]). Запишем данный результат в виде теоремы.

Теорема 4.8 ([41])**.** *Пусть $A = A^* \geqslant 0$, $A \in \mathbb{C}^{N\times N}$. Тогда на основе строк C^* и столбцов C, соответствующих подматрице $\hat{A}$, такой, что*

$$\max_{ij}\left(|\left(\hat{A}^{-1}R\right)_{ij}|^2 + \hat{A}^{-1}_{ii}\left(A - C\hat{A}^{-1}R\right)_{jj}\right) \leqslant \rho$$

(то есть обладающей ρ-локально максимальным объемом во всей матрице) можно построить сильное RRLU с

$$f = \rho,$$
$$p_1(r, M, N) = p_2(r, M, N) = \sqrt{1 + \rho^2 r(N - r)}.$$

Причем найти подматрицу с ρ-локальным максимальным объемом можно за $O\left(Nr^2\log_\rho r\right)$ операций.

4.4.3. Поиск прямоугольных подматриц

Объединяя вместе алгоритмы $\mathrm{dominant}$ и SRRQR, мы можем быстро искать подматрицы локально максимального объема произвольной формы.

Предлагаемый алгоритм 4.9 состоит из двух частей. В первой он ищет подматрицу локально максимального объема в фиксированных строках, а в другой подматрицу локально максимального объема в фиксированных столбцах (не меняя её размеры), см. рисунок 4.2). Без ограничения общности будем считать, что число строк m больше числа столбцов r. Иначе алгоритм отличается лишь транспонированием.

Алгоритм 4.9 Поиск прямоугольной матрицы локально максимального объема

Вход: Матрица $A \in \mathbb{C}^{M \times N}$, стартовые наборы индексов строк $\mathcal{I}$ и столбцов $\mathcal{J}$ размера m и r соответственно.

Выход: Факторы крестовой аппроксимации $C\hat{A}^{+}R$ ранга r.

while были перестановки строк или столбцов **do**
  $\hat{A} := A_{\mathcal{I},\mathcal{J}}$
  $C := A_{:,\mathcal{J}}$
  while $\max\limits_{i,j} \left[\left| \left(C\hat{A}^{+} \right)_{ij} \right|^2 + \left(1 + \left\| \left(C\hat{A}^{+} \right)_{i,:} \right\|_2^2 \right) \left(1 - \left\| \left(C\hat{A}^{+} \right)_{j,:} \right\|_2^2 \right) \right] > 1$ **do**
    Замена j на i в $\mathcal{I}$, i и j соответствуют максимуму выше
    Обновление $C\hat{A}^{+}$
  end while
  $\hat{A} := A_{\mathcal{I},\mathcal{J}}$
  Берем Q из QR разложения $\hat{A}$
  $R := A_{\mathcal{I},:}$
  while $\max\limits_{i,j} \left[\left| \left(\hat{A}^{+}R \right)_{ij} \right|^2 + \left\| \hat{A}^{+}_{i,:} \right\|_2^2 \left\| (R - QQ^{*}R)_{:,j} \right\|_2^2 \right] > 1$ **do**
    Замена i на j в $\mathcal{J}$, i и j соответствуют максимуму выше
    Обновление $\hat{A}^{+}R$, $\hat{A}^{+}$ и $Q^{*}R$
  end while
end while

Если число замен строк и столбцов ограничено m и r соответственно (по аналогии с maxvol), а число переходов между строками и столбцами ограничено константой, то общая сложность алгоритма составит $O\left(Mm^2 + Nmr\right)$.

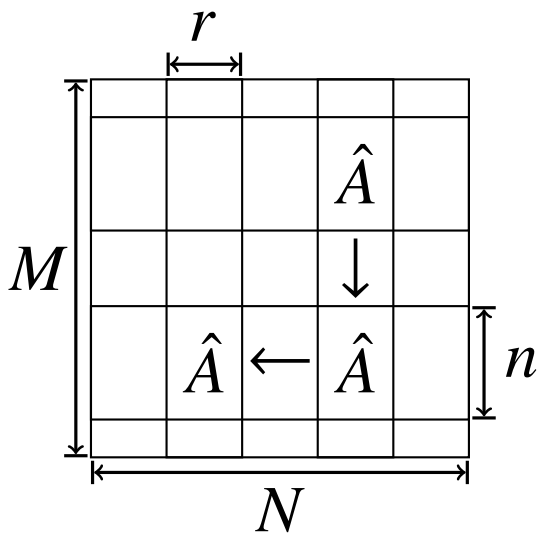


Рис. 4.2: Обновление текущей подматрицы $\hat{A} = A_{\mathcal{I},\mathcal{J}}$ в процессе поиска прямоугольной подматрицы локально максимального объема.

4.5. Поиск подматриц большого проективного объема

Поиск подматрицы локально максимального объема из предыдущих разделов позволяет строить аппроксимации вида $C\hat{A}^{+}R$. Однако, согласно результатам теорем 2.3 и 3.4, имеет смысл выбрать число строк и столбцов большим, чем требуемый ранг, и искать подматри-

цу большого r-проективного объема. В этом случае вместо $C\hat{A}^{+}R$ можно использовать крестовую аппроксимацию вида $C\hat{A}_r^{+}R$, где $\hat{A}_r^{+} = \left(\hat{A}_r\right)^{+}$ – r-псевдообратная матрица, которую можно получить с помощью псевдообращения сокращенного сингулярного разложения $\hat{A}$.

Простейшим способом получить подматрицу большого проективного объема является расширение $r \times r$ подматрицы с помощью алгоритма rect-maxvol. Этот вариант является для нас приемлемым, поскольку не существует быстрых алгоритмов поиска локально максимального проективного объема (даже при использовании быстрого пересчета сингулярного разложения выбор наилучшей замены в текущих m строках займет не меньше Nnm операций; возможно искать в подпространстве r наибольших сингулярных чисел фиксированных строк/столбцов, но это тоже дорого, и нет доказательств, что такой алгоритм обязательно остановится).

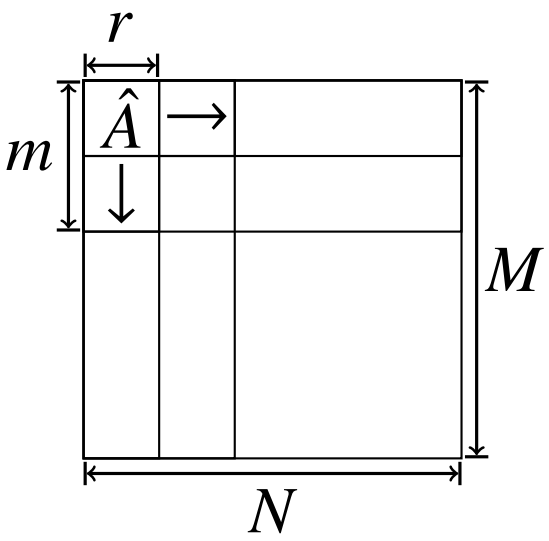


Рис. 4.3: Быстрый поиск подматрицы большого проективного объема: расширение $\hat{A} = A_{I,\mathcal{J}}$ (полученной с помощью maxvol) после двух применений rect-maxvol.

Идея расширения квадратной подматрицы показана на рисунке 4.3.

Алгоритм 4.10 Быстрый поиск подматрицы большого проективного объема

Вход: Матрица $A \in \mathbb{C}^{M\times N}$, стартовые наборы индексов строк $\mathcal{I}$ и столбцов $\mathcal{J}$ размера r, финальные числа n и m столбцов C и строк R соответственно.

Выход: Факторы крестовой аппроксимации $C\hat{A}_r^{+}R$.

1: Поиск подматрицы локально максимального объема $A_{\mathcal{I},\mathcal{J}} \in \mathbb{C}^{r\times r}$ с помощью maxvol (алгоритм 4.5)
2: $C := A_{:,\mathcal{J}}$
3: $R := A_{\mathcal{I},:}$
4: Добавление $n - r$ столбцов из R с помощью rect-maxvol (алгоритм 4.2), их индексы добавляются в $\mathcal{J}$
5: Добавление $m - r$ строк из C с помощью rect-maxvol, их индексы добавляются в $\mathcal{I}$
6: $\hat{A} = A_{\mathcal{I},\mathcal{J}} \in \mathbb{C}^{m\times n}$
7: $\hat{A}_r^{+} = \left(\hat{A}_r\right)^{+}$
8: $C := A_{:,\mathcal{J}}$
9: $R := A_{\mathcal{I},:}$

Основным преимуществом данного метода является то, что rect-maxvol нами был суще-

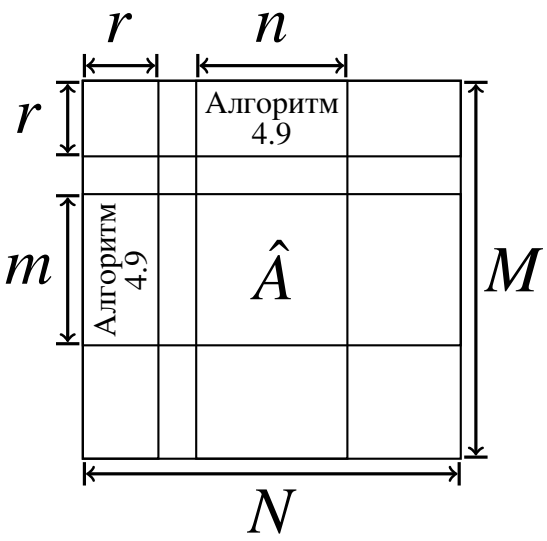


Рис. 4.4: Подматрица $\hat{A} = A_{I,\mathcal{J}}$, возвращаемая алгоритмом maxvol-proj.

ственно ускорен (см. раздел 4.2, алгоритм 4.2), так что $n - r$ столбцов могут быть добавлены за $O\left(Nr(n-r)\right)$ операций, а $m-r$ строк за $O\left(Mr(m-r)\right)$ операций, так что общая сложность алгоритма составляет $O\left(Mmr + Nnr + (m+n)^2 \min(m,n)\right)$, где последнее слагаемое есть сложность сингулярного разложения $m \times n$ подматрицы.

С другой стороны, из-за отсутствия перестановок строк/столбцов, не стоит ожидать, что данные алгоритм будет всегда давать высокую точность аппроксимации. Данную проблему можно частично решить, используя алгоритм поиска подматриц локально максимального объема (алгоритм 4.9). А именно, мы можем найти $m \times n$ подматрицу большого проективного объема, используя $r \times n$ и $m \times r$ подматрицы локально максимального объема. Полученный алгоритм назовем maxvol-proj. Его идея проиллюстрирована на рисунке 4.4.

Алгоритм 4.11 maxvol-proj

Вход: Матрица $A \in \mathbb{C}^{M \times N}$, требуемый ранг r, число строк m и столбцов n.

Выход: Факторы крестовой аппроксимации $C\hat{A}_r^+ R$.

1: Поиск подматрицы локально максимального объема $A_{\mathcal{I}_1,\mathcal{J}_1} \in \mathbb{C}^{m \times r}$ с помощью алгоритма 4.9
2:
3: Поиск подматрицы локально максимального объема $A_{\mathcal{I}_2,\mathcal{J}_2} \in \mathbb{C}^{r \times n}$ с помощью алгоритма 4.9
4: $\hat{A} = A_{\mathcal{I}_1,\mathcal{J}_2} \in \mathbb{C}^{m \times n}$
5: $\hat{A}_r^+ = \left(\hat{A}_r\right)^+$
6: $C = A_{:,\mathcal{J}_2}$
7: $R = A_{\mathcal{I}_1,:}$

Такой подход можно обосновать следующим образом. Пусть $\operatorname{rank} A = r$, и некоторые её подматрицы размера $r \times n$ и $m \times r$ обладают локально максимальным объемом. Тогда легко

проверить, что подматрица $\hat{A}$ на пересечении соответствующих им m строк и n столбцов обладает локально максимальным r-проективным объемом.

Утверждение 4.14. *Пусть $A_{\mathcal{I}_1,\mathcal{J}_1} \in \mathbb{C}^{m\times r}$ и $A_{\mathcal{I}_2,\mathcal{J}_2} \in \mathbb{C}^{r\times n}$ – подматрицы локально максимального объема в матрице $A \in \mathbb{C}^{M\times N}$ ранга r, заданные наборами индексов строк $\mathcal{I}_1$, $\mathcal{I}_2$ и столбцов $\mathcal{J}_1$, $\mathcal{J}_2$. Тогда $\hat{A} = A_{\mathcal{I}_1,\mathcal{J}_2}$ – подматрица локально максимального r-проективного объема.*

Доказательство. Согласно одному из опредеелений ранга, существует разложение

$$A = UV, \quad U \in \mathbb{C}^{M\times r}, \quad V \in \mathbb{C}^{r\times N}.$$

Тогда данные по условию подматрицы можно записать в виде:

$$\begin{gathered} A_{\mathcal{I}_1,\mathcal{J}_1} = \hat{U}\tilde{V}, \\ A_{\mathcal{I}_2,\mathcal{J}_2} = \tilde{U}\hat{V}, \\ \hat{A} = A_{\mathcal{I}_1,\mathcal{J}_2} = \hat{U}\hat{V}, \\ \hat{U} \in \mathbb{C}^{m\times r},\ \tilde{V} \in \mathbb{C}^{r\times r},\ \tilde{U} \in \mathbb{C}^{r\times r},\ \hat{V} \in \mathbb{C}^{r\times n}. \end{gathered}$$

Рассмотрим QR разложение $\hat{U} = Q_U R_U$ и LQ разложение $\hat{V} = L_V Q_V$. Поскольку унитарные матрицы не меняют сингулярных чисел, получаем

$$\mathcal{V}_r\left(\hat{A}\right) = \mathcal{V}_r\left(\hat{U}\hat{V}\right) = \mathcal{V}_r\left(Q_U R_U L_V Q_V\right) = \mathcal{V}_r\left(R_U L_V\right) = \mathcal{V}\left(R_U L_V\right) = \left|\det\left(RL\right)\right|.$$

Рассмотрим подматрицу $\hat{A}' = A_{\mathcal{I}_1',\mathcal{J}_2} = \hat{U}'\hat{V}$ в тех же столбцах, что и $\hat{A}$, но с другими строками. Используя $\hat{U}' = Q_U' R_U'$,

$$\mathcal{V}_r\left(\hat{A}'\right) = \left|\det\left(R_U' L_V\right)\right| = \mathcal{V}_r\left(\hat{A}\right)\frac{\left|\det R_U'\right|}{\left|\det R_U\right|}.$$

Если её проективный объем превосходит $\mathcal{V}_r\left(\hat{A}\right)$, то найдется подматрица $A_{\mathcal{I}_1',\mathcal{J}_1}$ в тех же строках, что и A', и такая, что

$$\mathcal{V}\left(A_{\mathcal{I}_1',\mathcal{J}_1}\right) = \left|\det\left(R_U'\tilde{V}\right)\right| > \left|\det\left(R_U\tilde{V}\right)\right| = \mathcal{V}\left(A_{\mathcal{I}_1,\mathcal{J}_1}\right),$$

что противоречит предположению о том, что $A_{\mathcal{I}_1',\mathcal{J}_1}$ обладает локально максимальным объемом.

Таким образом, мы показали, что $\hat{A}$ обладает локально максимальным объемом в своих столбцах. Аналогично доказывается, что она обладает локально максимальным объемом в своих строках. Можно также рассмотреть одновременную замену строки и столбца и доказать, что $\hat{A}$ – подматрица локально максимального r-проективного объема во всей матрице. □

Получается, что в случае, когда $\operatorname{rank} A = r$, поиск подматрицы локально максимального объема сводится к поиску двух меньших подматриц локально максимального объема. Естественно, если ранг A не равен r, то нет никаких гарантий на то, что подматрица на пересечении будет обладать локально максимальным проективным объемом.

4.6. Другие методы поиска невырожденных подматриц

Здесь мы рассмотрим другие алгоритмы с гарантиями для $\|\hat{R}^+\|_2/\|R^+\|_2$ и $\|\hat{R}^+\|_F/\|R^+\|_F$. Известные и новые результаты приведены в таблице 4.1.

Таблица 4.1: Сложность и точность при поиске сильно невырожденных $\hat{R} \in \mathbb{R}^{r\times n}$ подматриц в произвольных строках $R \in \mathbb{R}^{r\times N}$. ε обозначает машинную точность, если число итераций от нее зависит (при этом все еще считается, что стандартные арифметические операции стоят $O(1)$). Первая ссылка в источниках содержит оценки точности и сложности, а вторая алгоритм, для (возможно, модификации) которого эти оценки получены.

Источник	$\|\hat{R}^+\|_F^2/\|R^+\|_F^2$	$\|\hat{R}^+\|_2^2/\|R^+\|_2^2$	Сложность
SRRQR [52, 39], $n = r$	$\left(1+\rho^2(N-r)\right)\frac{r\|R^+\|_2}{\|R^+\|_F}$	$1+\rho^2 r(N-r)$	$O(Nr^2\log N/\log\rho)$
MVEE ($\varepsilon = 1$) [97, 76], $n \geqslant 4r\ln\ln r + 42r$	$\left(1+\frac{2}{r}(N-n)\right)\frac{r\|R^+\|_2}{\|R^+\|_F}$	$1+2(N-n)$	$O(Nnr)$
Теорема 3.7 ($\delta = 1/2$) из [44, 98], $n \geqslant 32r\ln(4r)$	$4N$	$4N$	$O(Nr^2 + n\log n)$
Теорема 3.11 ($\delta = 1/2$) из [44, 46], $n = r$	$\rho^2(N-r+1)$	$\rho^2 r(N-r+1)$	$O(Nr^3/\log\rho)$
Теорема 3.5 из [44, 99], $n > r$	$\frac{(1+\sqrt{\frac{N}{n}})^2}{(1-\sqrt{\frac{r}{n}})^2}$	$\frac{(1+\sqrt{\frac{N}{n}})^2}{(1-\sqrt{\frac{r}{n}})^2}$	$O(Nnr^2)$
Теорема 3.1 из [44, 51], $n \geqslant r$	$\frac{N-r+1}{n-r+1}$	$r\frac{N-r+1}{n-r+1}$	$O\left(N(N-n)r\log\varepsilon^{-1}\right)$
Следствие 3.3 из [44, 51], $n \geqslant r$	$\frac{N-r+1}{n-r+1}\cdot\frac{r\|R^+\|_2}{\|R^+\|_F}$	$1+r\frac{N-n}{n-r+1}$	$O\left(N(N-n)r\log\varepsilon^{-1}\right)$
НОВЫЕ ОЦЕНКИ			
Теорема 4.10, $n \geqslant r$	$\frac{N+2-2\sqrt{\frac{r}{n+1}}}{\left(\sqrt{n+1}-\sqrt{r}\right)^2}$	$\frac{N+2-2\sqrt{\frac{r}{n+1}}}{\left(\sqrt{n+1}-\sqrt{r}\right)^2}$	$O(Nnr^2)$
Теорема 4.11, $n = r$	$(N-r+1)\frac{r\|R^+\|_2^2}{\|R^+\|_F^2}$	$1+r(N-r)$	$O(Nr^2)$
Теорема 4.13, $n \geqslant r$	$\frac{N-r+1}{n-r+1}$	$1+\frac{\|R^+\|_F^2}{\|R^+\|_2^2}\cdot\frac{N-n}{n-r+1}$	$O\left(N(N-n)r\right)$
Алгоритм 4.12, $n \geqslant r$	$\left(1+\frac{\rho^2(N-n)}{n-r+1}\right)\frac{r\|R^+\|_2^2}{\|R^+\|_F^2}$	$1+\frac{\rho^2 r}{n-r+1}(N-n)$	$O\left(Nr^2\log\log r + Nnr/\log\rho\right)$
Алгоритм A.6, $n \geqslant r$	$\frac{N-r+1}{n-r+1}\frac{r\|R^+\|_2^2}{\|R^+\|_F^2}$	$1+\frac{r}{n-r+1}(N-n)$	$O\left(Nnr^2\right)$

Алгоритм 4.12 соответствует ускоренному алгоритму максимизации объема подматрицы, который мы рассмотрим в следующем подразделе 4.7. Алгоритм A.6 основан на дерандомизации

выборки по объему [80] и выписан в Приложении А.

Отношение $\|\hat{R}^+\|_2/\|R^+\|_2$, как уже было показано выше в утверждении 4.1, определяется величиной $t(r,n,N)$, определение 2.3. Сравнивая результат теоремы 3.1 из [44] в таблице 4.1 и утверждение 4.1, получаем, что оптимальное значение $\|\hat{R}^+\|_F/\|R^+\|_F = \frac{N-r+1}{n-r+1}$ является достижимым.

Заметим также, что с точки зрения нормы Фробениуса на практике не менее интересна величина $\left\|\hat{R}^+R\right\|_F = \left\|\hat{U}^+\right\|_F$ (лемма 2.4), которая соответствует частному случаю, когда все сингулярные числа R равны. Для нее, как видно из таблицы 4.1, особенно эффективны методы, основанные на поиске подматриц локально максимального объема (SRRQR и алгоритм 4.12) и позволяющие достичь верхней оценки вплоть до $\sqrt{r\frac{N-r+1}{n-r+1}}$. При этом в утверждении 2.6 была доказана близкая к ней нижняя оценка $\sqrt{r\frac{N}{n}}$.

Начнем с того, что усовершенствуем результат теоремы 3.5 из [44].

Теорема 4.9 ([44])**.**

$$t(r,n,N) \leqslant \frac{1+\sqrt{N/n}}{1-\sqrt{r/n}}.$$

Соответствующая подматрица может быть найдена за $O(Nnr^2)$ операций.

Сразу заметим, что в [44] производился один лишний шаг, а потому в теореме 4.9 значение n можно заменить на $n+1$. В нашей усовершенствованной оценке в знаменателе также будет $\sqrt{n+1}$ вместо $\sqrt{n}$, что позволит применять её даже для случая $n = r$ (хоть результат при $n = r$ и будет хуже, чем для локально максимального объема).

Теорема 4.9 основана на следующей лемме из [99] (здесь мы используем комплексную версию, что является более общим случаем).

Лемма 4.9 ([99])**.** *Пусть для $A \in \mathbb{C}^{r\times r}$, $A = A^*$ справедливо $\lambda_{\min}(A) > l$, $\operatorname{tr}\left((A-lI)^{-1}\right) \leqslant 1/\delta_l$. Пусть $v \in \mathbb{C}^r$ – произвольный вектор.*

Тогда $\lambda_{\min}(A) > l+\delta_l$, и если

$$0 < \frac{1}{t} \leqslant \frac{\left\|(A-(l+\delta_l)I)^{-1}v\right\|_2^2}{\delta_l \operatorname{tr}\left((A-(l+\delta_l)I)^{-1}(A-lI)^{-1}\right)} - v^*(A-(l+\delta_l)I)^{-1}v \stackrel{def}{=} L_A(v),$$

то

$$\operatorname{tr}\left((A+tvv^*-(l+\delta_l)I)^{-1}\right) \leqslant \operatorname{tr}\left((A-lI)^{-1}\right) \leqslant 1/\delta_l.$$

Мы воспользуемся леммой 4.9, чтобы улучшить оценку теоремы 4.9.

В отличие от [44], мы будем использовать лишь одно множество векторов, и потребуем, чтобы выбранные векторы были различны, а все коэффициенты были равны 1.

Доказательство основано на следующей лемме.

Лемма 4.10. *Пусть в условиях леммы 4.9 $A = \hat{U}\hat{U}^*$, где $\hat{U} \in \mathbb{C}^{r\times k}$ является подматрицей матрицы $U \in \mathbb{C}^{r\times n}$, $UU^* = I$. Пусть* $\operatorname{tr}\left((A - lI)^{-1}\right) \leqslant -r/l_0 < 1/\delta_l$, $l = l_0 + k\delta_l$, $l + \delta_l \leqslant 1$, $\lambda_{\min}(A) > l$. *Пусть вектор v выбирается равновероятно из множества столбцов матрицы U, не входящих в $\hat{U}$. Тогда*

$$\mathbb{E}L_A(v) \geqslant \frac{(1/\delta_l - 2)\left(1 + r\frac{\delta_l}{l_0}\right) - (k-1)r\frac{\delta_l}{l_0} - k}{N - k} \tag{4.61}$$

или

$$\operatorname{tr}\left((A - (l+\delta_l)I)^{-1}\right) \leqslant -r/l_0. \tag{4.62}$$

Доказательство. В доказательстве леммы 7.5 из [68] было получено неравенство

$$\sum_{v\in U} L_A(v) > 1/\delta_l - \operatorname{tr}\left((A - lI)^{-1}\right). \tag{4.63}$$

При этом сумму по столбцам $v \in \hat{U}$ можно оценить как

$$\begin{aligned}
\sum_{v\in\hat{U}} L_A(v) &= \frac{\left\|(A-(l+\delta_l)I)^{-1}\hat{U}\right\|_F^2}{\delta_l \operatorname{tr}\left((A-(l+\delta_l)I)^{-1}(A-lI)^{-1}\right)} - \operatorname{tr}\left(\hat{U}^*(A-(l+\delta_l)I)^{-1}\hat{U}\right) \\
&= \frac{\operatorname{tr}\left((A-(l+\delta_l)I)^{-2}A\right)}{\delta_l \operatorname{tr}\left((A-(l+\delta_l)I)^{-1}(A-lI)^{-1}\right)} - \operatorname{tr}\left((A-(l+\delta_l)I)^{-1}A\right) \\
&= \frac{\operatorname{tr}\left((A-(l+\delta_l)I)^{-1}\right) + (l+\delta_l)\operatorname{tr}\left((A-(l+\delta_l)I)^{-2}\right)}{\delta_l \operatorname{tr}\left((A-(l+\delta_l)I)^{-1}(A-lI)^{-1}\right)} \\
&\quad - r - (l+\delta_l)\operatorname{tr}\left((A-(l+\delta_l)I)^{-1}\right) \\
&= -r + (l+\delta_l)\sum_{v\in U} L_A(v) + \frac{\operatorname{tr}\left((A-(l+\delta_l)I)^{-1}\right)}{\delta_l \operatorname{tr}\left((A-(l+\delta_l)I)^{-1}(A-lI)^{-1}\right)} \\
&= -r + (l+\delta_l)\sum_{v\in U} L_A(v) + 1 + \frac{\operatorname{tr}\left((A-lI)^{-1}\right)}{\delta_l \operatorname{tr}\left((A-(l+\delta_l)I)^{-1}(A-lI)^{-1}\right)} \\
&\leqslant -r + (l+\delta_l)\sum_{v\in U} L_A(v) + 1 + \frac{r\operatorname{tr}\left((A-lI)^{-1}\right)}{\delta_l \operatorname{tr}\left((A-(l+\delta_l)I)^{-1}\right)\operatorname{tr}\left((A-lI)^{-1}\right)} \\
&= -r + (l+\delta_l)\sum_{v\in U} L_A(v) + 1 + \frac{r}{\delta_l \operatorname{tr}\left((A-(l+\delta_l)I)^{-1}\right)} \\
&\leqslant -r + (l+\delta_l)\sum_{v\in U} L_A(v) + 1 - \frac{l_0}{\delta_l},
\end{aligned}$$

где последнее неравенство получено в предположении, что (4.62) нарушено.

Подстановка в (4.63) приводит нас к

$$\begin{aligned}\mathbb{E}L_A(v) &\geqslant \frac{(1-l-\delta_l)\left(1/\delta_l - \operatorname{tr}\left((A-lI)^{-1}\right)\right) + r - 1 + \frac{l_0}{\delta_l}}{N-k} \\ &\geqslant \frac{(1-k\delta_l-\delta_l)\left(1/\delta_l + \frac{r}{l_0}\right) - 1}{N-k} \\ &= \frac{(1/\delta_l - 2)\left(1 + r\frac{\delta_l}{l_0}\right) - (k-1)r\frac{\delta_l}{l_0} - k}{N-k}.\end{aligned}$$

□

Теперь мы можем воспользоваться леммами 4.9 и 4.10 для улучшения оценки теоремы 4.9.

Теорема 4.10.

$$t(r,n,N) \leqslant \frac{\sqrt{N}}{\sqrt{n+1}-\sqrt{r}} + \sqrt{\frac{1}{N(n+1)}}.$$

Соответствующая подматрица может быть найдена за $O(Nnr^2)$ операций.

Данная оценка была нами учтена в таблице 2.1, использующей аппроксимации по спектральной норме из раздела 2.2.1.

Доказательство. Выберем

$$\delta_l = \frac{1}{\frac{N}{1-\sqrt{\frac{r}{n+1}}} + 2} \tag{4.64}$$

и

$$l_0 = -\frac{\sqrt{r(n+1)}}{\frac{N}{1-\sqrt{\frac{r}{n+1}}} + 2}. \tag{4.65}$$

На шаге $k \geqslant 0$ мы будем иметь $l = l_0 + k\delta_l$. Условие леммы 4.10 выполняется на нулевом шаге и будет продолжать выполняться пока либо $\operatorname{tr}\left((A-(l+\delta_l))^{-1}\right) \leqslant -r/l_0$ даже без добавления нового столбца в $\hat{U}$, либо возможен выбор $t = 1$ в лемме 4.9, что гарантируется заданными δ_l (4.64) и l_0 (4.65) при их подстановке в (4.61). А именно, всегда найдется такой столбец v, что

$$\begin{aligned}L_A(v) \geqslant \mathbb{E}L_A(v) &\geqslant \frac{(1/\delta_l - 2)\left(1 + r\frac{\delta_l}{l_0}\right) - (k-1)r\frac{\delta_l}{l_0} - k}{N-k} \\ &= \frac{N - (k-1)r\frac{\delta_l}{l_0} - k}{N-k} \geqslant 1\end{aligned}$$

при $k \geqslant 1$, а при $k = 0$ подстановка $A = 0$ дает

$$\mathbb{E}L_A(v) = \frac{r + \frac{l_0}{\delta_l}}{N(l_0 + \delta_l)} = \frac{\left(r - \sqrt{r(n+1)}\right)\left(2 + \frac{N}{1-\sqrt{\frac{r}{n+1}}}\right)}{N\left(1 - \sqrt{r(n+1)}\right)}$$
$$\geqslant \frac{\left(r - \sqrt{r(n+1)}\right)\left(0 + \frac{N}{1-\sqrt{\frac{r}{n+1}}}\right)}{N\left(0 - \sqrt{r(n+1)}\right)} \geqslant 1.$$

После $k = n$ шагов согласно лемме 4.9 мы будем иметь оценку

$$t^{-1}(r, n, N) \geqslant \sqrt{\lambda_{\min}(\hat{U}\hat{U}^*)} = \sqrt{\lambda_{\min}(A)} \geqslant \sqrt{l_0 + (n+1)\delta_l}.$$

Подставляя в нее (4.64) и (4.65), получаем

$$t(r, n, N) \leqslant \frac{\sqrt{N + \frac{2}{\sqrt{n+1}}\left(\sqrt{n+1} - \sqrt{r}\right)}}{\sqrt{n+1} - \sqrt{r}} \leqslant \frac{\sqrt{N}}{\sqrt{n+1} - \sqrt{r}} + \sqrt{\frac{1}{N(n+1)}}.$$

□

Замечание 4.3. Скорее всего, результат теоремы является довольно точным, однако отметим несколько мест, где он может быть улучшен.

Во-первых, на последнем шаге можно прибавлять $-r/l_0$ вместо δ_l, что приведет к неравенству $t^{-1}(r, n, N) \geqslant \sqrt{l_0 + n\delta_l - l_0/r}$.

Во-вторых, можно оценить отношение $\frac{r}{\delta_l \operatorname{tr}\left((A-(l+\delta_l)I)^{-1}\right)}$ более точно, используя неравенство

$$\operatorname{tr}\left((A - (l+\delta_l)I)^{-1}\right) \geqslant -r/l_0 + \frac{\mathbb{E}v^*(A - (l+\delta_l)I)^{-2}v}{1 + \mathbb{E}v^*(A - (l+\delta_l)I)^{-1}v}$$
$$\geqslant -r/l_0 + \frac{\operatorname{tr}\left((A - (l+\delta_l)I)^{-1}\right)}{r} \cdot \frac{(1 - l - \delta_l)\operatorname{tr}\left((A - (l+\delta_l)I)^{-1}\right) - r}{(1 - l - \delta_l)\operatorname{tr}\left((A - (l+\delta_l)I)^{-1}\right) - r + N - k}.$$

Последнее неравенство можно улучшить, выписав квадратное уравнение на $\operatorname{tr}\left((A - (l+\delta_l)I)^{-1}\right)$.

Наконец, неравенство (4.63) также можно улучшить до

$$\sum_{v \in U} L_A(v) > \left(1/\delta_l - \operatorname{tr}\left((A - lI)^{-1}\right)\right)\left(1 + \frac{\delta_l}{r}\operatorname{tr}\left((A - (l+\delta_l)I)^{-1}\right)\right).$$

Итоговые формулы, однако, получаются более громоздкими, например

$$t(r, n, N) \leqslant \sqrt{\frac{N + 1 - \sqrt{\frac{r}{n+1}}}{\left(n - (r-1)\sqrt{\frac{n+1}{r}}\right)\left(1 - \sqrt{\frac{r}{n+1}}\right)\left(1 + \frac{1}{\sqrt{r(n+1)}}\right)}}.$$

Последнюю формулу можно огрубить до

$$t(r, n, N) \leqslant \frac{\sqrt{N}}{\sqrt{n+1} - \sqrt{r}}$$

при

$$\delta_l = \frac{1 - \sqrt{\frac{r}{n+1}}}{N}$$

и

$$l_0 = \frac{r - \sqrt{r(n+1)}}{N}.$$

Заметим, что во всех результатах из [44] и [68] также можно заменить n на $n+1$ в оценках, так как авторы не учли отсутствие необходимости прибавлять tvv^* для оценки на $\lambda_{\min}$ на последнем шаге.

Пусть теперь мы хотим получить на практике оценки на норму псевдообратных как для (квадратных) подматриц локально максимального объема. Как мы показали в разделе 4.3, алгоритм maxvol [11] требует $O\left(Nr^2 \log_\rho r\right)$ операций и гарантирует оценки на норму $\hat{R}^{-1}$ только с точностью до множителя $\rho > 1$. Оказывается, существует способ гарантировать те же оценки, что и для локально максимального объема, но используя жадный алгоритм со сложностью $O\left(Nr^2\right)$. Данный алгоритм (без оценок на норму обратной к подматрице) был впервые предложен в контексте QR разложений в [100].

Заметим, что обычный жадный выбор ведущих столбцов для этого не подходит, так как может давать норму псевдообратной, отличающуюся от наилучшей на множитель порядка 2^r (пример есть в [39]).

Теорема 4.11. *Пусть даны ортонормированные строки $\hat{V} \in \mathbb{C}^{r \times N}$. Тогда за $O(Nr^2)$ можно найти подматрицу $\hat{V} \in \mathbb{C}^{r \times r}$ матрицы V такую, что*

$$\left\|\hat{V}^{-1}\right\|_F \leqslant \sqrt{r\,(N - r + 1)} \tag{4.66}$$

и

$$\left\|\hat{V}^{-1}\right\|_2 \leqslant \sqrt{1 + r\,(N - r)}. \tag{4.67}$$

Доказательство. Алгоритм который мы будем строить, соответствует выбору $A = I$ в теореме 2.15, а V – вектору правых сингулярных векторов Z. Здесь мы выпишем соответствующий алгоритм и напрямую докажем, что он гарантирует (4.66). При этом оценка на 2-норму (4.67)

прямо следует из оценки на норму Фробениуса:

$$\begin{aligned}
\left\|\hat{V}^{-1}\right\|_2^2 &= \left\|\hat{V}^{-1}\right\|_F^2 - \sum_{k=2}^{r} \sigma_k^2\left(\hat{V}^{-1}\right) \\
&\leqslant \left\|\hat{V}^{-1}\right\|_F^2 - (r-1)\sigma_r^2\left(\hat{V}^{-1}\right) \\
&\leqslant \left\|\hat{V}^{-1}\right\|_F^2 - (r-1)/\left\|\hat{V}\right\|_2^2 \\
&\leqslant \left\|\hat{V}^{-1}\right\|_F^2 - (r-1)/\|V\|_2^2 \\
&= \left\|\hat{V}^{-1}\right\|_F^2 - r + 1 \\
&\leqslant r\,(N-r+1) - r + 1 \\
&= 1 + r\,(N-r)\,.
\end{aligned}$$

Итак, пусть мы находимся на k-м шаге, и добавляем $k+1$-й столбец. Для того чтобы упростить его выбор, будем использовать вращения матрицы V слева так, чтобы после k шагов матрица $\hat{V}$ имела вид

$$\hat{V} = \begin{bmatrix} \hat{V}_1 \\ 0 \end{bmatrix} \in \mathbb{C}^{r\times k}, \tag{4.68}$$

где $\hat{V}_1 \in \mathbb{C}^{k\times k}$ – квадратная матрица. Такой вид можно получить, например, после k шагов модифицированного алгоритма Грамма-Шмидта, а потому после r шагов получим полную стоимость $O\left(Nr^2\right)$.

Разбиению (4.68) соответствует разбиение матрицы V:

$$V = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \in \mathbb{C}^{r\times N}, \quad V_1 \in \mathbb{C}^{k\times N},\ V_2 \in \mathbb{C}^{(r-k)\times N}.$$

Пусть мы добавляем к $\hat{V}$ столбец $V_j = \begin{bmatrix} V_{1,j} \\ V_{2,j} \end{bmatrix} \in \mathbb{C}^r$ матрицы V. Оценим норму Фробениуса псевдообратной полученной матрицы:

$$\begin{aligned}
\left\|\begin{bmatrix} \hat{V} & V_j \end{bmatrix}^+\right\|_F^2 &= \left\|\begin{bmatrix} \hat{V}_1 & V_{1,j} \\ 0 & V_{2,j} \end{bmatrix}^+\right\|_F^2 \\
&= \left\|\begin{bmatrix} \hat{V}_1^{-1} & \frac{1}{\|V_{2,j}\|_2^2}\hat{V}_1^{-1}V_{1,j}V_{2,j}^* \\ 0 & \frac{1}{\|V_{2,j}\|_2^2}V_{2,j}^* \end{bmatrix}\right\|_F^2 \\
&= \left\|\hat{V}_1^{-1}\right\|_F^2 + \frac{1}{\|V_{2,j}\|_2^4}\left\|\hat{V}_1^{-1}V_{1,j}V_{2,j}^*\right\|_F^2 + \frac{1}{\|V_{2,j}\|_2^4}\left\|V_{2,j}^*\right\|_2^2 \\
&= \left\|\hat{V}^+\right\|_F^2 + \frac{\left\|\hat{V}_1^{-1}V_{1,j}\right\|_2^2}{\|V_{2,j}\|_2^2} + \frac{1}{\|V_{2,j}\|_2^2}.
\end{aligned} \tag{4.69}$$

Столбец j будем выбирать из условия

$$j = \arg\min_{j>k} \left(1 + \left\|\hat{V}_1^{-1} V_{1,j}\right\|_2^2\right) / \|V_{2,j}\|_2^2. \tag{4.70}$$

Такой выбор занимает $O(Nr)$ на каждом из r шагов, если матрица $\hat{V}_1^{-1} V_1$ известна. Её обновление занимает $O(Nk)$, так как $\hat{V}_1$ и V_1 получают только один новый столбец и новую строку соответственно (то есть ранг обновления для них обоих равен 1). Таким образом, полная вычислительная сложность поска $r \times r$ подматрицы составит $O\left(Nr^2\right)$.

Используя оценки на матожидания (при равновероятном выборе столбцов)

$$\mathbb{E}_{j>k} \left\|\hat{V}_1^{-1} V_{1,j}\right\|_2^2 = \frac{\left\|\hat{V}_1^{-1} V_1\right\|_F^2 - \left\|\hat{V}_1^{-1} \hat{V}_1\right\|_F^2}{N-k} = \frac{\left\|\hat{V}_1^{-1}\right\|_F^2 - k}{N-k}$$

и

$$\mathbb{E}_{j>k} \|V_{2,j}\|_2^2 = \frac{\|V_{2,j}\|_F^2}{N-k} = \frac{r-k}{N-k},$$

получаем, оценивая минимум отношения через отношение средних:

$$\min_j \left(1 + \left\|\hat{V}_1^{-1} V_{1,j}\right\|_F^2\right) / \|V_{2,j}\|_2^2 \leqslant \frac{1 + \mathbb{E}_{j>k} \left\|\hat{V}_1^{-1} V_{1,j}\right\|_F^2}{\mathbb{E}_{j>k} \left\|V_{2,j}\right\|_2^2} \leqslant \frac{N - 2k + \left\|\hat{V}_1^{-1}\right\|_F^2}{r-k} = \frac{N - 2k + \left\|\hat{V}^{+}\right\|_F^2}{r-k} \tag{4.71}$$

Доказывать оценку (4.66) будем по индукции, используя следующее предположение индукции для k столбцов:

$$\left\|\hat{V}^{+}\right\|_F^2 \leqslant k \frac{N-k+1}{r-k+1}. \tag{4.72}$$

При $k = 0$ в (4.70) определим числитель как 1, когда V_1 не содержит ни одной строки, что приведет к выбору столбца V_j с наибольшей нормой и даст

$$\left\|V_j^{+}\right\|_F^2 \leqslant \frac{1}{\mathbb{E}_j \left\|V_j\right\|_2^2} = \frac{N}{\|V\|_F^2} = \frac{N}{r}.$$

Тем самым, мы доказали базу индукции.

При переходе к $k+1$ столбцам, объединяя (4.69), (4.71) и (4.72), получаем:

$$\begin{aligned}
\left\|\begin{bmatrix} \hat{V} & V_j \end{bmatrix}^{+}\right\|_F^2 &\leqslant \left\|\hat{V}^{+}\right\|_F^2 + \frac{N - 2k + \left\|\hat{V}^{+}\right\|_F^2}{r-k} \\
&= \frac{N - 2k + \left\|\hat{V}^{+}\right\|_F^2 (r-k+1)}{r-k} \\
&\leqslant \frac{N - 2k + k\frac{N-k+1}{r-k+1} (r-k+1)}{r-k} \\
&= \frac{N - 2k + k\,(N-k+1)}{r-k} \\
&= (k+1)\,\frac{N-(k+1)}{r-(k+1)+1}.
\end{aligned}$$

И, в итоге, после r шагов, мы получим

$$\left\|\hat{V}^{-1}\right\|_F^2 \leqslant r\,(N-r+1)\,.$$

□

Следствие 4.4. Из доказательства также следует, что за $k \leqslant r$ шагов (требующих $O(Nkr)$ операций) в матрице U можно найти подматрицу $\hat{U} \in \mathbb{C}^{r\times k}$ такую, что

$$\left\|\hat{V}^{+}\right\|_F \leqslant \sqrt{k + \frac{k}{r-k+1}\,(N-r)} = \sqrt{k\frac{N-k+1}{r-k+1}}$$

и

$$\left\|\hat{V}^{+}\right\|_2 \leqslant \sqrt{1 + \frac{k}{r-k+1}\,(N-r)}.$$

Заметим, что оценки теоремы 4.11 не гарантируют высокой близости подматрицы $\hat{V}$ к подматрице максимального объема. Однако, слишком малым объем подматрицы $\hat{V}$ также быть не может, что следует из утверждения ниже.

Утверждение 4.15. *Пусть подматрица $\hat{V} \in \mathbb{C}^{r\times r}$ ищется в матрице $V \in \mathbb{C}^{r\times N}$, $VV^* = I$, с помощью метода, описанного при доказательстве теоремы 4.11. Пусть $V_M \in \mathbb{C}^{r\times r}$ – подматрица максимального объема. Тогда*

$$\mathcal{V}\left(\hat{V}\right) \geqslant \max\left((N-r+1)^{-r/2}, \left(\frac{4^r-1}{3}\right)^{-r/2}\right)\mathcal{V}\left(V_M\right). \tag{4.73}$$

Доказательство. Для получения оценок на объем воспользуемся неравенством

$$\mathcal{V}\left(V_M\right)/\mathcal{V}\left(\hat{V}\right) = \mathcal{V}\left(\hat{V}^{-1}V_M\right) = \left|\det\left(\hat{V}^{-1}V_M\right)\right| \leqslant \left(\frac{\left\|\hat{V}^{-1}V_M\right\|_F^2}{r}\right)^{r/2}, \tag{4.74}$$

которое следует из неравенства между средним геометрическим (квадратом модуля определителя) и средним арифметическим (квадратом нормы Фробениуса) для квадратов сингулярных чисел квадратной матрицы $\hat{V}^{-1}V_M$.

Из теоремы 4.11 сразу следует

$$\left\|\hat{V}^{-1}V_M\right\|_F^2 \leqslant \left\|\hat{V}^{-1}V\right\|_F^2 = \left\|\hat{V}^{-1}\right\|_F^2 \leqslant r\,(N-r+1)\,,$$

откуда вместе с (4.74) получаем

$$\mathcal{V}\left(\hat{V}\right) \geqslant (N-r+1)^{-r/2}\,\mathcal{V}\left(V_M\right).$$

Теперь получим оценку, в которую N не входит. Для этого оценим $\max\limits_i \left\|\hat{V}^{+}V_i\right\|_2^2$, где индексом i обозначим i-й столбец матрицы V, $\hat{V} \in \mathbb{C}^{r\times k}$ – подматрица из первых k набранных столбцов.

Обозначим $x_k = \max\limits_i \left\|\hat{V}^+ V_i\right\|_2^2$ и оставим обозначение i для столбца, на котором этот максимум достигается. Пусть добавляемый на $k+1$-м шаге столбец имеет индекс j. Тогда, согласно (4.70)

$$\frac{1 + \left\|\hat{V}_1^{-1} V_{1,j}\right\|_2^2}{\left\|V_{2,j}\right\|_2^2} \leqslant \frac{1 + \left\|\hat{V}_1^{-1} V_{1,i}\right\|_2^2}{\left\|V_{2,i}\right\|_2^2}, \tag{4.75}$$

где мы используем то же разбиение матрицы V, что и в теореме 4.11.

Оценим x_{k+1}, используя (4.75):

$$\begin{aligned}
x_{k+1} &= \left\| \begin{bmatrix} \hat{V}_1^{-1} & \frac{1}{\|V_{2,j}\|_2^2} \hat{V}_1^{-1} V_{1,j} V_{2,j}^* \\ 0 & \frac{1}{\|V_{2,j}\|_2^2} V_{2,j}^* \end{bmatrix} \begin{bmatrix} V_{1,i} \\ V_{2,i} \end{bmatrix} \right\|_2^2 \\
&\leqslant \left(\left\|\hat{V}_1^{-1} V_{1,i}\right\|_2 + \frac{\left\|\hat{V}_1^{-1} V_{1,j}\right\|_2}{\left\|V_{2,j}\right\|_2^2} \left| V_{2,j}^* V_{2,i} \right| \right)^2 + \frac{\left| V_{2,j}^* V_{2,i} \right|^2}{\left\|V_{2,j}\right\|_2^4} \\
&\leqslant \left(\left\|\hat{V}_1^{-1} V_{1,i}\right\|_2 + \frac{\left\|\hat{V}_1^{-1} V_{1,j}\right\|_2}{\left\|V_{2,j}\right\|_2} \left\|V_{2,i}\right\|_2 \right)^2 + \frac{\left\|V_{2,i}\right\|_2^2}{\left\|V_{2,j}\right\|_2^2} \\
&\leqslant \left\|\hat{V}_1^{-1} V_{1,i}\right\|_2^2 + 2 \left\|\hat{V}_1^{-1} V_{1,i}\right\|_2 \frac{\left\|\hat{V}_1^{-1} V_{1,j}\right\|_2}{\left\|V_{2,j}\right\|_2} \left\|V_{2,i}\right\|_2 + \frac{\left\|\hat{V}_1^{-1} V_{1,j}\right\|_2^2}{\left\|V_{2,j}\right\|_2^2} \left\|V_{2,i}\right\|_2^2 + \frac{\left\|V_{2,i}\right\|_2^2}{\left\|V_{2,j}\right\|_2^2} \\
&\leqslant \left\|\hat{V}_1^{-1} V_{1,i}\right\|_2^2 + 2 \left\|\hat{V}_1^{-1} V_{1,i}\right\|_2 \frac{\left\|\hat{V}_1^{-1} V_{1,j}\right\|_2}{\sqrt{1 + \left\|\hat{V}_1^{-1} V_{1,j}\right\|_2}} \sqrt{1 + \left\|\hat{V}_1^{-1} V_{1,i}\right\|_2^2} + \left(1 + \left\|\hat{V}_1^{-1} V_{1,i}\right\|_2^2\right) \\
&\leqslant x_k + 2x_k + (1 + x_k) \\
&= 4x_k + 1.
\end{aligned}$$

С учетом $x_1 \leqslant 1$, получаем, решая рекуррентное соотношение, что

$$x_r \leqslant \frac{4^r - 1}{3},$$

откуда

$$\left\|\hat{V}^{-1} V_M\right\|_F^2 \leqslant r \max_i \left\|\hat{V}^+ V_i\right\|_2^2 \leqslant r x_r = r \frac{4^r - 1}{3}. \tag{4.76}$$

Подставляя (4.76) в (4.74), получаем часть оценки (4.73), не содержащую N. □

Здесь мы жадно набирали столбцы, однако можно также рассмотреть аналогичный алгоритм для удаления столбцов. В [44] предложен алгоритм, работающий за $O\left(N^2 r \log^2 \varepsilon\right)$ операций, где ε – машинная точность (предполагается $r < n \ll N$). Здесь мы предложим алгоритм, не содержащий дополнительного логарифмического множителя, зависящего от точности. Для этого обобщим результат из [44] о пересчете нормы Фробениуса псевдообратной матрицы.

Лемма 4.11. *Для* $B = [\ A\ b\ ] \in \mathbb{C}^{r\times(n+1)}$, $n \geqslant r$, *выполнено*

$$\|B^+\|_F^2 = \|A^+\|_F^2 - \frac{\left\|(AA^*)^{-1}\, b\right\|_2^2}{1 + \|A^+b\|_2^2} = \|A^+\|_F^2 - \frac{\left\|(BB^*)^{-1}\, b\right\|_2^2}{1 - \|B^+b\|_2^2}. \tag{4.77}$$

Доказательство. Воспользуемся формулой пересчета обратной матрицы на основе симметричного обновления ранга 1:

$$(BB^*)^{-1} = (AA^*)^{-1}\left(I - \frac{bb^*\,(AA^*)^{-1}}{1 + \|A^+b\|_2^2}\right). \tag{4.78}$$

Данную формулу легко проверить, умножив правую часть на $BB^* = AA^* + bb^*$, что даст единичную матрицу.

Из (4.78) следует

$$\begin{aligned}\operatorname{tr}(BB^*)^{-1} &= \operatorname{tr}(AA^*)^{-1} - \operatorname{tr}\left((AA^*)^{-1}\, bb^*\,(AA^*)^{-1}\right) / \left(1 + \left\|A^+b\right\|_2^2\right) \\ &= \operatorname{tr}(AA^*)^{-1} - \operatorname{tr}\left(b^*\,(AA^*)^{-1}\,(AA^*)^{-1}\, b\right) / \left(1 + \left\|A^+b\right\|_2^2\right).\end{aligned}$$

Используя тот факт, что для произвольной матрицы $\operatorname{tr}(AA^*)^{-1} = \|A^+\|_F^2$, получаем левое равенство в (4.77).

Аналогично, используя

$$(AA^*)^{-1} = (BB^*)^{-1}\left(I + \frac{bb^*\,(BB^*)^{-1}}{1 - \|B^+b\|_2^2}\right),$$

получаем правое равенство в (4.77). □

Теорема 4.12. *Один шаг жадного набора или удаления столбца из подматрицы* $A \in \mathbb{C}^{r\times n}$ *матрицы* $B \in \mathbb{C}^{r\times N}$, *минимизирующего норму фробениуса псевдообратной к полученной подматрице, занимает* $O(Nr)$ *операций.*

Доказательство. Для определения того, какой столбец брать на следующем шаге, достаточно построить быстрый (за $O(Nr)$) пересчет норм столбцов матрицы $C = A^+B$ и матрицы $D = (AA^*)^{-1}B$. Пересчет D осуществляется путем пересчета $(AA^*)^{-1}$, который уже был описан ранее.

Пересчет нормы столбцов C за $O(Nr)$ описан в разделе 4.2 (алгоритм 4.2). А именно, обозначим через $l_i = \|(A^+B)_i\|_2^2$ квадрат нормы i-го столбца в матрице A^+B. Пусть $A = QX,\ QQ^* = I,\ Q \in \mathbb{C}^{r\times n},\ X \in \mathbb{C}^{r\times r}$. Обозначим $C = X^{-1}B \in \mathbb{C}^{r\times N}$. Тогда

$$l_i = \left\|\left(A^+B\right)_i\right\|_2^2 = \left\|\left(Q^*A^+B\right)_i\right\|_2^2 = \left\|\left(X^{-1}B\right)_i\right\|_2^2 = \|C_i\|_2^2\,.$$

Как показано в разделе 4.2, уравнение (4.4), при добавлении i-го столбца полученную матрицу C' можно записать как

$$C' = C - \left(1 - \frac{1}{\sqrt{1 + l_i}}\right) \frac{C_i C_i^*}{l_i} C,$$

а новые квадраты норм l'_j записываются как (4.5):

$$l'_j = l_j - \frac{\left|C_i C_j^*\right|^2}{1 + l_i}.$$

Аналогично, при удалении столбца (выражая величины без штрихов через величины со штрихами):

$$C = C' + \left(\frac{1}{\sqrt{1 - l'_i}} - 1\right) \frac{C'_i C'^*_i}{l'_i} C'$$

а

$$l_j = l'_j + \frac{\left|C'_i C'^*_j\right|^2}{1 - l'_i}.$$

В обоих случаях формулы пересчета требуют $O(Nr)$ операций. □

Следствие 4.5. Набор n столбцов из нулевой подматрицы занимает $O(Nnr)$ операций. Последовательное удаление столбцов из полной матрицы занимает $O\left(N^2 r\right)$ операций.

Замечание 4.4. Аналогично можно заменять столбцы, чтобы получить подматрицу A' того же размера, что и A, сначала добавляя наилучший i-й столбец, а потом удаляя наихудший j-й. Или воспользоваться критерием замены, см. А.

Теперь, когда у нас есть жадные алгоритмы, выведем оценки для них.

Теорема 4.13. *При последовательном жадном наборе столбцов в ортонормированных строках $U \in \mathbb{C}^{r\times N}$, начиная с подматрицы $\hat{U}_r \in \mathbb{C}^{r\times r}$ для которой верно неравенство $\|\hat{U}_r^{-1}\|_F \leqslant \sqrt{r(N - r + 1)}$, выполнено*

$$\left\|\hat{U}^+\right\|_F \leqslant \sqrt{\frac{r^2}{n}(N - r + 1)} \left\|R^+\right\|_F. \tag{4.79}$$

где новая подматрица $\hat{U}$ имеет n столбцов. Общая сложность набора столбцов составляет $O(Nnr)$.

При последовательном удалении столбцов из произвольных строк $R \in \mathbb{C}^{r\times N}$ до размера $A \in \mathbb{C}^{r\times n}$ справедлива оценка

$$\|A\|_F \leqslant \sqrt{\frac{N - r + 1}{n - r + 1}} \|R\|_F. \tag{4.80}$$

Для унитарной матрицы $R = U$, $A = \hat{U}$, получим

$$\left\|\hat{U}^+\right\|_F \leqslant \sqrt{r\frac{N - r + 1}{n - r + 1}}. \tag{4.81}$$

Общая сложность удаления столбцов составляет $O\left(Nr^2 + Nr\,(N-n)\right)$.

Предполагается, что $n \geqslant r$.

Доказательство. Сначала докажем (4.80), которое уже было доказано другим способом в [44]. Пусть на текущем шаге мы имеем подматрицу B с $M+1$ столбцом. Согласно лемме 4.11, при удалении одного столбца из $B = [\ A\ b\ ] \in \mathbb{C}^{r\times(M+1)}$ имеем

$$\begin{aligned}
\min_b \|A^+\|_F^2 &= \|B^+\|_F^2 + \min_b \frac{\left\|(BB^*)^{-1} b\right\|_2^2}{1 - \|B^+ b\|_2^2} \\
&\leqslant \|B^+\|_F^2 + \frac{\mathbb{E}_b \left\|(BB^*)^{-1} b\right\|_2^2}{1 - \mathbb{E}_b \|B^+ b\|_2^2} \\
&= \|B^+\|_F^2 + \frac{\|B^+\|_F^2}{M+1-r} \\
&= \frac{M-r+2}{M-r+1}\|B^+\|_F^2.
\end{aligned}$$

Перемножая коэффициенты для M от $N-1$ до n, получаем оценку (4.80). Неравенство (4.81) является прямым следствием (4.80), так как $\|U^+\|_F^2 = r$.

Теперь докажем (4.79). Обозначим $x_k = \|\hat{U}_k\|_F^2$, где $\hat{U}_k \in \mathbb{C}^{r\times k}$ получена из $\hat{U}_r$ за $k-r$ добавлений столбцов. Согласно лемме 4.11, x_{k+1} оценивается как

$$\begin{aligned}
x_{k+1} &= x_k - \min_{b\in U/\hat{U}_k} \frac{\left\|\left(\hat{U}_k\hat{U}_k^*\right)^{-1} b\right\|_2^2}{1 + \left\|\hat{U}_k^+ b\right\|_2^2} \\
&\leqslant x_k - \frac{\mathbb{E}_{b\in U/\hat{U}_k} \left\|\left(\hat{U}_k\hat{U}_k^*\right)^{-1} b\right\|_2^2}{1 + \mathbb{E}_{b\in U/\hat{U}_k} \left\|\hat{U}_k^+ b\right\|_2^2} \\
&= x_k - \frac{\frac{\left\|\left(\hat{U}_k\hat{U}_k^*\right)^{-1}U\right\|_F^2 - \left\|\left(\hat{U}_k\hat{U}_k^*\right)^{-1}\hat{U}_k\right\|_F^2}{N-k}}{1 + \frac{\|\hat{U}_k^+ U\|_F^2 - \|\hat{U}_k^+\hat{U}_k\|_F^2}{N-k}} \\
&= x_k - \frac{\left\|\left(\hat{U}_k\hat{U}_k^*\right)^{-1}\right\|_F^2 - x_k}{N-k+(x_k-r)} \\
&\leqslant x_k - \frac{x_k^2/r - x_k}{N-k+(x_k-r)} \\
&= x_k\left(1 - \frac{(x_k-r)/r}{N-k+(x_k-r)}\right),
\end{aligned} \tag{4.82}$$

где мы учли, что $\left\|\left(U_kU_k^*\right)^{-1}\right\|_F^2 \geqslant \frac{1}{r}\left\|U_k^+\right\|_F^4 = x_k^2/r$.

Заметим, что правая часть (4.82) монотонно возрастает при росте x_k начиная с $x_k \geqslant r$ (проверяется, например, заменой $x_k = r + \alpha r\,(N-k)$ и дифференцированием по α с учетом

$k < N$). Таким образом, можно заменить x_k на его верхнюю границу (4.79), которую мы возьмем за предположение индукции (база $k = r$ верна по условию) и далее будем использовать (4.82), чтобы доказать шаг индукции по k.

При любом фиксированном x_k вне скобок в правой части (4.82), выражение в скобках $1 - \frac{(x_k - r)/r}{N-k+(x_k-r)}$, наоборот, убывает с ростом $x_k \geqslant r$, поэтому мы только огрубим оценку, если внутри скобок заменим x_k на не большую величину $r + \frac{r}{k-r+1}(N-k) \leqslant \frac{r^2}{k}(N-k+1)$. После подстановки получим

$$\begin{aligned} x_{k+1} &\leqslant \frac{r^2}{k}(N-r+1)\left(1 - \frac{\frac{1}{k-r+1}(N-k)}{N-k+\frac{r}{n-r+1}(N-k)}\right) \\ &= \frac{r^2}{k}(N-r+1)\left(1 - \frac{1}{k+1}\right) \\ &= \frac{r^2}{k+1}(N-r+1)\,. \end{aligned}$$

Таким образом, предположение индукции верно и на следующем шаге, что доказывает (4.79). □

При этом оценка на отношение спектральных норм в таблице 4.1 следует из

$$\frac{\left\|\hat{R}^+\right\|_2^2}{\|R^+\|_2^2} = \frac{\left\|\hat{R}^+\right\|_F^2 - \left(\left\|\hat{R}^+\right\|_F^2 - \left\|\hat{R}^+\right\|_2^2\right)}{\|R^+\|_2^2} \leqslant \frac{\left\|\hat{R}^+\right\|_F^2 - \left(\|R^+\|_F^2 - \|R^+\|_2^2\right)}{\|R^+\|_2^2} = 1 + \left(\frac{\left\|\hat{R}^+\right\|_F^2}{\|R^+\|_F^2} - 1\right)\cdot\frac{\|R^+\|_F^2}{\|R^+\|_2^2},$$

куда затем подставляется оценка $\frac{\left\|\hat{R}^+\right\|_F^2}{\|R^+\|_F^2} \leqslant \frac{N-r+1}{n-r+1}$.

4.7. Жадный обмен столбцов для максимизации объема

В данном разделе покажем, что алгоритм rect-maxvol для набора столбцов может быть эффективно использован и для удаления столбцов, а также приводить к быстрой сходимости с точки зрения норм столбцов $\hat{R}^+R$ (предпоследняя строка таблицы 4.1). Для этого нам потребуется следующая теорема, обобщающая результат автора из [60, 83] (схожий результат был позже независимо доказан в [101]).

Теорема 4.14. *Пусть $\hat{A} \in \mathbb{C}^{n\times r}$ – подматрица в $A \in \mathbb{C}^{N\times r}$ ρ-локально максимального объема. Пусть $A_M \in \mathbb{C}^{n\times r}$ – подматрица максимального объема в A. Тогда*

$$\mathcal{V}(A_M) \leqslant \mathcal{V}\left(\hat{A}\right)\cdot\left(\frac{n+1}{n-r+1}\left(1+\left(\rho^2-1\right)\frac{n}{r}\right)\right)^{r/2}. \tag{4.83}$$

Если $\frac{r+(\rho^2-1)n}{n-r+1} \geqslant 1$ (это, в частности, выполнено всегда, когда $n \leqslant 2r-1$), то

$$\mathcal{V}(A_M) \leqslant \mathcal{V}\left(\hat{A}\right)\cdot\left(\frac{n}{n-r+1}\left(1+\left(\rho^2-1\right)\frac{n}{r}\right)\right)^{r/2}. \tag{4.84}$$

Доказательство. Сначала докажем результат для $n \leqslant 2r-1$. Согласно теореме 4.1, для произвольной строки $a \in \mathbb{C}^r$ из A верно

$$\left\|a\hat{A}^+\right\|_2 \leqslant \frac{r+\left(\rho^2-1\right)n}{n-r+1}.$$

Данное неравенство распространяется и на тот случай, когда a – строка $\hat{A}$, поскольку любая строка $\hat{A}\hat{A}^+$ по норме не больше 1:

$$\left\|a\hat{A}^+\right\|_2^+ \leqslant 1 \leqslant \frac{r+\left(\rho^2-1\right)n}{r} \leqslant \frac{r+\left(\rho^2-1\right)n}{n-r+1},$$

где в конце мы воспользовались $n \leqslant 2r-1$.

Отсюда

$$\left\|\hat{A}_M\hat{A}^+\right\|_F^2 \leqslant n\frac{r+\left(\rho^2-1\right)n}{n-r+1}.$$

Максимальный объем достигается, когда все сингулярные числа равны, что приводит к оценке

$$\mathcal{V}\left(A_M\right)/\mathcal{V}\left(\hat{A}\right) = \mathcal{V}\left(A_M\hat{A}^+\right) \leqslant \left(\frac{n}{r}\cdot\frac{r+\left(\rho^2-1\right)n}{n-r+1}\right)^{r/2} = \left(\frac{n}{n-r+1}\left(1+\left(\rho^2-1\right)\frac{n}{r}\right)\right)^{r/2}.$$

Точно такое же рассуждение можно провести и для $n > 2r-1$, если A_M и $\hat{A}$ не содержат общих строк.

Теперь рассмотрим случай, когда A_M и $\hat{A}$ содержат $k > 0$ общих строк.

Согласно лемме 4.4 при замене только одной строки (что даст какую-то подматрицу $\tilde{A}$) выполнено неравенство

$$\rho^2 \geqslant \mathcal{V}\left(\tilde{A}^2\right)/\mathcal{V}^2\left(\hat{A}\right) = \left|C_{i,j}\right|^2 + \left(1+\|C_i\|_2^2\right)\left(1-\left\|C_j\right\|_2^2\right) = \left|C_{i,j}\right|^2 + \left(1+l_i\right)\left(1-l_j\right), \qquad (4.85)$$

где $C = A\hat{A}^+$, j – индекс заменяемого столбца $\hat{A}$, а i – индекс новой строки.

Пусть $x = \max\limits_i l_i$. Поскольку все выражение (4.85) не превосходит ρ^2, то и второе слагаемое $\left(1+\|C_i\|_2^2\right)\left(1-\left\|C_j\right\|_2^2\right) = \left(1+l_i\right)\left(1-l_j\right)$ также не больше ρ^2. Подставляя $l_i = x$, находим, что

$$\forall j = \overline{1,n}: \quad (1-l_j)(1+x) \leqslant \rho^2,$$

то есть

$$\forall j = \overline{1,n}: \quad l_j \geqslant \max\left(0, 1-\frac{\rho^2}{1+x}\right). \qquad (4.86)$$

Напомним наше предположение, что A_M содержит $k > 0$ строк из $\hat{A}$. Сумма квадратов длин всех строк C, соответствующих $\hat{C}$ (то есть строк $\hat{C} = \hat{A}\hat{A}^+$) равна r. С другой стороны, сумма квадратов норм оставшихся $n-k$ строк согласно (4.86) не меньше $(n-k)\max\left(0, 1-\frac{\rho^2}{1+x}\right)$. Поэтому норма Фробениуса общих строк не превосходит

$$r-(n-k)\max\left(0, 1-\frac{\rho^2}{1+x}\right).$$

Тогда квадрат нормы Фробениуса матрицы $C_M = A_M \hat{A}^+$ ограничен величиной

$$\|C_M\|_F^2 \leqslant (n-k)x + \left(r - (n-k)\max\left(0, 1 - \frac{\rho^2}{1+x}\right)\right) \leqslant r + (n-k)\frac{x^2 - 1 + \rho^2}{x+1}, \quad 1 - \frac{\rho^2}{1+x} \geqslant 0. \tag{4.87}$$

По k максимум достигается при $k = 1$. По x вторая производная неотрицательна, так что максимум достигается на границе допустимых x. Левой границей является случай, когда $1 - \frac{\rho^2}{1+x} = 0$, то есть $x = \rho^2 - 1$, что дает оценку

$$\|C_M\|_F^2 \leqslant r + (n-1)\left(\rho^2 - 1\right),$$

что ниже оценки для случая $k = 0$.

Правой границей для x согласно теореме 4.1 является

$$x = \frac{r + \left(\rho^2 - 1\right)n}{n - r + 1}.$$

Сразу заметим, что если $x > 1$, то $k = 0$ оптимально (выгодно брать строки большей длины), что дает ту же оценку, что и при $n \leqslant 2r - 1$.

Подставляя это значение в (4.87) (в знаменателе используем $\rho = 1$ в качестве оценки всего выражения сверху), получаем

$$\begin{aligned}
\|C_M\|_F^2 &\leqslant r + \frac{r^2(n-1)}{(n-r+1)(n+1)} + \frac{2\left(\rho^2-1\right)nr(n-1)}{(n-r+1)(n+1)} + \frac{\left(\rho^2-1\right)^2 n^2 (n-1)}{(n-r+1)(n+1)} \\
&\leqslant r\frac{n+1+2\left(\rho^2-1\right)(n-1) + \left(\rho^2-1\right)^2 n(n-1)/r}{n-r+1} \\
&\leqslant r\frac{n+1}{n-r+1}\left(1 + 2\left(\rho^2-1\right) + \left(\rho^2-1\right)^2\frac{n-1}{r}\right).
\end{aligned}$$

Воспользовавшись неравенствами $x \leqslant 1$, можно оценить $\rho^2 \leqslant 2 - \frac{2r-1}{n}$ и избавиться от $(\rho^2-1)^2$:

$$\|C_M\|_F^2 \leqslant r\frac{n+1}{n-r+1}\left(1 + (\rho^2-1)\frac{n}{r}\right)$$

Оценивая отношение объемов по аналогии со случаем $n \leqslant 2r - 1$, получаем (4.83). □

Замечание 4.5. В доказательстве были использованы только два факта:

$$\max_i l_i \leqslant \frac{r + (\rho^2 - 1)n}{n - r + 1} \tag{4.88}$$

и

$$\left(1 - \min_j l_j\right)\left(1 + \max_i l_i\right) \leqslant \rho^2. \tag{4.89}$$

Они включают в себя только l_i и l_j, которые пересчитываются за $O(Nr)$ с помощью быстрой версии rect-maxvol (алгоритм 4.2). Таким образом, если отношение объемов превышает границу теоремы, то объем можно увеличить минимум в ρ раз с помощью замены, которую можно найти за $O(Nr)$ операций. Это позволяет быстро достичь объема, близкого к максимальному, а также постоянной ρ, близкой к 1.

В обоих случаях (4.88) и (4.89) замена происходит на столбец максимальной длины l_i, то есть другие значения i рассматривать не требуется. В итоге получаем следующий алгоритм 4.12, позволяющий заменять столбцы за $O(Nr)$, чтобы в итоге гарантировать (4.88), (4.89) и (4.83) или (4.84).

Алгоритм 4.12

Вход: Строки $R \in \mathbb{C}^{r\times N}$, стартовый набор индексов столбцов $\mathcal{I}$ размера n, параметр $\rho \geqslant 1$.

Выход: Подматрица $\hat{A} = R_{:,\mathcal{I}}$ большого объема.

1: $Y := I$

2: $R_{:,\mathcal{I}} = LQ$

3: $C := L^{-1}R$

4: $\hat{C} = Q$

5: **for** $j := 1$ **to** N **do**

6: $\quad l_j := \|C_{:,j}\|_2^2$

7: **end for**

8: $i := \arg\max\limits_{i, i\notin\mathcal{I}} l_i$

9: **while** $\max\limits_{j, j\in\mathcal{I}} \left(\left|C^*_{j,:} Y C_{:,i}\right|^2 + \left(1 - l_j\right)\left(1 + l_i\right)\right) > \rho^2$ **do**

10: $\quad$ Индекс j из $\mathcal{I}$ меняется на i

11: $\quad$ Обновление $Y = \left(\hat{C}\hat{C}^*\right)^{-1}$

12: $\quad$ Обновление l_j, $j = \overline{1, N}$

13: **end while**

Использование данного алгоритма вместо алгоритма $\mathrm{dominant}$ приводит к вычислительной сложности $O\left(Nnr + Nr^2 \log_\rho n\right)$ (оценка на число шагов остается той же) для достижения (4.88). Если требуется оценка с $\rho' = 1 + \frac{r}{n}\rho$ (как в таблице 4.1), то, подставляя ρ' в оценку сложности, получаем сложность $O\left(Nnr + Nnr \log_\rho n\right)$.

Утверждение 4.16. *Пусть $n \leqslant 2r - 1$, тогда за $O\left(Nr^2 \log\log r + Nr^2/\log\rho\right)$ можно найти подматрицу, объем которой отличается от максимального не более, чем в $\rho^r \sqrt{nC^n_{2r-1}/r}$ раз.*

При $n = r$ отличие от максимального объема будет менее, чем в $(2\rho)^r / (4\pi r)^{1/4}$ раз. Это улучшает результат для локально максимального объема из [11] и является альтернативой полиномиальному алгоритму из [102], который позволяет достичь отношения к максимальному объему не более, чем в $e^{r/2+o(r)}$ раз. Такая стартовая подматрица может быть полезна, например, при построении минимального охватывающего эллипсоида [76]. Заметим, что за полиномиальное время (если P $\neq$ NP) для прямоугольных столбцов (в $ar \times br$ матрице, $a, b > 1$) нельзя гарантировать отличие от максимального объема менее, чем экспоненциальное по r, что доказано в [103].

Доказательство. Выберем стартовую подматрицу, как для $\mathrm{dominant}$, так что отношение максимального объема к начальному не более $(kr)^{r/2}$ для подматрицы размера $r \times k$. Пусть V_s – отношение максимального объема к текущему после s шагов. Тогда текущее максимальное отношение V_{s+1}/V_s определяется формулой (4.83), откуда получаем

$$\frac{k-r+1}{k+1} V_s^{2/r} \leqslant 1 + \left(\frac{V_s^2}{V_{s+1}^2} - 1 \right) \frac{k}{r} \leqslant\leqslant \frac{V_s^2}{V_{s+1}^2} \cdot \frac{k}{r}.$$

Отделив V_{s+1} в левую часть, получаем

$$V_{s+1} \leqslant \sqrt{\frac{k(k+1)}{r(k-r+1)}} V_s^{1-1/r}.$$

Обозначим $V'_s = V_s \left(\frac{k(k+1)}{r(k-r+1)} \right)^{-r/2}$, что приведет к

$$V'_{s+1} \leqslant \left(V'_s \right)^{1-1/r}.$$

Начальное значение равно $V'_0 = \left(\frac{r^2(k-r+1)}{k+1} \right)^{r/2} \leqslant r^r$. В качестве конечного значения выберем, например $V_s = e^r$. Тогда число шагов оценивается как

$$\begin{aligned} r^{r \cdot (1-1/r)^s} &\leqslant e^r, \\ r \cdot (1-1/r)^s \log r &\leqslant r, \\ e^{-s/r} &\leqslant \log^{-1} r, \\ s &\leqslant r \log\log r. \end{aligned}$$

После этого будем менять столбцы, пока норма каждого окажется не больше $\frac{cr}{k-r+1}$. Это потребует еще $\log_{1+(\rho^2-1)\frac{r}{k}} \left(e^2 \frac{k(k+1)}{r(k-r+1)} \right)^{r/2} = O\left(r/\log\rho\right)$ шагов при $k = 2r-1$. После этого квадрат длины каждого столбца не больше ρ^2, а потому отличие объема текущей $r \times 2r-1$ матрицы к объему матрицы размера $r \times r$ не больше ρ^r. Удалив $(2r-1)-n$ столбцов (каждый раз жадно выбирая столбец так, чтобы максимизировать объем оставшейся подматрицы), мы уменьшим объем не более, чем в $\sqrt{C_{2r-1}^n}$ раз, что завершает доказательство. □

Следствие 4.6. За $O\left(Nr^2 \log\log r + Nnr\right)$ операций можно достичь подматрицы, объем которой отличается от максимального не более, чем в $6^{r/2}$ раз.

Доказательство. Для $n \leqslant 2r-1$ мы уже получили оценку не хуже в утверждении 4.16. Пусть теперь $n > 2r-1$, и мы (согласно утверждению 4.16) начинаем с $r \times (2r-1)$ подматрицы $6^{r/2}$-максимального объема (для чего достаточно выбрать в нем $\rho \leqslant \sqrt{3}$). Будем далее набирать жадно столбцы. Если при этом всегда на шаге k найдется $l_i \geqslant r/(k-r+1)$, то объем будет расти в $1+l_i \geqslant (k+1)/(k-r+1)$ раз. С другой стороны, объем подматрицы с $k+1$ столбцами максимального

объема превосходит объем подматрицы с k столбцами не более, чем в $(k+1)/(k-r+1)$, что мы уже ни раз проверяли (например, при доказательстве теоремы 2.3). Таким образом, в этом случае отношение объема к максимальному не растет, и в итоге мы получим $6^{r/2}$-максимальный объем.

Пусть теперь на некотором шаге $l_i \leqslant r/(k-r+1)$, а после этого шага неравенство снова выполнено в другую сторону (если еще шаги останутся). В этом случае достаточно показать, что $l_i \leqslant r/(k-r+1)$ гарантирует 6^r-максимальный объем, так как далее отношение не вырастет (не считая этого одного шага). Итак, если $l_i \leqslant r/(k-r+1)$, то уравнение (4.89) выполнено для $\rho^2 = 1+l_i$. Поэтому итоговое отношение квадратов объемов (согласно теореме 4.14) оценивается как $\left(\frac{k+1}{k-r+1}\left(1+l_i\frac{k}{r}\right)\right)^r \cdot (1+l_i)^{\frac{k+1}{k-r+1}} \leqslant \left(\frac{k+1}{k-r+1}\left(1+\frac{k}{k-r+1}\right)\right)^r \leqslant \left(\frac{2r}{r}\left(1+\frac{2r-1}{r}\right)\right)^r \leqslant 6^r$. □

Отсюда сразу следует, что при оценке числа шагов исчезнет множитель $\log n$, что и записано в таблице 4.1. Несмотря на это, на практике число шагов почти всегда существенно меньше r после выполнения жадного набора. Например, при поиске 50×50 подматрицы локально максимального объема в случайных ортонормированных строках $U \in \mathbb{R}^{50\times5000}$ требуется в среднем 1,2 замены (среднее за 100 генераций). То есть из 50 жадно набранных столбцов в среднем достаточно заменить всего 1, чтобы достичь локально максимального объема. Поэтому набирать 100, а потом сокращать до 50 столбцов ради потенциально возможного худшего случая на практике оказывается невыгодно.

Оценку можно существенно улучшить, если разрешить выбирать те же столбцы по несколько раз. Такая оценка уже была получена в [101]. Здесь мы покажем, что можно обойтись без $r\log\log r$ лишних столбцов (поскольку мы уже заранее сделали соответствующие замены), немного улучшим точность оценки, а также покажем, как можно использовать её, чтобы найти $r\times r$ подматрицу $e^{r/2}$-максимального объема.

Утверждение 4.17. *За $O\left(Nr^2\log\log r + Nr^2\log\varepsilon^{-1}/\varepsilon\right)$ операций в матрице $R \in \mathbb{C}^{r\times N}$ можно найти $r\times r$ подматрицу $e^{r/2+r\varepsilon}/(2\pi r)^{1/4}$-максимального объема, $\varepsilon \leqslant$ const.*

Доказательство. Как и ранее, начнем с $2r-1$ столбцов с $\rho^2 = 2$, которых можно быстро достичь с помощью утверждения 4.16.

Раз квадрат длины каждого столбца теперь не больше 2, то отношение объема относительно $r\times n$ матрицы максимального объема не больше $\left(2\frac{n}{r}\right)^{r/2}$. Но нас будет интересовать именно изменение гарантированного минимального значения квадрата максимальной длины, которое равно 2. Если для $r\times n$ подматрицы оно достигнет некоторого числа l_n, то отношение максимального объема к данному будет $\left(l_n\frac{n}{r}\right)^{r/2}$. Само l_n меняется также, как и объем:

$$l_{k+1} \leqslant l_k/(1+l_k)^{1/r}.$$

Введя $x = k/r$ получаем, что l_k меняется быстрее, чем при следующем дифференциальном уравнении, которое соответствует $r \to \infty$:

$$\frac{d \ln l(x)}{dx} = -\ln(1 + l(x)) \leqslant -l + l^2/2.$$

Его решение при $l(0) = 1$ удовлетворяет неравенству

$$l \leqslant \frac{1}{x + 1 + \frac{1}{2}\ln\left(\frac{l}{2-l}\right)} \leqslant \frac{1}{x + 1 - \frac{1}{2}\ln(1/2 + x/2)}.$$

Решив изначальное уравнение с единицей, получаем, что достижение $l = 1$ потребует не более $\lceil 0{,}8r \rceil \leqslant r$ столбцов, так что можно выбрать $x = \frac{n}{r} - \frac{2r-1}{r} - \frac{r}{r} = \frac{n+1}{r} - 3$. В итоге получаем отношение к максимальному объему не больше

$$\frac{\mathcal{V}(A_M)}{\mathcal{V}\left(\hat{A}\right)} \leqslant \left(\frac{1}{\frac{n+1}{r} - 2 - \frac{1}{2}\ln\left(\frac{n+1}{2r} - 1\right)} \cdot \frac{n}{r}\right)^{r/2}.$$

Эта формула верна для $n \geqslant 2r - 1 + \lceil 0{,}8r \rceil$. В противном случае можно пользоваться оценкой следствия 4.6.

Разность логарифмов максимального и текущего объема в этом случае $O\left(r^2 \log n/n\right)$. Таким образом, используя далее $\rho = 1 + \frac{r^2}{n^2}$, достигаем $\left(\frac{n+r}{n-r+1}\right)^{r/2}$-максимальный объем за $O\left(Nnr \log_\rho n\right)$ ($\rho \leqslant \text{const}$). Жадно удаляя столбцы вплоть до r (что потребует не больше времени), получаем $r \times r$ подматрицу, объем которой отличается от максимального не более, чем в

$$\sqrt{C_n^r \left(\frac{r(1 + r/n)}{n - r + 1}\right)^r} \leqslant e^{r/2} \cdot \left(1 + \frac{2r-1}{n-r+1}\right)^{r/2} / (2\pi r)^{1/4}$$

раз.

При $n \sim r/\varepsilon$ получаем требуемый результат. □

Это немного улучшает результат $e^{r/2+o(r)}$ из [102], позволяя, если требуется, избавиться от $o(r)$ за полиномиальное время, а также представляет из себя эффективный алгоритм, применимый на практике. После него можно также применить алгоритм maxvol, найдя подматрицу $1 + \varepsilon/\log \varepsilon^{-1}$-локально максимального объема, что не испортит асимптотику.

4.8. Связь с поиском минимального охватывающего эллипсоида

В задачах статистики [104], кластеризации и распознавания образов [105] и компьютерной графике [106] может потребоваться упростить представление некоторого большого набора точек $x_i \in \mathbb{R}^r$, $i = \overline{1, N}$, чтобы с ним было далее удобно работать. Например, было бы удобно заменить их выпуклую оболочку на некоторое другое гладкое выпуклое множество, принадлежность

которому легко проверять, и внутри которого легко проводить оптимизацию. Подходящим кандидатом в этом случае является многомерный эллипсоид минимального объема, содержащий внутри себя все точки. Такой выбор также привлекателен со статистической точки зрения: если x_i получены из некоторого многомерного распределения Гаусса, то эллипсоид будет достаточно хорошо приближать такое распределение, поскольку линии уровня распределения Гаусса являются эллипсоидами. Соответствующая задача называется задачей поиска минимального охватывающего эллипсоида.

Минимальный охватывающий эллипсоид можно найти с помощью симметричной положительно определенной матрицы матрицы $H \in \mathbb{R}^{r\times r}$, являющейся решением

$$\begin{aligned} &-\ln\det H \to \min, \\ &x_i^T H x \leqslant 1, \quad i = \overline{1,N}, \end{aligned} \tag{4.90}$$

где $x_i \in \mathbb{R}^r$ – набор точек в r-мерном пространстве, а фактор L разложения Холецкого $H = LL^T$ задает преобразование единичного шара в эллипсоид, содержащий все точки x_i. Здесь мы рассматриваем центрально симметричный случай, так как к нему можно свести общую задачу с заранее неизвестным центром $\tilde{x} \in \mathbb{R}^r$ с помощью следующего преобразования [76]:

$$Y = \frac{1}{r+1}\begin{bmatrix} rX \\ 1_N \end{bmatrix} \in \mathbb{R}^{(r+1)\times N},$$

где Y является новым множеством точек, а искомая матрица H является ведущей $r \times r$ подматрицей $H_Y \in \mathbb{R}^{(r+1)\times(r+1)}$.

В задаче (4.90) вместо самого определителя оптимизируется его логарифм, поскольку в этом случае оптимизируемая функция является строго выпуклой.

Задаче (4.90) также соответствует следующая двойственная задача:

$$\begin{aligned} &\ln\det\left(XUX^T\right) \to \max, \\ &\operatorname{tr} U = r,\ U \geqslant 0, \end{aligned} \tag{4.91}$$

где мы объединили векторы x_i в матрицу $X \in \mathbb{R}^{r\times N}$, а $U \in \mathbb{R}^{N\times N}$ – диагональная матрица. В этом случае $H = XUX^T$ является решением прямой задачи.

Интересно также отметить, что решение данной задачи минимизирует максимальную взвешенную норму столбцов X (при тех же дополнительных условиях) [107]

$$U = \arg\min_U \max_i \left\| \left(X^T U X\right)^{-1/2} x_i \right\|_2,$$

что напоминает по форме свойства подматриц локально максимального объема в лемме 1.3. Данная связь становится еще более явной, если учесть, что матрицу $H = X^T U X$ можно задать с помощью $n \leqslant r(r+1)/2$ параметров: достаточно использовать n точек, формирующих некоторую

подматрицу $\hat{X} \in \mathbb{R}^{r\times n}$, а все остальные элементы U будут нулевыми. Такой набор столбцов, дающих точное (или, далее, приближенное) решение называется базовым (core set). Стоит отметить, что базовые наборы небольшого размера позволяют строить столбцовые аппроксимации высокой точности в p-норме [97]. В частности, предложенный здесь способ набора столбцов дает лучшие гарантии на их количество, чем алгоритмы из [76], а потому также улучшает соответствующие оценки из [97].

Точное решение как прямой, так и двойственной задачи найти достаточно тяжело, поэтому вместо точного решения обычно ищется приближенное.

Определение 4.2. Матрица H задает ε-оптимальный охватывающий эллипсоид, если

$$\left(X^T H^{-1} X\right)_{ii} \leqslant 1 + \varepsilon \ \forall i,$$

а сама матрица H представима в виде

$$H = XUX^T,$$

где $U \in \mathbb{R}^{M\times M}$, $\operatorname{tr} U = r$, $U \geqslant 0$ является диагональной матрицей. То есть U является допустимым в двойственной задаче (4.91).

Если ненулевые элементы U соответствуют подматрице $\hat{X} \in \mathbb{R}^{r\times N}$ матрицы X, то такую подматрицу вместе с соответствующими ей весами $S \in \mathbb{R}^{n\times n}$, $S = \hat{U}^{1/2}$, будем называть ε-решением.

Из двойственности задач получаем, что взвешенный в $\sqrt{1+\varepsilon}$ раз ε-оптимальный эллипсоид не более, чем в $(1+\varepsilon)^{r/2}$ раз больше оптимального (по объему).

Из определения также следует, что ε-решению с диагональной матрицей S необходимо и достаточно удовлетворять условиям

$$\left\|\left(\hat{X}S\right)^+ X_i\right\|_2^2 \leqslant 1 + \varepsilon \ \forall i, \|S\|_F^2 = r.$$

Заметим, что данным условиям будет удовлетворять подматрица локально максимального объема, если в S записывать сколько раз мы берем одинаковые столбцы.

Утверждение 4.18. *Для любого n существует базовый набор из не более, чем n столбцов, на основе которого можно построить ε-решение с $\varepsilon \leqslant \frac{r-1}{n-r+1}$.*

Доказательство. Выберем подматрицу $\hat{Y} \in \mathbb{R}^{r\times n}$ локально максимального объема в матрице $Y = [X\ X\ \ldots X] \in \mathbb{R}^{r\times \lfloor \frac{n}{r} \rfloor N}$. Пусть nS_{kk}^2/r равно числу раз k-й столбец $\hat{X}$ присутствует в $\hat{Y}$. Если Y содержит копию некоторого столбца $X_{:,i}$ вне $\hat{Y}$, то согласно лемме 1.3

$$\left\|\left(\hat{X}S\right)^+ X_{:,i}\right\|_2^2 = \frac{n}{r}\left\|\left(\hat{X}\sqrt{n}S/\sqrt{r}\right)^+ X_{:,i}\right\|_2^2 = \frac{n}{r}\left\|\hat{Y}^+ X_{:,i}\right\|_2^2 \leqslant \frac{n}{n-r+1}.$$

Если на данном столбце i достигается максимум, то левая часть равна $1+\varepsilon$, откуда находим $\varepsilon \leqslant \frac{r-1}{n-r+1}$.

Если же некоторый столбец $X_{:,i}$ встречается в $\hat{Y}$ все $\left\lfloor \frac{n}{r} \right\rfloor$ раз, так что его копии нет вне $\hat{Y}$, то

$$\left\|\hat{Y}^{+} X_{:,i}\right\|_2^2 \leqslant \left\|\left(\left\lfloor \frac{n}{r} \right\rfloor X_{:,i}\right)^{+} X_{:,i}\right\|_2^2 = \frac{1}{\left\lfloor \frac{n}{r} \right\rfloor} \leqslant \frac{r}{n-r+1},$$

что дает в итоге ту же самую оценку на ε. □

Замечание 4.6. Поскольку нет необходимости выбирать тот же столбец более $\left\lfloor \frac{n}{r} \right\rfloor$, то получаем

$$\|S\|_C^2 \leqslant \left\lfloor \frac{n}{r} \right\rfloor.$$

Данный результат асимптотически превосходит оценку из [76] (см. также вторую строку таблицы 4.1), поскольку не содержит слагаемого порядка $r \log\log r$. Однако, формально мы не можем гарантировать достижение локально максимального объема за полиномиальное время. Поэтому далее покажем, как этой оценки можно достичь алгоритмически. К сожалению, при этом число столбцов возрастет примерно в 2 раза, и нам дополнительно все равно потребуется порядка $r \log\log r$ шагов.

Для этого воспользуемся следующей леммой, доказанной в [76].

Лемма 4.12 ([76])**.** *Пусть* $\max\limits_i \left\|\left(\hat{X}S\right)^{+} X_{:,i}\right\|_2^2 = l$*. Тогда добавление* i*-го столбца в базовый набор с определенным весом увеличивает объем текущего* эллипсоида *не менее, чем в* $le^{1/l-1}$ *раз.*

Утверждение 4.19. ε*-решение для* $\varepsilon \leqslant 3$ *можно построить за* $O\left(Nr^2 \log\log r + Nr^2/\varepsilon\right)$ *операций с помощью* $n \leqslant \left\lceil 3{,}45r + \frac{4r}{\varepsilon} + \frac{2}{3} r \log \varepsilon^{-1} \right\rceil$ *столбцов. При этом кроме добавления дополнительно потребуется не более* $\left\lceil r \ln\log_2 r + 2r \right\rceil$ *замен столбцов.*

Доказательство. Первые $2r$ столбцов наберем жадно. После первых r столбцов отношение текущего объема к максимальному объему среди $r \times 2r$ подматриц будет не больше $V_{\max}/V(r) \leqslant (2r)^{r/2}$. Обозначим $l = (V_{\max}/V)^{2/r}$ и $x = k/r$, где k – текущее число столбцов. Согласно лемме 1.2 каждый новый столбец уменьшает отношение к максимальному объему хотя бы в $1 + l/2$ раз, так как точно найдется столбец с квадратом длины не меньше $l/2$ (иначе отношение объемов было бы меньше). Тогда на $l(x)$ получаем дифференциальное уравнение

$$\frac{d \ln l}{dx} \leqslant -\ln(1 + l/2),$$
$$\frac{dl}{dx} \leqslant -l \ln(1 + l/2) \leqslant -\frac{l}{2/l + 1/2}.$$

Для $l(1)$, соответствующего $2r$ столбцам, при $l(0) = 2r$ получаем

$$l(1) \leqslant \frac{4}{W\left(\frac{2e^{2+2/r}}{r}\right)} \leqslant \frac{2r-4}{e^2} + 4.$$

Итого отношение объема к максимальному составляет

$$V_{\max}/V(2r) \leqslant l^{r/2}(1) \leqslant \left(\frac{2r-4+4e^2}{e^2}\right)^{r/2}.$$

Далее поступим аналогично доказательству утверждения 4.16. При $k = 2r$ столбцах имеем

$$V_{s+1} \leqslant 2V_s^{1-1/r}.$$

При $V'_s = 2^{-r/2}V_s$ получаем

$$V'_{s+1} \leqslant V_s^{1-1/r}.$$

Начальное значение равно $V'_0 = \left(\frac{r-2+2e^2}{e^2}\right)^{r/2}$. В качестве конечного выберем $V'_s = 2^{r/2}$ (считаем $r \geqslant 2$). Тогда число замен до достижения V'_s оценивается как

$$\left(\frac{r-2+2e^2}{e^2}\right)^{r/2\cdot(1-1/r)^s} = 2^{r/2},$$
$$r/2 \cdot (1-1/r)^s \ln\frac{r-2+2e^2}{e^2} = r\log 2/2,$$
$$e^{-s/r}\log\frac{r-2+2e^2}{e^2} = \ln 2,$$
$$s = r\ln\frac{\ln\left(r-2+2e^2\right)-2}{\ln 2}$$

с округлением вверх. После этого до достижения $\rho = \sqrt{2}$ каждая замена увеличивает объем в $\sqrt{2}$ раз, поэтому потребуется еще $\log_{\sqrt{2}} 2^r \leqslant 2r$ замен. Итого мы сделали $\left\lceil r\ln\frac{\ln\left(r-2+2e^2\right)-2}{\ln 2}\right\rceil + 2r \leqslant \left\lceil r\ln\log_2 r + 2r\right\rceil$ замен.

Теперь пусть $V_{\max}$ – объем эллипсоида минимального объема. Используя то же обозначение для l, используя утверждение 4.18, получаем, что для любого конкретного отношения l найдется столбец $X_{:,i}$ с квадратом длины хотя бы l (иначе отношение объемов было бы меньше), причем для $2r$ столбцов мы получили $l_0 = 4$.

Теперь начнем набирать столбцы, используя лемму 4.12. Снова составим соответствующее дифференциальное уравнение на l:

$$\frac{d\ln l}{dx} \leqslant -\log l - \frac{1}{l} + 1. \tag{4.92}$$

На этот раз нам известно начальное значение $l(0) = 4$ и конечное значение $l(x) = 1+\varepsilon$. Решением

(4.92) тогда будет

$$\begin{aligned} x &= \int\limits_{1+\varepsilon}^{4} \frac{dl}{l\log l - l + 1} \\ &= \int\limits_{\varepsilon}^{3} \frac{dt}{(1+t)\log(1+t) - t} \\ &= \int\limits_{\varepsilon}^{3} \left(\frac{1}{(1+t)\log(1+t) - t} - \frac{2}{t^2} - \frac{2}{3t} + \frac{1}{9}\right) + \int\limits_{\varepsilon}^{3} \left(\frac{2}{t^2} + \frac{2}{3t} - \frac{1}{9}\right) dt \\ &\leqslant \int\limits_{0}^{3} \left(\frac{1}{(1+t)\log(1+t) - t} - \frac{2}{t^2} - \frac{2}{3t} + \frac{1}{9}\right) dt + 2/\varepsilon - 2/3 + \frac{2}{3}\log 3 + \frac{2}{3}\log\varepsilon^{-1} + \frac{1}{3}. \end{aligned} \tag{4.93}$$

Наконец, после этого каждый шаг не меньше $(1+\varepsilon)e^{1/(1+\varepsilon)-1}$, а отношение квадратов объемов составляет не больше $(1+\varepsilon)^r$, поэтому потребуется еще не более $r\frac{\log(1+\varepsilon)}{\log\left((1+\varepsilon)e^{1/(1+\varepsilon)-1}\right)} \leqslant \frac{2r}{\varepsilon} + \frac{5}{3}r$ шагов. Объединяя с первыми $2r$ столбцами и оценив интеграл в (4.93) численно, получаем, что итоговое число столбцов не превосходит

$$n \leqslant \left\lceil 3{,}45r + \frac{4r}{\varepsilon} + \frac{2}{3}r\log\varepsilon^{-1} \right\rceil.$$

□

Полученный результат превосходит предыдущую наилучшую оценку из [76] (см. также MVEE в таблице 4.1). При $\varepsilon = 1$ предложенный алгоритм потребует не больше $\lceil 7{,}45r \rceil$ столбцов по сравнению с предыдущей лучшей оценкой $4r\ln\ln r + 42r$.

Кроме того, для достижения ρ-локально максимального объема нам понадобилось всего $O\left(r^2\log_\rho(n/r)/n\right)$ замен: то есть, если при поиске без повторений не выгодно добавлять один из уже выбранных столбцов второй раз, то чем больше столбцов набрано, тем меньше замен потребуется для достижения ρ-локально максимального объема.

Глава 5. Эффективность поиска локально максимального объема в почти малоранговых матрицах

Рассмотрим теперь вопрос сходимости алгоритма maxvol (алгоритм 4.5) к подматрице ρ-локально максимального объема, если не совершать одновременных замен строк и столбцов. Тогда поиск оптимальной замены составляет $O(Nr)$ операций, что существенно сокращает стоимость алгоритма по сравнению с описанным в разделе 4.4.2. В этом случае можно лишь гарантировать, что найденная подматрица будет обладать локально максимальным объемом в своих строках и столбцах, но не во всей матрице. А значит нельзя, например, гарантировать оценки теоремы 4.5. В связи с этим возникает вопрос: в каких случаях локальная максимальность в текущих строках и столбцах будет гарантировать ρ-локальную максимальность (с небольшим ρ) во всей матрице?

Отметим, что сходимость поиска подматриц локально максимального объема крайне тяжело показать в общем случае. Данный вопрос впервые был рассмотрен в [86], где была доказана следующая теорема. В ней предполагается, что в приближении ранга 1 правый сингулярный вектор v является случайным и доказывается, что тогда при достаточно малой погрешности алгоритм maxvol (который для ранга 1 ищет максимальный по модулю элемент) находит элемент, близкий к максимуму во всей матрице.

Теорема 5.1 ([86])**.** *Пусть $A = Z + E$, $Z = u\sigma v^*$, $\sigma > 0$, $u^*u = v^*v = 1$, $A \in \mathbb{R}^{M\times N}$, где $v \in \mathbb{R}^N$ – случайный единичный вектор в $\mathbb{R}^N$. Обозначим*

$$\delta = \|E\|_C.$$

Пусть

$$\varepsilon = \frac{\|E\|_C}{\|Z\|_C} \leqslant \frac{1}{8}.$$

Пусть

$$\beta = \frac{4\varepsilon \|v\|_C \sqrt{N + 2\sqrt{cN} + 2c}}{\sqrt{2\pi}}.$$

Пусть алгоритм maxvol*, на первом шаге которого выбирается максимальный по модулю элемент среди среди k столбцов, возвратил после остановки или хотя бы 3-х шагов подматрицу $\hat{A} \in \mathbb{R}^{1\times 1}$, находящуюся на пересечении строки $R \in \mathbb{R}^{1\times N}$ и столбца $C \in \mathbb{R}^{M\times 1}$. Тогда с вероятностью $1 - e^{-c} - \beta^k$ справедлива оценка*

$$\left\|A - C\hat{A}^{-1}R\right\|_C \leqslant 4\,(1 + 16\varepsilon)\,\|E\|_C\,.$$

При этом найденный элемент будет отличаться от максимального (по модулю) не более, чем в $1 + 16\varepsilon$ раз (и не более, чем на 16δ).

Чтобы вероятность β стартовать с «плохого» столбца была меньше 1, необходимо $\|E\|_C \lesssim \frac{\sqrt{MN}}{\sqrt{\log N}}\sigma$, то есть, если погрешность распределена относительно равномерно и с учетом того, что для случайного v в среднем $\|v\|_C \sim \sqrt{\frac{\log N}{N}}$,

$$\|E\|_F \lesssim \frac{\|Z\|_F}{\sqrt{\log N}}.$$

Таким образом, хоть требование на величину погрешности и зависит от N, эта зависимость очень слабая, и отличия в несколько раз уже достаточно для сходимости maxvol всего за несколько шагов, без необходимости просмотра всей матрицы (то есть за $O\left(kM+N\right)$ операций).

Аналогичную оценку можно получить для аппроксимаций тензоров тензорным поездом ранга 1 (как показано в докладе автора на конференции CRCNAA 2017), но в ней требование на величину погрешности экспоненциально зависит от размерности тензора d (погрешность должна быть порядка $\log^{-\frac{d-1}{2}} N$).

Теорема 5.2. *Пусть задан тензор* $A\left(i_1,...,i_d\right)=\sigma u_1...u_d+E$, $\sigma>0$, $u_i\in\mathbb{R}^N$, $\|u_i\|_2=1$, $i=\overline{1,d}$. *Пусть все векторы* u_i *равномерно распределены на сфере в* $\mathbb{R}^N$. *Обозначим*

$$\delta=\|E\|_C,$$

$$\varepsilon=\frac{\|E\|_C}{\|A-E\|_C}=\frac{\|E\|_C}{\sigma\|u_1\|_\infty...\|u_d\|_\infty}\leqslant\frac{1}{d2^d},$$

$$\mu=2c_1\ln N,$$

c_1 *- произвольная константа. Далее* c *и* c_2 *- также произвольные константы. Обозначим*

$$\beta=\sqrt{1+2\sqrt{\frac{c_2}{N}}+2\frac{c_2}{N}}.$$

Пусть для некоторого β' *выполнены неравенства*

$$0<\frac{\beta\mu^{1/2}\sqrt{\frac{1}{2\pi}}}{\frac{1}{4\varepsilon}\mu^{-\frac{d-1}{2}}(d-1)^{-\frac{2c}{3N}}e^{-\frac{2c^2}{3N^2}-s\sqrt{\frac{2c(d-1)}{N}}-\frac{(\gamma+\ln 2)}{2}(d-1)}-1}=\beta'\leqslant\beta\mu^{1/2}\sqrt{\frac{1}{2\pi}},\tag{5.1}$$

$$s=\sqrt{\frac{\pi^2}{8}+\ln^2\beta+2\ln\beta+\frac{4}{N-2}+\frac{(\gamma+\ln 2)}{2N}},$$

γ *– постоянная Эйлера. Пусть мы применяем алгоритм из [108] (где* maxvol *применяется к* N *столбцам каждой развертки, причем индексы выбранных для каждой развертки столбцов не пересекаются). Пусть в каждой развертке алгоритм* maxvol *стартует с максимального (по модулю) элемента среди* k *случайных строк.*

Тогда с вероятностью

$$1 - d\sqrt{\frac{2}{\mu}}e^{-\frac{\mu}{2}(1-\frac{1}{N})} - de^{-c_2} - (d-1)\beta'^k - (d-1)e^{1-c}$$

алгоритм остановится на элементе, который отличается от максимального не более, чем на $2de\delta$.

Вернемся к матричному случаю. Пусть теперь ранг больше 1. В этом случае вместо поиска максимального по модулю элемента мы ищем подматрицу объема близкого к локально максимальному.

Пусть дано некоторое приближение Z ранга r. Рассмотрим произвольную подматрицу $r \times r$ в первых r строках матрицы $A = Z + E$ размера $M \times N$. Если соответствующие Z правые сингулярные векторы v_i, $i = \overline{1, r}$ распределены равномерно на сфере в $\mathbb{R}^N$ и перпендикулярны, то с точностью до произведения на некоторую квадратную матрицу их можно считать случайными гауссовыми векторами. В этом случае, так как нас везде интересует отношение объема к максимуму, то он целиком определяется этими r гауссовыми векторами.

Итак, нас интересует то, как распределен объем $r \times r$ гауссовой матрицы.

Утверждение 5.1. *Квадрат определителя гауссовой матрицы размера* $r \times r$ *обладает распределением*

$$\prod_{k=1}^{r} \chi^2(k).$$

Доказательство. Будем считать модуль определителя как объем параллелепипеда в $\mathbb{C}^r$. Квадрат длины первого вектора распределен как $\chi^2(r)$. Если мы знаем объем k-мерного параллелепипеда, построенного на первых k векторах, то объем $k+1$-мерного будет во столько раз больше, какую проекцию имеет $k+1$-ый вектор на перпендикулярное первым k векторам $r-k$-мерное подпространство. А в этом подпространстве его квадрат распределен как $\chi^2(r-k)$. Взяв произведение по всем k, получим требуемый ответ. □

Следствие 5.1. Пусть $V \in \mathbb{R}^{r\times r}$ – случайная гауссова матрица. Тогда

$$\mathbb{E}\,|\det V|^2 = r!$$

Утверждение 5.2. *Пусть* $V \in \mathbb{R}^{r\times r}$ *– случайная гауссова матрица. Тогда*

$$\mathcal{P}\left(|\det V| \leqslant \alpha\sqrt{r!}\right) < \alpha r^{3/4}.$$

Доказательство. Плотностью вероятности $\chi^2(k)$ распределения является $\frac{1}{2^{k/2}\Gamma(\frac{k}{2})}x_k^{k/2-1}e^{-x_k/2}$. Чтобы оценить искомую вероятность, проинтегрируем все распределения вероятностей до $+\infty$, а распределение для $k = 1$ до $\frac{\alpha^2 r!}{x_2\cdot\ldots\cdot x_r}$.

$$\int\limits_0^{\frac{\alpha^2 r!}{x_2 \dots x_r}} \frac{1}{\sqrt{2}\Gamma\left(\frac{1}{2}\right)} x^{-\frac{1}{2}} e^{-\frac{x}{2}} dx \leqslant \int\limits_0^{\frac{\alpha^2 r!}{x_2 \dots x_r}} \frac{1}{\sqrt{2}\Gamma\left(\frac{1}{2}\right)} x^{-\frac{1}{2}} dx = \frac{\alpha\sqrt{2}\sqrt{r!}}{\Gamma\left(\frac{1}{2}\right)\sqrt{x_2 \dots x_r}}.$$

Теперь в каждый интеграл можно подставить $\frac{1}{\sqrt{x_k}}$. Тогда $r-1$-кратный интеграл разбивается в произведение $r-1$ интегралов. Интеграл для $\chi^2(k)$ будет иметь вид

$$\int\limits_0^{+\infty} \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k-1}{2}-1} e^{-\frac{x}{2}} dx = \frac{\Gamma\left(\frac{k-1}{2}\right)}{\sqrt{2}\Gamma\left(\frac{k}{2}\right)} \int\limits_0^{+\infty} \frac{1}{2^{\frac{k-1}{2}}\Gamma\left(\frac{k-1}{2}\right)} x^{\frac{k-1}{2}-1} e^{-\frac{x}{2}} dx = \frac{\Gamma\left(\frac{k-1}{2}\right)}{\sqrt{2}\Gamma\left(\frac{k}{2}\right)}.$$

Взяв произведение по всем k и воспользовавшись оценками снизу и сверху для Гамма-функции, мы получим

$$\frac{\alpha\sqrt{r!}}{2^{\frac{r}{2}-1}\Gamma\left(\frac{r}{2}\right)} \leqslant \frac{\alpha\sqrt{\sqrt{2\pi r}\left(\frac{r}{e}\right)^n \left(1+\frac{1}{12}+\frac{1}{288}\right)}}{\frac{2^{\frac{r}{2}}}{r}\sqrt{\pi r}\left(\frac{r}{2e}\right)^{\frac{r}{2}}} < \alpha r^{\frac{3}{4}}.$$

□

Теперь получим оценку сверху для максимального объема в фиксированных строках.

Утверждение 5.3. *Пусть $V \in \mathbb{R}^{r\times N}$ – случайная гауссова матрица, а $\hat{V} \in \mathbb{R}^{r\times r}$ – произвольная её квадратная подматрица. Тогда*

$$\mathcal{P}\left(\max|\det V| \geqslant \left(e + 2e\sqrt{\frac{c}{r}} + \frac{2ce}{r}\right)^{r/2} \sqrt{r!}\right) \leqslant e^{\ln N - c}, \tag{5.2}$$

где c – произвольная константа.

Доказательство. Чтобы каждый определитель матрицы A не превосходил $\alpha^{r/2}\sqrt{r!}$, достаточно, чтобы квадрат нормы $l_i = \left\|V_{:,i}\right\|_2^2$ каждого её столбца i не превосходил $\alpha r/e$. Квадрат каждого столбца имеет распределение $\chi^2(r)$, для хвоста которого можно использовать оценку из [87] (см. также лемму 3.1):

$$\mathcal{P}\left(l_i > r + 2\sqrt{cr} + 2c\right) \leqslant e^{-c}.$$

Применив эту оценку для каждого из N столбцов, получим (5.2) с оценкой максимального объема $\left(e + 2e\sqrt{\frac{c}{r}} + \frac{2ce}{r}\right)^{r/2} \sqrt{r!}$. □

Замечание 5.1. Утверждениями 5.2 и 5.3 можно пользоваться для вычисления отношения объема случайной подматрицы к максимальному также и в случайных ортонормированных строках, поскольку правые сингулярные векторы случайной гауссовой матрицы задают случайную унитарной матрицу (с равномерным распределнием Хаара).

Теорема 5.3. *Пусть* $A = Z + E$, $Z = U\Sigma V^*$, $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$, $\sigma_i > 0$, $U^*U = V^*V = I$, $A \in \mathbb{R}^{M\times N}$. *Пусть* $V \in \mathbb{R}^{r\times N}$ *– случайные ортонормированные строки. Обозначим*

$$\delta = \|E\|_C.$$

Пусть

$$\varepsilon = \sum_{i=1}^{r} \frac{r^2\delta\sqrt{MN}}{\sigma_i\sqrt{r-i+1}}. \tag{5.3}$$

Пусть

$$\varepsilon' = \frac{9\cdot 2^r}{8}\left(e^{\varepsilon}-1\right) \leqslant \frac{1}{8},$$

$$\beta = 4r^{3/4}\varepsilon'\left(e + 2e\sqrt{\frac{c}{r}} + \frac{2ce}{r}\right)^{r/2}.$$

Пусть алгоритм maxvol *(со стартовой подматрицей на каждом шаге с объемом не меньше, чем* 2^{-r} *от максимального, что можно гарантировать, используя утверждение 4.16), на первом шаге которого выбирается матрица с максимальным модулем определителя среди подматриц локально максимального объема среди* k *блоков с* r *различными столбцами в каждом, возвратил после остановки подматрицу* $\hat{A} \in \mathbb{C}^{r\times r}$, *находящуюся на пересечении строк* $R \in \mathbb{C}^{r\times N}$ *и столбцов* $C \in \mathbb{C}^{M\times r}$. *Тогда с вероятностью* $1 - e^{\ln N - c} - \beta^k$ *справедлива оценка*

$$\left\|A - C\hat{A}^{-1}R\right\|_C \leqslant (1+16\varepsilon')\,(r+1)^2\,\|E\|_C\,. \tag{5.4}$$

Замечание 5.2. К сожалению, из-за необходимости поиска подматрицы локально максимального объема вместо максимального объема мы не можем гарантировать достижение высокой точности за небольшое число шагов, что можно было сделать в случае ранга 1 (теорема 5.1). Для того, чтобы число шагов было ограниченным, может потребоваться использовать одновременные замены (утверждение 4.12) или пересчитывать одновременно $C\hat{A}^{-1}$ и $\hat{A}^{-1}R$ после каждой замены.

Доказательство. Если C-норма погрешности равна δ, то норма Фробениуса погрешности для матрицы размера $r\times r$ не превосходит $r\delta$. Рассмотрим подматрицу $\hat{Z} = \hat{U}\hat{\Sigma}\hat{V}^*$ матрицы $U\Sigma V^*$ и соответствующую ей подматрицу $\hat{E}$ матрицы E. Тогда

$$\det\left(\hat{A}\right) = \det\left(\hat{Z}+\hat{E}\right) = \det\left(\hat{\Sigma}+\hat{U}^*\hat{E}\hat{V}\right).$$

Оценим, насколько $\hat{E}$ может изменить $\det\hat{Z}$.

Обозначим через α_i 2-норму i-й строки $\hat{U}^*\hat{E}\hat{V}$. Так как унитарные матрицы не меняют норму Фробениуса, а $\|\hat{E}\|_F \leqslant r\delta$, то

$$\sum_{i=1}^{r} {\alpha_i}^2 \leqslant r^2\delta^2. \tag{5.5}$$

Погрешность определителя будем оценивать через произведение 2-норм строк. Из неравенства треугольника получаем, что 2-норма i-й строки не превосходит $\hat{\sigma}_i + \alpha_i$, где $\hat{\sigma}_i$ – i-е сингулярное число $\hat{Z}$. Поэтому определитель возрастет не более, чем на

$$\prod_{i=1}^{r} (\hat{\sigma}_i + \alpha_i) - \prod_{i=1}^{r} \hat{\sigma}_i. \tag{5.6}$$

Обозначим оценку (относительно подматриц $\tilde{A}$, отличающихся от текущей в одной строке или столбце) погрешности определителя через ε'. Сразу учтем, что можно гарантировать, что объем $\tilde{A}$ не меньше 2^{-r} от максимального, если стартовать как в утверждении 4.16. При этом, как мы потребуем далее, ε' меняет объем матрицы максимального объема не более чем на $1/8$, а потому, раскрыв скобки в (5.6), получаем

$$\varepsilon' = \frac{1}{\max\limits_{\hat{A}} \prod\limits_{i=1}^{r} \tilde{\sigma}_i} \sum_{k=1}^{r} \sum_{i} \hat{\sigma}_{i_1} \ldots \hat{\sigma}_{i_{r-k}} \alpha_{j_1} \ldots \alpha_{j_k} \leqslant \frac{9}{8} \cdot \frac{2^r}{\max\limits_{\hat{Z}} \prod\limits_{i=1}^{r} \hat{\sigma}_i} \sum_{k=1}^{r} \sum_{i} \hat{\sigma}_{i_1} \ldots \hat{\sigma}_{i_{r-k}} \alpha_{j_1} \ldots \alpha_{j_k}, \tag{5.7}$$

где i во внутренней сумме пробегает все возможные возрастающие подпоследовательности длины $r - k$. Через j здесь обозначена подпоследовательность оставшихся k индексов.

Такое рассмотрение позволяет также оценить изменение определителя снизу: в этом случае сумма в (5.7) будет содержать коэффициенты вида $(-1)^k$, а потому будет не больше полученной оценки. То есть формула (5.7) дает не только верхнюю, но и нижнюю оценку на изменение определителя, то есть представляет из себя оценку на максимальное относительное изменение его модуля (то есть максимальное относительное изменение объема).

Из выражения (5.6) ясно, что если $\hat{\sigma}_i \geqslant \hat{\sigma}_j$, то максимум достигается при $\alpha_i \leqslant \alpha_j$. Иначе можно переставить строки i и j местами, что изменение определителя не уменьшит. Это позволяет получить оценку для α_k. Хотя все $\alpha_i, i < k$ в худшем случае равны 0, все $\alpha_i, i > k$ будут не меньше α_k. Подставляя $\alpha_i = \alpha_k, i > k$ в (5.5), получаем, что

$$\alpha_k \leqslant \frac{r}{\sqrt{r - k + 1}} \delta. \tag{5.8}$$

Домножим и разделим (5.7) на $\sqrt{r!}$ и применим оценку (5.8). В этом случае входящие в $\sqrt{r!}$ множители $\sqrt{r - j_1 + 1}, \ldots, \sqrt{r - j_k + 1}$ сократятся, и мы получим

$$\varepsilon' \leqslant \frac{9 \cdot 2^r}{8\sqrt{r!} \max\limits_{\hat{Z}} \prod\limits_{i=1}^{r} \hat{\sigma}_i} \sum_{k=1}^{r} \sum_{i} \hat{\sigma}_{i_1} \sqrt{r - i_1 + 1} \ldots \hat{\sigma}_{i_{r-k}} \sqrt{r - i_{r-k} + 1} (r\delta)^k. \tag{5.9}$$

Оценим произведение сингулярных чисел во внутренней сумме. Для этого сначала обозначим через $\hat{U}$ и $\hat{V}$ подматрицы из U и V, соответствующие $\hat{A}$. Оценив произведения сингулярных чисел

$\hat{Z}$ через максимально возможные, используя максимум по $\hat{U}$ и $\hat{V}$, получим

$$\begin{aligned}
&\sum_i \hat{\sigma}_{i_1}\sqrt{r-i_1+1}\ldots\hat{\sigma}_{i_{r-k}}\sqrt{r-i_{r-k}+1} \leqslant \\
&\leqslant \sum_i \sigma_{i_1}\sqrt{r-i_1+1}\ldots\sigma_{i_{r-k}}\sqrt{r-i_{r-k}+1}\max_{\hat{U}}\prod_{i=1}^{r-k}\sigma_i\left(\hat{U}\right)\max_{\hat{V}}\prod_{i=1}^{r-k}\sigma_i\left(\hat{V}\right) \leqslant \\
&\leqslant \frac{1}{k!}\prod_{i=1}^{r}\sigma_i\sqrt{r-i+1}\left(\sum_{i=1}^{r}\frac{1}{\sigma_i\sqrt{r-i+1}}\right)^k\max_{\hat{U}}\prod_{i=1}^{r-k}\sigma_i\left(\hat{U}\right)\max_{\hat{V}}\prod_{i=1}^{r-k}\sigma_i\left(\hat{V}\right).
\end{aligned} \tag{5.10}$$

В любой матрице $\hat{U}\in\mathbb{C}^{r\times r}$ можно найти подматрицу из $r-k$ строк с максимальным объемом. Согласно лемме 1.1, он будет отличаться от произведения $r-k$ сингулярных чисел исходной матрицы не более, чем в $\sqrt{C_r^{r-k}}=\sqrt{C_r^k}$ раз (так как в лемме 1.1 будет ровно C_r^{r-k} слагаемых, каждое из которых по объему не превосходит максимального). При этом $r-i+1$-ю строку всегда можно добавить так, что объем увеличится минимум в $\sqrt{\frac{i}{M}}$ раз. Для этого достаточно умножением справа на унитарную матрицу занулить последние i столбцов в уже выбранной подматрице. Тогда, так как сумма квадратов элементов в i последних столбцах всей матрицы не изменится, найдется строка, сумма квадратов элементов которой в этих столбцах не меньше $\frac{i}{M}$.

Добавляя таким образом строки к уже фиксированным $r-k$ строкам, мы получим, что

$$\max_{\hat{U}}\prod_{i=1}^{r-k}\sigma_i\left(\hat{U}\right)\leqslant\sqrt{C_r^kM^k}\max_{\hat{U}}\prod_{i=1}^{r}\sigma_i\left(\hat{U}\right). \tag{5.11}$$

Относительную погрешность ε' мы находим подстановкой (5.11) и аналогичной оценки для $\hat{V}$

$$\max_{\hat{V}}\prod_{i=1}^{r-k}\sigma_i\left(\hat{V}\right)\leqslant\sqrt{C_r^kN^k}\max_{\hat{V}}\prod_{i=1}^{r}\sigma_i\left(\hat{V}\right)$$

в (5.10), а затем (5.10) в (5.9). С учетом $\max\limits_{\hat{U}}\prod\limits_{i=1}^{r}\sigma_i\left(\hat{U}\right)\max\limits_{\hat{V}}\prod\limits_{i=1}^{r}\sigma_i\left(\hat{V}\right)\prod\limits_{i=1}^{r}\sigma_i=\max\limits_{\hat{Z}}\prod\limits_{i=1}^{r}\hat{\sigma}_i$, получаем

$$\varepsilon'=\frac{9\cdot2^r}{8}\sum_{k=1}^{r}\frac{1}{k!}\left(\sum_{i=1}^{r}\frac{n\delta}{\sigma_i\sqrt{r-i+1}}\right)^kC_r^k\left(\sqrt{MN}\right)^k\leqslant\frac{9\cdot2^r}{8}\sum_{k=1}^{\infty}\frac{1}{k!}\left(\sum_{i=1}^{r}\frac{r^2\delta\sqrt{MN}}{\sigma_i\sqrt{r-i+1}}\right)^k.$$

Используя введенное в условии определение ε (5.3),

$$\varepsilon'\leqslant\frac{9\cdot2^r}{8}\left(e^{\varepsilon}-1\right).$$

Перейдем теперь к оценке того, насколько сильным может быть отличие от локально максимального объема в Z, если мы нашли подматрицу локально максимального объема в A.

Введем μ как отношение текущего объема без погрешности (соответствующего подматрице локально максимального объема) к максимальному объему (без погрешности) среди подматриц,

отличающихся от данной одной строкой или столбцом. Тогда для роста объема в Z достаточно выполнение неравенства

$$\begin{aligned} \mu^2 + \varepsilon' &\leqslant \mu - \varepsilon', \\ \mu^2 - \mu + 2\varepsilon' &\leqslant 0, \end{aligned} \tag{5.12}$$

где слева μ^2 – оценка отношения текущего объема без погрешности к максимальному объему без погрешности среди подматриц, отличающихся от данной одной строкой и/или столбцом, а μ – максимально достижимое отношение при замене одной строки или столбца. Если относительная погрешность ε' не понижает максимально достижимое значение (μ) ниже текущего (μ^2), то рост объема все еще возможен.

Решая неравенство (5.12), получаем, что рост происходит в промежутке

$$\mu \in \left[\frac{1 - \sqrt{1 - 8\varepsilon'}}{2},\ \frac{1 + \sqrt{1 - 8\varepsilon'}}{2} \right], \tag{5.13}$$

откуда получаем условие $\varepsilon' \leqslant 1/8$.

Следовательно, начиная с $\mu_0 \geqslant 4\varepsilon' \geqslant \frac{1-\sqrt{1-8\varepsilon'}}{2}$, мы гарантируем, используя maxvol, рост вплоть до $\mu = \frac{1+\sqrt{1-8\varepsilon'}}{2}$. В результате мы получим матрицу $\frac{1+\varepsilon'}{\mu^2+\varepsilon'}$-локально максимального объема. Обозначим этот коэффициент через c и оценим его, подставив правую границу (5.13):

$$c = \frac{1 + \varepsilon'}{\mu^2 + \varepsilon'} \leqslant 1 + 16\varepsilon'.$$

Согласно замечанию 2.1, из теоремы 2.1 получаем оценку погрешности

$$\left\| A - C\hat{A}^{-1}R \right\|_C \leqslant (1 + 16\varepsilon') \, (r + 1)^2 \, \|E\|_C \, .$$

Обозначим через $\beta = \alpha r^{3/4} \cdot (ce + e)^{r/2}$ условную вероятность выбрать r столбцов с объемом не меньше α от максимального, согласно утверждениям 5.2 и 5.3. Нам необходимо начать с $\alpha = \mu_0 \geqslant 4\varepsilon'$, а потому вероятность оценивается как

$$\beta = 4r^{3/4}\varepsilon' \, (ce + e)^{r/2} \, .$$

Применяя алгоритм maxvol k раз для каждого такого выбора, получаем вероятность ни разу не стартовать со столбцов с объемом не меньшим $4\varepsilon'$ не больше β^k. Учитывая дополнительно вероятность того, что максимальный объем не слишком сильно отличается от среднего (из утверждения 5.3) получаем итоговую оценку вероятности достижения (5.4).

Наконец, заметим, что не обязательно использовать maxvol k раз до достижения подматрицы локально максимального объема в строках и столбцах, а достаточно выбрать подматрицу максимального объема среди найденных подматриц локально максимального объема в k блоках из r столбцов. Действительно, если среди них есть блок, соответствующий $\mu_0 \geqslant 4\varepsilon'$, то в нем мы гарантированно найдем подматрицу с объемом не меньше $\mu_1 \geqslant \mu_0$ относительно локально

максимального (согласно границам (5.13)). При этом для любой пары подматриц большего объема (отличающихся друг от друга в одной строке) оценка влияния погрешности будет меньше, а значит оценка для достигаемого μ_1 будет выше. Таким образом, при выборе максимального объема среди найденных подматриц в k блоках-столбцах, мы найдем подматрицу в r строках, соответствующих $\mu_1 \geqslant 4\varepsilon'$, что гарантирует дальнейший рост μ как минимум до верхней границы (5.13), пока алгоритм не остановится. □

Замечание 5.3. Чтобы итоговая вероятность успеха была близка к 1, необходимо использовать $c = \max\left(\text{const}, \frac{\ln N}{r}\right)$. Поскольку необходимо $\beta < 1$, с учетом $\|E\|_F \sim \|E\|_C \sqrt{MN}$ для случайного шума и используя (5.3), получаем условие на величину шума

$$\|E\|_F \lesssim \frac{\sigma_r}{\left(\max\left(\text{const}, \sqrt{\frac{\log N}{r}}\right)\right)^r}.$$

Для малых рангов эта оценка логарифмически зависит от числа столбцов (для $r = 1$ асимптотика та же, что и в теореме 5.1), а потому, хоть и является грубой, говорит о том, что не слишком высокий шум обычно не способен повлиять на поиск подматрицы локально максимального объема, даже если не делать одновременные замены строк и столбцов (как в разделе 4.4.2), а использовать алгоритм maxvol в строках и столбцах поочередно, не рассматривая всю матрицу целиком.

5.1. Выбор ранга аппроксимации

На практике часто требуемый ранг аппроксимации заранее неизвестен. Вместо этого накладываются некоторые требования по точности вида $\|A - CGR\| \leqslant \varepsilon$ в некоторой норме. В целом, достичь точности ε можно, например, удваивая ранг, пока не будет достигнута нужная точность. После получения аппроксимации большего ранга, его всегда легко уменьшить, поскольку сокращенное сингулярное разложение крестовой аппроксимации выполняется за $O\left(Mnr + Nmr + mn \min\left(m, n\right)\right)$ операций, см. конец раздела 1.1.

Однако, стартовый (возможно, завышенный) ранг можно также набрать с помощью простых адаптивных методов крестовой аппроксимации, основынных на неполном LU разложении с выбором ведущего элемента в текущих строках и/или столбцах. В основном для этого используется упомянутый ранее алгоритм Cross 2D [4], впервые примененный еще до публикации в контексте мозаично-скелетонных аппроксимаций [1]. Его единственное отличие от неполного разложения с выбором ведущего элемента по столбцам (строкам) состоит в том, что алгоритм делает дополнительный шаг по строкам (столбцам) для выбора нового ведущего столбца (строки) на следующий шаг. Данный подход эвристический и не всегда работает на практике, поскольку выбранный на каждом шаге элемент почти никогда не является максимальным в своей строке

(столбце), а потому даже по отдельности каждый шаг не гарантирует высокой точности аппроксимации ранга 1.

Как следует из теоремы 5.1, для достижения близкого к максимальному элемента обычно требуется три шага. Кроме того, из нее также следует, что крайне полезным оказывается использование части столбцов в качестве «тестовых» (стартовых): выбирать на первом шаге максимум не из одного столбца, а из нескольких, что существенно увеличивает вероятность быстрого достижения максимума в дальнейшем. В связи с этим предлагается использовать описанную в алгоритме 5.1 модификацию, где число тестовых столбцов t (которое также можно увеличивать с ростом ранга), максимальное число шагов $s_{\max}$ и константа ρ, определяющая близость к максимуму, выбираются пользователем.

Во время выполнения алгоритма сохраняются номера уже выбранных строк и столбцов, так что в них $(A - UV)_{ij} = 0$, что позволяет избежать повторного выбора из-за ошибок округления. Основная асимптотика в сложности возникает в обновлениях ранга 1 вида $A - UV$. Всего в алгоритме требуется не больше s вычислений обновлений строк и столбцов на каждый ранг, а потому полное число умножений в случае $M = N$ будет не больше $Nsr^2/2 + O(Nrt)$.

Если число тестовых столбцов постоянно, а число шагов ограничено двумя, алгоритм обладает той же константой в асимптотике, что и Cross 2D, однако приводит к существенно меньшей дисперсии погрешности. Стоит отметить, что при фиксированном ранге данный подход все же уступает в точности алгоритмам maxvol, 4.3, и, тем более, maxvol-proj, 4.11, поскольку не позволяет менять уже набранные строки и столбцы. Таким образом, ранг полученной аппроксимации может быть завышен. Однако, как было уже упомянуто, его легко сократить. А если сингулярные числа матрицы быстро убывают, и перед этим применить один из алгоритмов поиска максимального объема или максимального проективного объема, то можно гарантировать высокую точность после такого сокращения, как следует из формул в конце раздела 1.1.

Эффективность аппроксимации данного алгоритма по достижении фиксированного ранга можно увидеть в таблице 5.1. Видно, что ключевую роль играет именно выбор нескольких тестовых столбцов, что позволяет улучшить точность метода, не приводя к существенному росту вычислительной сложности. Поскольку разница с более эффективными методами для фиксированного ранга обычно не превосходит двух раз, при фиксированной точности и данном распределении сингулярных чисел это приведет к увеличению ранга максимум на 1, что не важно в тех применениях, где нет необходимости строить аппроксимацию, близкую к сингулярному разложению. Код классического адаптивного крестового метода (ACA), использованный, в частности, в [18], предоставлен Станиславом Леонидовичем Ставцевым.

Алгоритм 5.1 Адаптивный крестовый метод

Вход: Матрица $R \in \mathbb{C}^{M \times N}$, требуемая точность ε ($\|A - CGR\|_C < \varepsilon$), перестановка столбцов $p \in \mathbb{N}^N$ для выбора тестовых столбцов, число тестовых столбцов t, максимальное число шагов на ранг $s_{\max}$, коэффициент отношения найденных элементов к максимальным $\rho \geqslant 1$.

Выход: Ранг r, факторы $U \in \mathbb{C}^{M \times r}$ и $V \in \mathbb{C}^{N \times r}$, задающие скелетное разложение $C\hat{A}^{-1}R = UV^T$.

$r := 0$

repeat

Если последний выбранный столбец находится среди t тестовых, он удаляется из множества тестовых, и к ним добавляется новый столбец согласно перестановке p, так что число тестовых столбцов всегда равно t.

Элементы каждого тестового столбца с индексом k обновляются путем вычитания из него $U_{:,r}V_{r,k}$.

Пусть $\left(A - UV^T\right)_{ij}$ – максимальный по модулю элемент среди тестовых столбцов.

$s := 1$

repeat

$j := \arg\max_j \left|(A - UV)_{ij}\right|$

$e_{\max} := \max_i \left|(A - UV)_{ij}\right|$

$s := s + 1$

if $s \geqslant s_{\max}$ **or** $e_{\max} \leqslant \rho \left|(A - UV)_{ij}\right|$ **then**

break

end if

$i := \arg\max_i \left|(A - UV)_{ij}\right|$

$e_{\max} := \max_j \left|(A - UV)_{ij}\right|$

$s := s + 1$

until $s \geqslant s_{\max}$ **or** $e_{\max} \leqslant \rho \left|(A - UV)_{ij}\right|$

$r := r + 1$

$U_{:,r} := (A - UV)_{:,j} \,/\sqrt{\left|(A - UV)_{ij}\right|}$

$V_{:,r} := (A - UV)^T_{i,:} \sqrt{\left|(A - UV)_{ij}\right|}/(A - UV)_{ij}$

until $e_{\max} < \varepsilon$

$r := r - 1$

Из U и V удаляется последний столбец.

Таблица 5.1: Средние погрешности и время вычисления скелетных аппроксимаций ранга $r = 20$ для 100 случайных квадратных randsvd матриц размера $N = 1000$ с сингулярными числами $\sigma_k = 1/2^k$. $LU = C\hat{A}^{-1}R$ – сокращенное LU разложение.

Метод	$\|A - LU\|_F$	$\|A - LU\|_2$	$\|A - LU\|_C$	Время, сек
ACA [18]	$3{,}37 \cdot 10^{-6}$	$3{,}13 \cdot 10^{-6}$	$3{,}88 \cdot 10^{-8}$	$1{,}14 \cdot 10^{-3}$
Алгоритм 5.1, 1 тестовый, 2 шага	$1{,}90 \cdot 10^{-6}$	$1{,}72 \cdot 10^{-6}$	$2{,}06 \cdot 10^{-8}$	$8{,}42 \cdot 10^{-4}$
Алгоритм 5.1, k тестовых, 2 шага	$1{,}52 \cdot 10^{-6}$	$1{,}35 \cdot 10^{-6}$	$1{,}64 \cdot 10^{-8}$	$1{,}24 \cdot 10^{-3}$
Алгоритм 5.1, 1 тестовый, до достижения $\rho = 1{,}1$	$1{,}59 \cdot 10^{-6}$	$1{,}42 \cdot 10^{-6}$	$1{,}76 \cdot 10^{-8}$	$1{,}27 \cdot 10^{-3}$
maxvol [11], r замен, 2 шага	$1{,}68 \cdot 10^{-6}$	$1{,}52 \cdot 10^{-6}$	$1{,}87 \cdot 10^{-8}$	$3{,}28 \cdot 10^{-3}$
maxvol [11], до остановки	$1{,}36 \cdot 10^{-6}$	$1{,}21 \cdot 10^{-6}$	$1{,}45 \cdot 10^{-8}$	$6{,}57 \cdot 10^{-3}$
Полный выбор ведущего элемента	$1{,}46 \cdot 10^{-6}$	$1{,}27 \cdot 10^{-6}$	$1{,}52 \cdot 10^{-8}$	$4{,}61 \cdot 10^{-2}$
Теорема 4.2	$1{,}23 \cdot 10^{-6}$	$1{,}07 \cdot 10^{-6}$	$1{,}31 \cdot 10^{-8}$	$6{,}78 \cdot 10^{-2}$
Алгоритм 4.8	$1{,}26 \cdot 10^{-6}$	$1{,}10 \cdot 10^{-6}$	$1{,}35 \cdot 10^{-8}$	$7{,}83 \cdot 10^{-2}$

Глава 6. Численные эксперименты

Алгоритмы maxvol и dominant позволяют находить подматрицы локально максимального объема в предписанных строках и/или столбцах, а потому формально позволяют достичь результатов из разделов 2.2.1 с $t(r, n, N)$ ограниченным следствием 1.6 и из главы 3, если матрица Z известна.

На практике, однако, наилучшее приближение $Z = A_r$ является неизвестным. Задача состоит именно в поиске приближения, близкого к наилучшему. Для этого поиск ведется в самой матрице A, а вместо проектора P в теореме 3.4 используется сокращенное сингулярное разложение подматрицы $\hat{A}$, что приводит к аппроксимации вида $C\left(\hat{A}_r\right)^{+}R$. В связи с тем, что доказанные в главе 3 результаты тогда уже не гарантируют оценки ошибки $\|A - C\hat{A}_r^{+}R\|_F$, интересно увидеть, насколько погрешность на практике близка к той, что указана в теоремах. А именно, выполняется ли неравенство

$$\mathbb{E}\|A - C\hat{A}_r^{+}R\|_F^2 \leqslant \frac{m+1}{m-r+1} \cdot \frac{n+1}{n-r+1}\|A - A_r\|_F^2. \tag{6.1}$$

Для более точного сравнения уточним смысл доказанных результатов. А именно, вместо того, чтобы явно оценивать $\|\hat{V}\|_F^2$ сверху, заменим оценку $\|\hat{V}\|_F^2 \leqslant r\frac{N-r+1}{n-r+1}$ напрямую на значение $\mathbb{E}_V\|\hat{V}\|_F^2$, где $\hat{V}$ ищется как подматрица с локально максимальным объемом. В этом случае

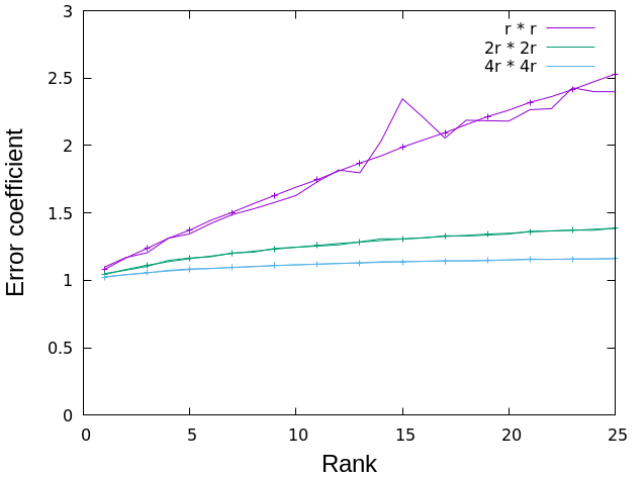


Рис. 6.1: Значения $\|A - C\hat{A}_r^+ R\|_F / \|A - A_r\|_F$ (линии без крестов) для случайных $A \in \mathbb{R}^{1000\times 1000}$ с сингулярными числами $\sigma_1 = \ldots = \sigma_r = 100\sigma_{r+1} = \ldots = 100\sigma_{1000}$ соответствующего ранга. Значения погрешности аппроксимации получены с помощью алгоритма maxvol-proj (алгоритм 4.11). Линии с крестами показывают ожидаемое значение коэффициента аппроксимации для каждого ранга и размера. Данный коэффициент равен $1+\frac{1}{N-r}\left(\mathbb{E}_V \left\|\hat{V}^+\right\|_F^2 - r\right)$. Различные линии показывают результаты для различных размеров подматрицы $\hat{A}$: $m = n = r$, $m = n = 2r$ и $m = n = 4r$.

коэффициент в 6.1 будет выглядеть следующим образом:

$$\mathbb{E}\|A - C\hat{A}_r^+ R\|_F^2 \approx \left(1 + \frac{\mathop{\mathbb{E}}_{\substack{U\in\mathbb{R}^{r\times M},\\ \hat{U}\in\mathbb{R}^{r\times m}}} \left\|\hat{U}^+\right\|_F^2 - r}{M - r}\right)\left(1 + \frac{\mathop{\mathbb{E}}_{\substack{V\in\mathbb{R}^{r\times N},\\ \hat{V}\in\mathbb{R}^{r\times n}}} \left\|\hat{V}^+\right\|_F^2 - r}{N - r}\right)\|A - A_r\|_F^2. \tag{6.2}$$

Выражение (6.2) является нашей гипотезой. Матожидания вида $\mathbb{E}_V\|\hat{V}\|_F^2$ можно оценить путем отдельной (независимой от A) генерации матриц U и V, поиска в них подматриц локально максимального объема и последующего усреднения.

На рисунке 6.1 показаны численные значения величины $\|A - C\hat{A}_r^+ R\|_F^2$ (где матрица A выбирается из RANDSVD ансамбля) в сравнении с матожиданием правой части (6.2). Видно, что величина погрешность полностью определяется $\left\|\hat{U}^+\right\|_F$ подматриц соответствующего размера. Таким образом, основное преимущество подматриц локально максимального объема состоит именно в том, что они позволяют ограничить $\left\|\hat{U}^+\right\|_F$. Кроме того, в отличие от других алгоритмов минимизации нормы Фробениуса псевдообратной матрицы, подматрицу локально максимального объема или большого проективного объема можно найти, постоянно увеличивая объем или проективный объем в строках и столбцах. Так как такое увеличение не может продолжаться бесконечно, алгоритм гарантированно остановится. С другой стороны, если напрямую минимизировать норму Фробениуса по очереди в строках и столбцах, нет очевидного способа гарантировать, что не возникнет цикла.

Можно также оценить эффективность поиска локально максимального объема с помощью алгоритмов maxvol и dominant в конкретных строках R.

В качестве теста рассмотрим выбор R из RANDSVD ансамбля (определение 3.1): её сингулярные числа фиксированы, а сингулярные векторы формируют случайные унитарные матрицы с мерой Хаара. Заметим, что выполнение алгоритмов maxvol и dominant полностью определяется правыми сингулярными векторами $V \in \mathbb{R}^{r\times N}$ сокращенного сингулярного разложения $R = U\Sigma V$,

поскольку

$$C = \hat{R}^+ R = \left(U\Sigma\hat{V}\right)^+ U\Sigma V = \left(\hat{V}\right)^+ V.$$

Таблица 6.1: Число шагов (замен столбцов) при использовании maxvol и dominant для $\rho = 1$. Первые r столбцов стартовой матрицы выбираются с помощью выбора ведущих столбцов (алгоритм 4.1). Показано среднее и максимальное число шагов среди 100 генераций $V \in \mathbb{R}^{r\times N}$, $VV^T = I$. $r = 50$, $N = 5000$.

Алгоритм	maxvol ($n = r = 50$)	dominant ($n = 100$)	dominant ($n = 500$)
Шагов, в среднем	1,2	81	437
Шагов, максимум	7	99	457

Согласно таблице 6.1, обычно нужно совсем немного замен для $n = r$ и не больше n замен для $n > r$. Тем не менее, это не доказывает, что n шагов всегда достаточно. Поэтому далее мы ограничимся $2n$ заменами, что все еще позволяет сохранить общую сложность $O\left(Nn^2\right)$.

Далее мы проверим полученные оценки на норму псевдообращения подматриц для нескольких специальных случаев распределения сингулярных чисел. Согласно оценкам следствия 4.1, худшим случаем является $\frac{\|R^+\|_2^2}{\|R^+\|_F^2} \approx 1$, т. е., когда только одно сингулярное число очень мало; а лучшим случаем является $\frac{\|R^+\|_2^2}{\|R^+\|_F^2} = 1/r$, т. е., когда все сингулярные числа совпадают. Поэтому именно эти случаи для нас наиболее интересны.

Согласно утверждению 4.1, нижней границей является $\frac{\left\|\hat{R}^+\right\|_{2,F}}{\|R^+\|_{2,F}} \geqslant \frac{N-r+1}{n-r+1}$ как для спектральной нормы, так и для нормы Фробениуса. Разумно спросить, достижима ли эта оценка с помощью локально максимального объема на практике, ведь оценка для худшего случая примерно в r раз больше. Чтобы ответить на этот вопрос, рассмотрим эффективность алгоритма dominant в среднем и в худшем случае. В таблице 6.2 мы рассматриваем два варианта числа столбцов: $n = r$ и $n = 2r - 1$.

При тестировании для случая равных сингулярных чисел достаточно задать $R = V$, где $V \in \mathbb{R}^{r\times N}$ – случайная (с мерой Хаара) матрица с ортонормированными строками. Чтобы получить любое другое распределение сингулярных чисел, достаточно выбрать $R = \Sigma V$, где Σ – диагональная матрица сингулярных чисел. Заметим, что умножать на левые сингулярные числа не требуется, поскольку они не повлияют ни на алгоритм (объемы подматриц от них не зависят), ни на нормы (спектральную и фробениусову) подматриц и самой матрицы R.

Итак, мы рассматриваем 2 случая:

$$\begin{aligned} &\text{Случай 1}: \quad \Sigma = I, \\ &\text{Случай 2}: \quad \Sigma = \operatorname{diag}\left(1, \ldots, 1, 10^{-10}\right). \end{aligned}$$

Для тестов фиксируем $r = 100$ и $N = 10099$. При таком выборе для $n = r$ следует ожидать отношение $\frac{\|\hat{R}^+\|_{2,F}}{\|R^+\|_{2,F}} \sim \sqrt{\frac{N-r+1}{n-r+1}} = 100$, а для $n = 2r-1$ отношение $\frac{\|\hat{R}^+\|_{2,F}}{\|R^+\|_{2,F}} \sim \sqrt{\frac{N-r+1}{n-r+1}} = 10$. С другой стороны, верхние границы таблицы 6.2 предсказывают в худшем случае отношения примерно в 10 раз выше для спектральной нормы (в обоих случаях) и для нормы Фробениуса в случае 2. При этом для случая 2 отношения спектральной нормы и нормы Фробениуса почти совпадают, поскольку R^+ почти ранга 1 (с небольшой относительной погрешностью).

Таблица 6.2: Результаты работы maxvol и dominant с точки зрения отношений норм псевдообратных матриц. Среднее значение и максимум получены исходя из 1000 генераций случайной матрицы V. $r = 100$, $N = 10099$.

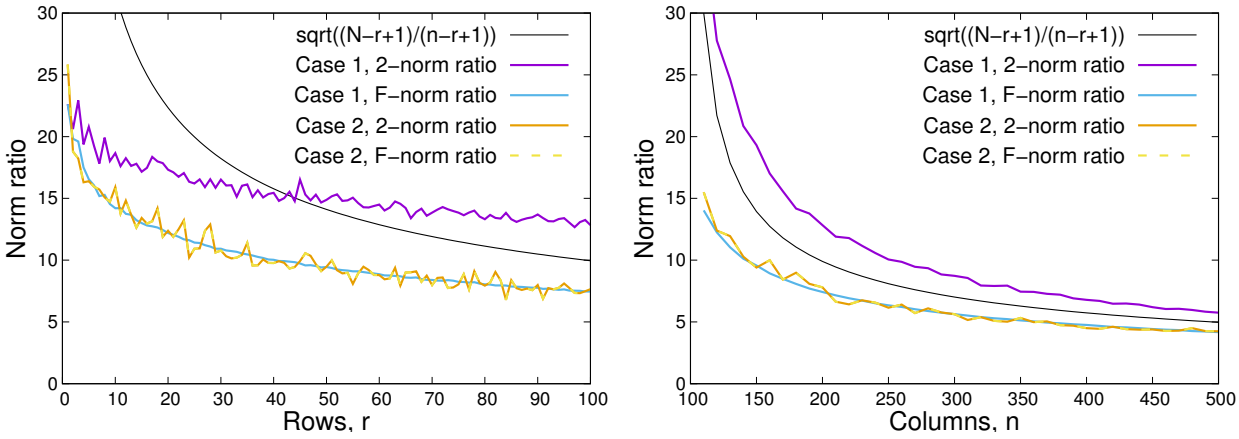
Columns	$n = r$	$n = 2r - 1$
$\|\hat{R}^+\|_F / \|R^+\|_F$, Случай 1, среднее	18,4	7,48
$\|\hat{R}^+\|_F / \|R^+\|_F$, Случай 1, максимум	19,1	7,52
$\|\hat{R}^+\|_F / \|R^+\|_F$, Случай 2, среднее	25,7	7,56
$\|\hat{R}^+\|_F / \|R^+\|_F$, Случай 2, максимум	38,3	8,39
$\|\hat{R}^+\|_2 / \|R^+\|_2$, Случай 1, среднее	64,9	13,0
$\|\hat{R}^+\|_2 / \|R^+\|_2$, Случай 1, максимум	77,7	13,6
$\|\hat{R}^+\|_2 / \|R^+\|_2$, Случай 2, среднее	25,7	7,56
$\|\hat{R}^+\|_2 / \|R^+\|_2$, Случай 2, максимум	38,3	8,39
$\sqrt{\frac{N-r+1}{n-r+1}}$	100	10

Как видно, верхние границы из таблицы 6.2, где отношение порядка $\sqrt{Nr} \approx 1000$, нам не встретились: все значения были того же порядка, что и нижняя граница $\sqrt{\frac{N-n+1}{n-r+1}}$. Случай 1 для спектральной нормы при $n = 2r - 1$ оказался наихудшим: для него отношение $\|\hat{R}^+\|_2 / \|R^+\|_2 \approx 13{,}6 > 10$. Однако, данный результат все еще почти в 2 раза меньше наилучшей верхней оценки $\frac{\sqrt{N+2-2\sqrt{\frac{r}{n+1}}}}{\sqrt{n+1}-\sqrt{r}} \approx 24{,}3$ (см. таблицу 4.1).

В качестве еще одной иллюстрации, рисунок 6.2a показывает как на спектральную норму и норму Фробениуса влияет ранг r, а рисунок 6.2b показывает как они меняются с изменением числа столбцов n.

Перейдем теперь напрямую к рассмотрению погрешностей крестовой аппроксимации. Будем генерировать квадратные матрицы, т. е. $M = N$. Для каждой матрицы $A \in \mathbb{R}^{N \times N}$ выбираются $n \geqslant r$ столбцов $C \in \mathbb{R}^{N \times n}$, $m \geqslant r$ строк $R \in \mathbb{R}^{m \times N}$ и строится аппроксимация вида CGR с генератором $G \in \mathbb{R}^{n \times m}$ ранга не больше r.

Одним из основных преимуществ использования подматриц большого проективного объема



(a) $n = 2r - 1$, r от 1 до 100.

(b) $r = 100$, n от 100 до 500.

Рис. 6.2: Результаты выполнения dominant с точки зрения отношения норм псевдообратных матриц для $R = \Sigma V$ со случайной $V \in \mathbb{R}^{r \times N}$, $N = 10000$.

является возможность достижения относительной точности $1 + \varepsilon$ по норме Фробениуса

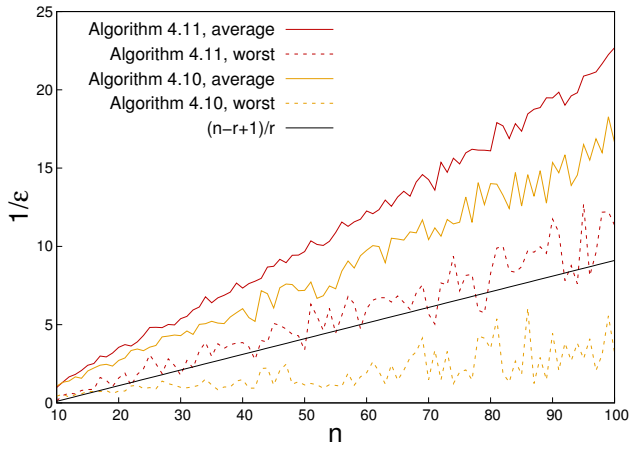
$$\|A - CGR\|_F = (1 + \varepsilon)\|A - A_r\|_F, \tag{6.3}$$

где в среднем $1 + \varepsilon \lesssim \frac{m+1}{m-r+1} \cdot \frac{n+1}{n-r+1}$ (6.1), согласно теореме 3.4. Таким образом, увеличивая m и n мы можем достичь сколь угодно малого ε. При этом $1/\varepsilon$ линейно зависит от размера подматрицы (при $m = n$).

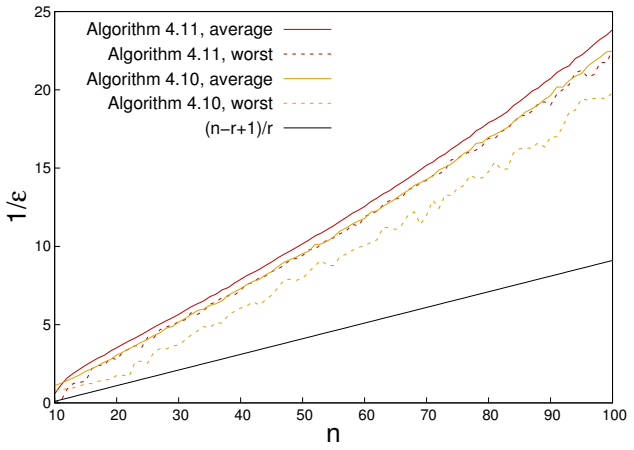
Далее ограничим число переходов между строками и столбцами: их будет не больше 4-х (две итерации внешнего цикла в алгоритмах 4.3 и 4.9). Перед каждым переходом ограничим число замен значениями m и n соответственно. Тогда, предполагая что алгоритм maxvol-proj достигает оценок (6.1), то есть $\varepsilon \leqslant \frac{r}{n-r+1}$, на это потребуется $O(Nr^2/\varepsilon^2)$ операций. Быстрый поиск большого проективного объема (алгоритм 4.10) строит аппроксимацию того же вида за $O(Nr^2/\varepsilon)$ операций, но должен быть использован с осторожностью, поскольку rect-maxvol не дает никаких гарантий на ρ-локально максимальный объем найденных подматриц. Тем не менее, как мы увидим, он работает почти так же эффективно, как и maxvol-proj. Напомним еще раз, что рассматриваемые здесь алгоритмы крестовой аппроксимации не используют всей матрицы, а потому для них не может существовать никаких теоретических гарантий в худшем случае.

Сначала проверим, что значение $1/\varepsilon$ действительно ведет себя как n/r для $n \times n$ подматриц при $n \gg r$. Эта гипотеза подтверждается рисунком 6.3, где значение $1/\varepsilon$ зависит линейно от размера подматрицы и выше ожидаемого $1/\varepsilon = \frac{n-r+1}{r}$. Тем не менее, встречаются и случаи, когда коэффициент $1+\varepsilon$ больше $1+\frac{r}{n-r+1}$ (т. е., когда $1/\varepsilon$ меньше $\frac{n-r+1}{r}$). Заметим, что они встречаются только когда сингулярные числа (после r-го) матрицы сильно отличаются друг от друга.

Увеличение числа строк и столбцов также существенно уменьшает дисперсию ошибки, так что вероятность большой погрешности падает (см. также теорему 3.6). Чтобы проверить этот факт, построим плотность вероятности погрешности крестовых аппроксимаций.



(a) $\sigma_k\ (A) = 1/2^k$.



(b) Первые $r = 10$ сингулярных чисел A равны 20, остальные равны 1.

Рис. 6.3: Зависимость $1/\varepsilon$ (6.3) от числа строк и столбцов подматрицы при использовании различных алгоритмов поиска большого проективного объема. Размер матрицы 400×400, ранг $r = 10$, сингулярные числа матрицы A выписаны под рисунками. Погрешность усреднена по 100 генерациям матрицы A. Меньшее значение $1/\varepsilon$ соответствует большей ошибке. Также показано наименьшее значение $1/\varepsilon$ за 100 генераций (наихудший случай, worst case).

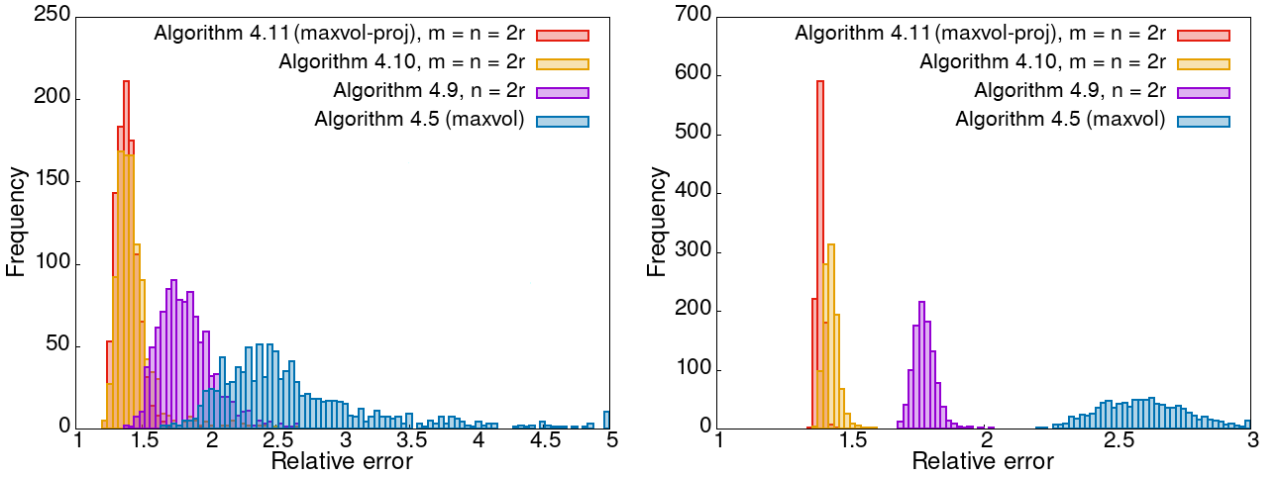
Соответствующие гистограммы показаны на рисунке 6.4. Легко видеть, что не только величина ошибки, но и её дисперсия существенно меньше при использовании $\mathrm{maxvol\text{-}proj}$ по сравнению с алгоритмом maxvol.

Все случаи, когда ошибка превзошла правую границу графика перенесены в последний столбец гистограмм. Такие случаи (когда погрешность превосходила сингулярное разложение более, чем в 4 раза) наблюдались только для алгоритма maxvol (7 случаев из 1000).

Как видно, крестовые аппроксимации на основе найденных подматриц приводят к приближениям по норме Фробениуса близких к оптимальному, заданному сокращенным сингулярным разложением, но при этом существенно быстрее. Несмотря на формальное отсутствие теоретических гарантий для них, мы видим, что идея поиска подматриц большого проективного объема приводит к падению погрешности и её дисперсии, как и предсказывают вероятностные оценки главы 3.

В таблице 6.3 мы сравниваем эффективность методов крестовой аппроксимации для матрицы A большего размера. Как видно, аппроксимация на основе проективного объема дает погрешность, близкую к сокращенному сингулярному разложению, и превосходит по точности и скорости столбцовую аппроксимацию на основе сильного выявляющего ранг QR разложения [39]. Заметим также, что алгоритм 4.10 быстрее и вносит меньшую погрешность, чем аппроксимация на основе прямоугольной подматрицы локально максимального объема (алгоритм 4.9). Таким образом, на практике нет преимущества в использовании прямоугольных матриц для построения крестовой аппроксимации.

Описанные алгоритмы и некоторые примеры доступны в GitHub:



(a) $\sigma_k\ (A) = 1/2^k$.

(b) Первые $r = 20$ сингулярных чисел A равны 20, остальные равны 1.

Рис. 6.4: Плотность вероятности относительной погрешности $\|A - CGR\|_F\ /\|A - A_r\|_F$ для различных алгоритмов крестовой аппроксимации. Размеры матрицы 400×400, ранг $r = 20$, сингулярные числа матрицы A указаны под рисунками. Распределение построено исходя из 1000 генераций матрицы A.

Таблица 6.3: Время выполнения и относительная (к сокращенному сингулярному разложению) погрешность по норме Фробениуса для различных алгоритмов крестовой и столбцовой аппроксимации. Размер матрицы $N = 5000$, ранг аппроксимации $r = 25$, $\sigma_k\ (A) = 2^{-k}$.

Метод	Алгоритм 4.11, $m = n = 2r$	Алгоритм 4.10, $m = n = 2r$	Алгоритм 4.9, $n = 2r$	Алгоритм 4.5	SRRQR [39]	SVD
Время, секунд	0,19	0,07	0,09	0,04	1,22	1286
Коэффициент погрешности	1,35	1,39	1,89	2,45	1,85	1

https://github.com/RodniO/Projective-volume-low-rank

Крестовые алгоритмы доступны в модуле ModAppr, а примеры их использования есть в файлах ExampleA.f90 и ExampleB.f90 папки incfiles.

Глава 7. Примеры задач, где необходимы быстрые крестовые аппроксимации

7.1. Уравнения Смолуховского

Данный раздел частично содержит результаты, описанные в кандидатской диссертации соискателя [109].

Классические (дискретные) уравнения Смолуховского представляют собой бесконечную систему дифференциальных уравнений вида

$$\frac{d}{dt}n_k = \frac{1}{2}\sum_{i+j=k} C_{ij}n_i n_j - \sum_{j=1}^{\infty} C_{kj}n_k n_j, \quad k = \overline{1,\infty}, \tag{7.1}$$

где n_i – концентрации частиц размера i, меняющиеся со временем, а C_{ij} – постоянное ядро агрегации. Уравнения Смолуховского применяются для описания агрегации и сборки белков и нанополимеров [110, 111], формирования дождя [112], агрегации сажи [113], формирования протопланет [114] и планетарных дисков [115].

Решение классических уравнений Смолуховского (7.1) само по себе уже является крайне сложной задачей. Численное решение бесконечной системы в принципе невозможно, в то время как для точного приближения решений могут потребоваться тысячи и миллионы (связанных между собой) уравнений. Поэтому на практике часто применяются методы Монте-Карло [116, 117, 118, 119], где система моделируется конечным числом частиц. Однако, в последние годы развитие получили также малоранговые методы решения [120], которые превосходят методы Монте-Карло одновременно в точности и вычислительной сложности. Они основаны на малоранговом разложении конечномерного ядра $C \in \mathbb{R}^{N\times N}$ ранга r. В этом случае сложность решения системы N уравнений падает с $O\left(N^2\right)$ до $O\left(Nr\log N\right)$ [20, 121, 122].

Ситуация, однако, усложняется, когда ядро C_{ij} меняется во времени. Такое происходит, например, если учесть эволюцию кинетических энергий (температур) частиц различного размера. Для этого вводится еще одна бесконечная система на парциальные температуры T_i [54]:

$$\frac{d}{dt}n_k = \frac{1}{2}\sum_{i+j=k} B_{ij}n_i n_j - \sum_{j=1}^{\infty} D_{kj}n_k n_j, \quad k = \overline{1,\infty}, \tag{7.2}$$

где ядра $B_{ij} = B_{ij}\left(T_i, T_j\right)$, $D_{ij} = D_{ij}\left(T_i, T_j\right)$ и ядро агрегации $C_{ij} = C_{ij}\left(T_i, T_j\right)$ теперь зависят от времени через парциальные температуры. Таким образом, теперь требуется строить аппроксимацию каждого из ядер на каждом шаге по времени, что требует очень быстрых и надежных методов малоранговой аппроксимации. Именно такие методы (основанные на крестовой аппроксимации) и рассмотрены в данной работе.

Для начала обозначим $A_{ij} = C_{ij} n_i n_j$ и построим малоранговую аппроксимацию матрицы $A \in \mathbb{R}^{N \times N}$:

$$A_{ij} \approx \sum_{l=1}^{r} U_{il} V_{jl}, \tag{7.3}$$

где r – ранг аппроксимации. Уравнение (7.3) также может быть записано в матричном виде

$$A \approx U V^T, \tag{7.4}$$

где $U, V \in \mathbb{R}^{N \times r}$ – факторы разложения.

Сразу заметим, что по сравнению с [20, 121, 122], мы аппроксимируем матрицу A, а не матрицу C частоты агрегации. Это позволяет учесть плотность числа частиц, сосредоточившись на наиболее «заселенной» части и игнорируя «хвост» распределения. Схожая идея, тем не менее, уже была использована при аппроксимации многокомпонентной агрегации [120].

Факторы U и V могут быть получены за $O(Nr^2)$ операций с помощью алгоритма maxvol (алгоритм 4.5). Его адаптация, позволяющая постепенно менять ранг (когда это требуется), будет описана далее (алгоритм 7.3). Аппроксимации матрицы $B_{ij} n_i n_j$ и $D_{ij} n_i n_j$ используются точно так же, как и аппроксимация $A_{ij} = C_{ij} n_i n_j$, поэтому далее мы сосредоточимся на матрице A.

Обозначим через

$$U_{i,:} = [U_{i1}, ..., U_{ir}], \quad V_{i,:} = [V_{i1}, ..., V_{ir}], \quad i = \overline{1, N} \tag{7.5}$$

строки матриц U и V.

Подставляя (7.4) в уравнения Смолуховского (7.1) и используя обозначение (7.5), получаем

$$\frac{d}{dt} n_k = \frac{1}{2} \sum_{i+j=k} U_{i,:} V_{j,:}^T - U_{k,:} \left(\sum_{j=1}^{N} V_{j,:} \right)^T, \quad k = \overline{1, N}. \tag{7.6}$$

Вычислив вектор сумм

$$s = \sum_{j=1}^{N} V_{j,:}, \quad s \in \mathbb{R}^r \tag{7.7}$$

заранее, мы уменьшаем сложность вычисления произведения $U_{k,:} \left(\sum_{j=1}^{N} V_{j,:} \right)^T$ с $O(N^2)$ до $O(Nr)$. Первое слагаемое в (7.6), однако, требует более детального анализа.

Первое слагаемое в правой части (7.6) можно записать как

$$\sum_{i+j=k} U_{i,:} V_{j,:}^T = \sum_{l=1}^{r} \sum_{i+j=k} U_{il} V_{jl} = \sum_{l=1}^{r} \sum_{i=1}^{k-1} U_{il} V_{k-i,l},$$

то есть как сумму r дискретных сверток векторов-столбцов $U_{:,l}$ и $V_{:,l}$.

Дискретные свертки могут быть вычислены с использованием быстрого преобразования Фурье (FFT). Аналогично (7.5) обозначим

$$U_{:,j} = \begin{bmatrix} U_{1j} \\ \vdots \\ U_{Nj} \end{bmatrix}, \quad V_{:,j} = \begin{bmatrix} V_{1j} \\ \vdots \\ V_{Nj} \end{bmatrix}, \quad j = \overline{1,r}$$

столбцы матриц U и V.

Сначала добавим нулевой столбец $0_N = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^N$, чтобы получить столбцы $\begin{bmatrix} U_{:,l} \\ 0_N \end{bmatrix} \in \mathbb{R}^{2N}$ и $\begin{bmatrix} V_{:,l} \\ 0_N \end{bmatrix} \in \mathbb{R}^{2N}$ для каждого $l = \overline{1,r}$.

Мы можем вычислить все r дискретных сверток, применив преобразование Фурье для $U_{:,l}$ и $V_{:,l}$, затем вычислив их поэлементное произведение и, наконец, применив обратное преобразование Фурье. Обозначим

$$u^{l,FFT} = \mathrm{FFT}\left(\begin{bmatrix} U_{:,l} \\ 0_N \end{bmatrix}\right) \in \mathbb{R}^{2N}, \quad v^{l,FFT} = \mathrm{FFT}\left(\begin{bmatrix} V_{:,l} \\ 0_N \end{bmatrix}\right) \in \mathbb{R}^{2N}, \quad l = \overline{1,r} \tag{7.8}$$

соответствующие преобразования Фурье. Используя обратное преобразование Фурье IFFT заменим первое слагаемое в (7.6):

$$\frac{d}{dt} n_k = \frac{1}{2}\left[\mathrm{IFFT}\left(\sum_{l=1}^{r} u^{l,FFT} \odot v^{l,FFT}\right)\right]_k - U_{k,:} s^T, \quad k = \overline{1,N}, \tag{7.9}$$

где $\odot$ обозначает поэлементное произведение. $2r$ быстрых преобразований Фурье и одно обратное преобразование Фурье могут быть вычислены за $O(Nr \log N)$.

После того как найдены факторы малорангового разложения и вычислены свертки можно использовать произвольную (явную) схему, чтобы вычислить (7.9). Следуя [20], мы будем использовать схему предиктор-корректор второго порядка.

Заметим, что на практике необходимо также избавиться от очень малых или отрицательных элементов, возникающих в результате неточности (машинного эпсилон) при реализации преобразования Фурье. В целом малоранговый метод для температурно-зависимых уравнений записан далее как алгоритм 7.1.

С течением времени частицы агрегируют, и их общее число снижается, как и общая частота столкновений. Поэтому важно обновлять шаг по времени с учетом данного факта.

Предполагая, что производные $n_i(t)$ и $n_i T_i(t)$ разных порядков при фиксированных i и t отличаются не сильно, имеет смысл выбирать шаг Δt так, чтобы отношение

$$\tau = \frac{\max\limits_i |n_i(t + \Delta t) - n_i(t)|}{\max\limits_i n_i(t)} \tag{7.10}$$

Алгоритм 7.1 Малоранговый метод для температурно-зависимых уравнений Смолуховского

Вход: Максимальный размер кластеров N.

Начальные плотности числа частиц каждого размера n_i, $\quad i = \overline{1, N}$.

Начальные парциальные температуры T_i, $\quad i = \overline{1, N}$.

Ядра C_{ij}, B_{ij} и D_{ij} как функции T_i и T_j.

Параметр шага τ (описан далее).

Конечное лабораторное время t_max.

Выход: T_i – температура частиц размера i в момент времени t_max.

n_i плотность числа частиц размера i в момент времени t_max.

1: $curtime := 0$

2: **while** $curtime < t\_max$ **do**

3: Строим малоранговые аппроксимации матриц с элементами $A_{ij} = C_{ij} n_i n_j$, $B_{ij} n_i n_j$ и $D_{ij} n_i n_j$

4: Вычисляем dn_i/dt и $d\left(n_i T_i\right)/dt$ с помощью уравнений (7.7)-(7.9)

5: $\Delta t := \tau \min\left(\frac{\max\limits_i n_i}{\max\limits_i |dn_i/dt|}, \frac{\max\limits_i (n_i T_i)}{\max\limits_i |d(n_i T_i)/dt|}\right)$

6: Вычисляем $n_i\left(t + \Delta t\right)$ и $\left(n_i T_i\right)\left(t + \Delta t\right)$ с помощью схемы предиктор-корректор второго порядка

7: $T_i := \left(n_i T_i\right)/n_i$

8: $curtime := curtime + \Delta t$

9: **end while**

оставалось постоянным (аналогично для $n_i T_i$). В тех же предположениях данный выбор шага применим и к схемам более высокого порядка. Данный адаптивный шаг был проверен для постоянного ядра, результаты записаны в таблицах 7.1 и 7.2.

Из таблиц видно, что наблюдается сходимость порядка $O\left(\tau^2\right)$. При этом адаптивный шаг не сильно увеличивает ошибку по сравнению с постоянным (минимальным) шагом. Погрешности по 2-норме для плотности числа частиц и масс были вычислены по формулам

$$\begin{aligned} n^{\text{error}} &= \sqrt{\frac{\sum\limits_{i=1}^{N}\left|n_i^{\text{num}}(t)-n_i\right|^2}{\sum\limits_{i=1}^{N} n_i^2(t)}}, \\ m^{\text{error}} &= \sqrt{\frac{\sum\limits_{i=1}^{N}\left|in_i^{\text{num}}(t)-in_i\right|^2}{\sum\limits_{i=1}^{N} (in_i)^2\,(t)}}, \end{aligned} \tag{7.11}$$

где n_i^{num} – численные решения. Аналитические решения в случае монодисперсных начальных условий $n_k\,(t=0)=\delta_{1k}$ записываются как [123]

$$\begin{aligned} n_k(t) &= \left(\frac{t}{t+1}\right)^{k-1}(1+t)^{-2}, \text{ если } C_{ij}=2, \\ n_k(t) &= \frac{k^{k-1}}{k!}e^{-t}\left(1-e^{-t}\right)^{k-1}e^{-k\left(1-e^{-t}\right)}, \text{ если } C_{ij}=i+j. \end{aligned} \tag{7.12}$$

Второе уравнение является решением для линейного ядра и используется ниже.

Таблица 7.1: Относительные ошибки (7.11) распределений плотности числа частиц n и масс m для решения классических уравнений Смолуховского с ядром $C_{ij}=2$ и шагом по времени, зависящим от τ. Размер системы $N=5000$, так что число частиц в «хвосте» распределения пренебрежимо мало.

Время t	τ	Погрешность n	Погрешность m
10	0,1	$6{,}38\cdot10^{-3}$	$1{,}54\cdot10^{-3}$
10	0,01	$5{,}89\cdot10^{-5}$	$1{,}47\cdot10^{-5}$
100	0,1	$6{,}62\cdot10^{-3}$	$1{,}30\cdot10^{-3}$
100	0,01	$6{,}05\cdot10^{-5}$	$1{,}22\cdot10^{-5}$
1000	0,1	$6{,}96\cdot10^{-3}$	$1{,}63\cdot10^{-3}$
1000	0,01	$6{,}43\cdot10^{-5}$	$1{,}54\cdot10^{-5}$

Далее мы будем использовать адаптивный шаг с $\tau=0{,}01$.

Стоит отметить, что для решения уравнений Смолуховского лучшей известной альтернативой малоранговому методу является метод конечных объемов с неравномерной сеткой (NFVS)

Таблица 7.2: Относительные ошибки (7.11) распределений плотности числа частиц n и масс m для решения классических уравнений Смолуховского с ядром $C_{ij} = 2$ и шагом по времени $\Delta t = \Delta t(\tau_0)$, где $\tau_0 = \tau(t = 0)$. Размер системы $N = 5000$, так что число частиц в «хвосте» распределения пренебрежимо мало.

Время t	τ_0	Погрешность n	Погрешность m
10	0,1	$1{,}14 \cdot 10^{-3}$	$1{,}55 \cdot 10^{-4}$
10	0,01	$1{,}07 \cdot 10^{-5}$	$1{,}79 \cdot 10^{-6}$
100	0,1	$5{,}88 \cdot 10^{-4}$	$1{,}05 \cdot 10^{-5}$
100	0,01	$4{,}14 \cdot 10^{-6}$	$8{,}70 \cdot 10^{-6}$
1000	0,1	$2{,}14 \cdot 10^{-4}$	$1{,}62 \cdot 10^{-4}$
1000	0,01	$1{,}82 \cdot 10^{-5}$	$1{,}55 \cdot 10^{-5}$

[124]. Он позволяет учесть «хвост» без существенного роста размера системы. Для малорангового метода в качестве альтернативы мы предлагаем аппроксимировать хвост распределения для размеров, больших N, с помощью экспоненты.

А именно, предположим, что концентрации можно аппроксимировать как

$$n_k \approx a x^b e^{-ck}$$

при достаточно больших k. Для оценки параметров a, b и c можно воспользоваться произвольными тремя значениями, близкими к N:

$$\begin{aligned} v_1 &= n_{N-2s} \\ v_2 &= n_{N-s} \\ v_3 &= n_N \end{aligned}$$

с небольшим шагом s (мы использовали $s = \lfloor \ln N \rfloor$). Тогда

$$\begin{aligned} b &= \frac{\ln \frac{v_1 v_3}{v_2^2}}{\ln\left(1 - \frac{s^2}{(n-s)^2}\right)}, \\ c &= \frac{\ln \frac{v_3}{v_2} - b \ln\left(1 + \frac{s}{n-s}\right)}{s}, \\ a &= \frac{v_3}{N^b} e^{-cN}. \end{aligned}$$

Если требуется оценить полное число частиц с размером, большим N, это можно сделать, заменив сумму неполной гамма-функцией:

$$\sum_{i=N+1}^{\infty} n_i \approx a c^{-b-1} \Gamma\left(b + 1, c\left(N + 0{,}5\right)\right).$$

Полная масса в «хвосте» может быть оценена, если использовать коэффициент $b + 1$ вместо b, что соответствует умножению n_i на i.

Далее нам требуется учесть, как данная аппроксимация влияет на размеры, меньшие N. Для этого применим аналогичную аппроксимацию для строк $C_{ij}n_in_j$ и оценим сумму $\sum_{j=N+1}^{\infty} C_{kj}n_kn_j$. При использовании малорангового разложения

$$C_{ij}n_in_j = UV^T, \quad U \in \mathbb{R}^{N\times r}, \quad V \in \mathbb{R}^{N\times r}$$

требуется только r вычислений неполной гамма-функции. В результате стоимость такого приближения пренебрежимо мала по сравнению с вычислительной сложностью всего алгоритма.

Можно также использовать более простую чисто экспоненциальную модель, для которой при $s = 1$

$$\sum_{i=N+1}^{\infty} n_i \approx \frac{n_N^2}{n_{N-1} - n_N}. \tag{7.13}$$

Важно заметить, что, в итоге, мы не требуем того, чтобы хвост был экспоненциальным, а лишь того, чтобы итоговый интеграл приближался экспонентой. Например, если бы $n_k \sim e^{-k^2}$, то её нельзя было бы приблизить с высокой относительной точностью, однако уравнение (7.13) в этом случае содержит относительную ошибку порядка $O(1/N)$.

Преимущества данного метода показаны в таблицах 7.3, 7.4, 7.5 и 7.6, где были использованы постоянное и линейное ядро. Метод аппроксимации хвоста сравнивается со схемой конечных объемов (NFVS) [124]. Как и в оригинальной работе [124], мы использовали экспоненциальные размеры ячеек, а именно $\Delta x_k = 2\lfloor 1{,}2^k \rfloor - 1$ для постоянного ядра и $\Delta x_k = 2\lfloor 1{,}05^k \rfloor - 1$ для линейного ядра, с тем же общим числом уравнений, что и для малорангового метода.

В таблицах 7.3 и 7.4 также показаны общие «потери» массы (масса в хвосте). Если бы использовался метод без аппроксимации хвоста, вся эта масса не учитывалась бы. Более того, она бы существенно повлияла на скорость агрегации частиц малого размера. В частности, для постоянного ядра в момент времени $t = 1000000$ понадобились бы миллионы уравнений, чтобы достичь той же относительной точности.

Время вычислений показано в таблицах 7.3 и 7.4. Постоянное и линейное ядро сами по себе являются малоранговыми, поэтому они были разложены аналитически, без использования крестовой аппроксимации. Заметим, что для линейного ядра адаптивный шаг остается почти постоянным: это связано с тем, что для него скорость агрегации не убывает со временем.

Как видно из таблицы 7.3, малоранговому методу потребовалось всего в 2 раза больше времени вычислений, чтобы достичь в 1000 раз большего лабораторного времени, чего не удалось бы сделать, если бы использовался постоянный шаг по времени.

На рисунке 7.1 численные решения малорангового метода и NFVS сравниваются с аналитическими решениями.

Таблица 7.3: Относительные ошибки (7.11) распределений плотности числа частиц n и масс m для решения классических уравнений Смолуховского с ядром $C_{ij} = 2$ для малорангового метода (Low-rank) и NFVS. Число уравнений $N = 100$ в обоих методах. Относительные ошибки вычислены для размеров $i \leqslant N$. См. также рисунки 7.1a-7.1b.

Метод, время t	Погрешность n	Погрешность m	Масса в хвосте, %	время выч-й, сек
Low-rank, $t = 10^3$	$1{,}47 \cdot 10^{-4}$	$1{,}13 \cdot 10^{-4}$	99,53%	$2{,}8 \cdot 10^{-2}$
NFVS, $t = 10^3$	$8{,}44 \cdot 10^{-2}$	$6{,}79 \cdot 10^{-2}$	–	$1{,}4 \cdot 10^{-1}$
Low-rank, $t = 10^6$	$6{,}33 \cdot 10^{-4}$	$6{,}57 \cdot 10^{-4}$	99,999999496%	$5{,}7 \cdot 10^{-2}$
NFVS, $t = 10^6$	$9{,}95 \cdot 10^{-2}$	$8{,}43 \cdot 10^{-2}$	–	$2{,}5 \cdot 10^{-1}$

Таблица 7.4: Относительные ошибки (7.11) распределений плотности числа частиц n и масс m для решения классических уравнений Смолуховского с ядром $C_{ij} = i + j$ для малорангового метода (Low-rank) и NFVS. Число уравнений $N = 100$ в обоих методах. Относительные ошибки вычислены для размеров $i \leqslant N$. См. также рисунки 7.1c-7.1d.

Метод, время t	Погрешность n	Погрешность m	Масса в хвосте, %	время выч-й, сек
Low-rank, $t = 2$	$5{,}71 \cdot 10^{-5}$	$4{,}60 \cdot 10^{-5}$	17%	$9{,}0 \cdot 10^{-3}$
NFVS, $t = 2$	$3{,}50 \cdot 10^{-3}$	$3{,}76 \cdot 10^{-2}$	–	$3{,}6 \cdot 10^{-2}$
Low-rank, $t = 3$	$3{,}53 \cdot 10^{-4}$	$2{,}93 \cdot 10^{-4}$	63%	$1{,}2 \cdot 10^{-2}$
NFVS, $t = 3$	$4{,}58 \cdot 10^{-3}$	$5{,}70 \cdot 10^{-2}$	–	$5{,}0 \cdot 10^{-2}$

Из результатов можно сделать вывод, что возможно получение решений, близких к аналитическим, даже когда существенная часть массы находится в «хвосте» и не учитывается напрямую в конечной системе.

Перейдем теперь к вопросу аппроксимации ядер $A_{ij} = C_{ij} n_i n_j$, $B_{ij} n_i n_j$ и $D_{ij} n_i n_j$. Заметим, что алгоритм maxvol (алгоритм 4.5) не гарантирует высокой точности, так как не видит всей матрицы. Поэтому, чтобы повысить вероятность построения аппроксимаций высокой точности, имеет смысл разбить исходные ядра на части. В частности, матрицу D можно разбить как

$$D_{ij} = D_{ij}^{agg} + D_{ij}^{res} + \left(D_{ij}^{exch} - D_{ji}^{exch} \right),$$

где D_{ij}^{agg} соответствует потерям кинетической энергии в результате агрегации, D_{ij}^{res} соответствует потерям в неупругих (реституционных) столкновениях, D_{ij}^{exch} соответствует теплообмену.

Конкретные значения ядер зависят от используемой модели столкновений. Например, в

Таблица 7.5: Относительные ошибки моментов решения уравнений Смолуховского с ядром $C_{ij} = 2$ для малорангового метода (Low-rank) и NFVS. Число уравнений $N = 100$.

Метод, время t	Погрешность 0-го момента (n)	Погрешность 1-го момента (m)	Погрешность 2-го момента
Low-rank, $t = 10^3$	$7{,}5 \cdot 10^{-6}$	$5{,}4 \cdot 10^{-5}$	$9{,}4 \cdot 10^{-5}$
NFVS, $t = 10^3$	$9{,}0 \cdot 10^{-4}$	$1{,}3 \cdot 10^{-9}$	$3{,}7 \cdot 10^{-3}$
Low-rank, $t = 10^6$	$7{,}5 \cdot 10^{-5}$	$1{,}1 \cdot 10^{-3}$	$2{,}2 \cdot 10^{-3}$
NFVS, $t = 10^6$	$6{,}2 \cdot 10^{-3}$	$2{,}5 \cdot 10^{-9}$	$2{,}7 \cdot 10^{-2}$

Таблица 7.6: Относительные ошибки моментов решения уравнений Смолуховского с ядром $C_{ij} = i + j$ для малорангового метода (Low-rank) и NFVS. Число уравнений $N = 100$.

Метод, время t	Погрешность 0-го момента (n)	Погрешность 1-го момента (m)	Погрешность 2-го момента
Tail, $t = 2$	$1{,}3 \cdot 10^{-5}$	$3{,}5 \cdot 10^{-6}$	$5{,}9 \cdot 10^{-5}$
NFVS, $t = 2$	$4{,}9 \cdot 10^{-3}$	$2{,}7 \cdot 10^{-9}$	$1{,}0 \cdot 10^{-1}$
Tail, $t = 3$	$2{,}9 \cdot 10^{-4}$	$9{,}4 \cdot 10^{-4}$	$4{,}7 \cdot 10^{-3}$
NFVS, $t = 3$	$1{,}1 \cdot 10^{-2}$	$5{,}1 \cdot 10^{-9}$	$2{,}1 \cdot 10^{-1}$

отсутствии агрегации при постоянном коэффициенте восстановления ε

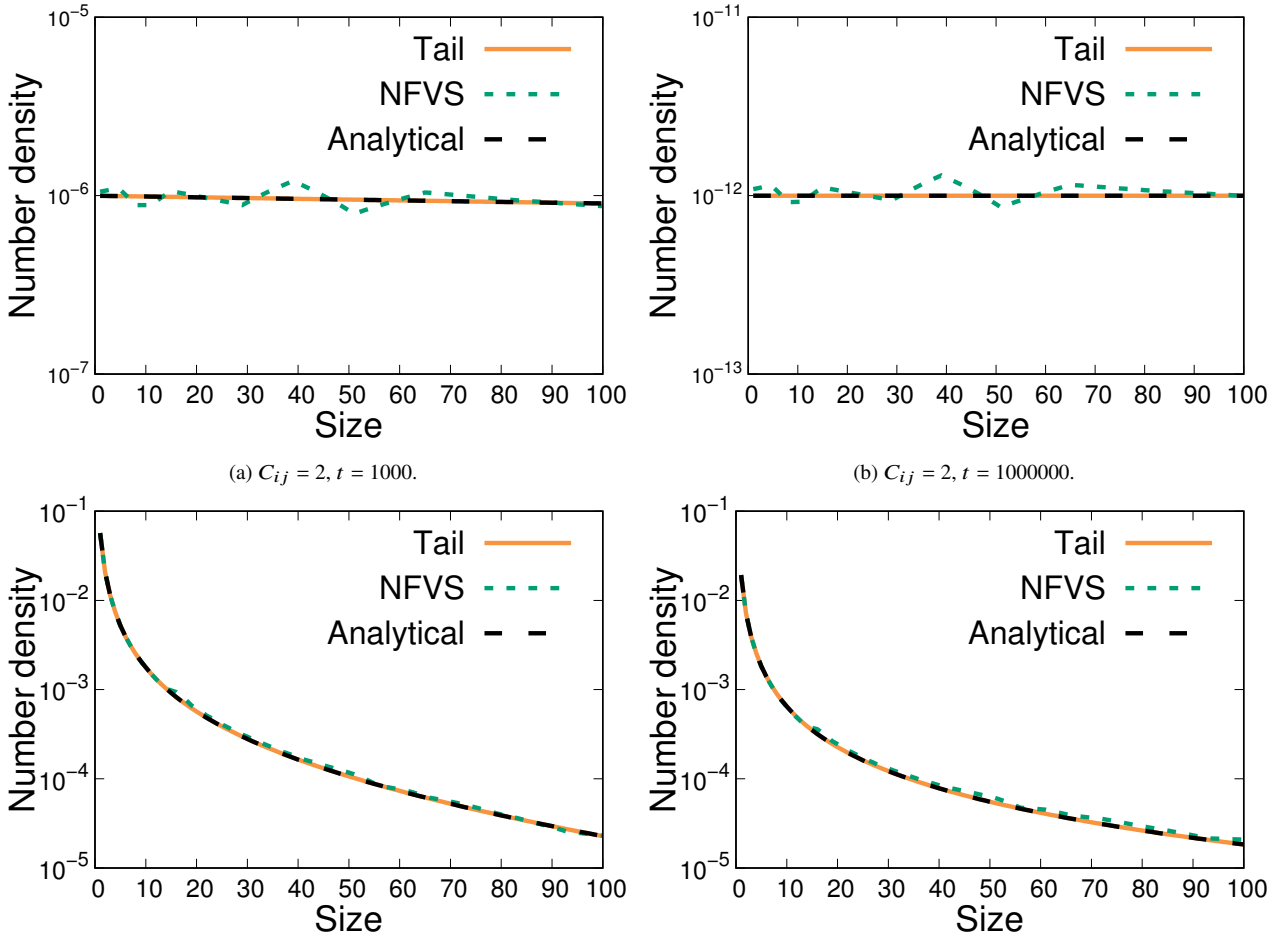
$$
\begin{aligned}
D_{ij}^{agg} &= 0, \\
D_{ij}^{res} &\sim \left(1 - \varepsilon^2\right) T_i \left(R_i + R_j\right)^2 \sqrt{\frac{T_i}{m_i} + \frac{T_j}{m_j}} \cdot \frac{m_j}{m_i + m_j}, \\
D_{ij}^{exch} &\sim (1 + \varepsilon)^2 T_i \left(R_i + R_j\right)^2 \sqrt{\frac{T_i}{m_i} + \frac{T_j}{m_j}} \cdot \frac{m_i m_j}{\left(m_i + m_j\right)^2}.
\end{aligned}
$$

Если же все столкновения приводят к агрегации, то

$$
D_{ij} = D_{ij}^{agg} \sim T_i \left(R_i + R_j\right)^2 \sqrt{\frac{T_i}{m_i} + \frac{T_j}{m_j}} \left(1 + \frac{1}{3} \frac{T_i/m_i}{T_i/m_i + T_j/m_j}\right).
$$

Последнее ядро также может быть разбито на два после раскрытия скобок. Данный подход может быть применен к произвольным ядрам, используемым в обобщенных уравнениях Смолуховского.

Данный метод решения также легко распараллелить. С точки зрения внутренних операций, вычисления в линейной алгебре и быстрое преобразование Фурье автоматически используют многопоточность, если их запускать через библиотеку Intel MKL [125]. Заметим также, что $2r$



(a) $C_{ij} = 2$, $t = 1000$.

(b) $C_{ij} = 2$, $t = 1000000$.

(c) $C_{ij} = i + j$, $t = 2$.

(d) $C_{ij} = i + j$, $t = 3$.

Рис. 7.1: Плотность числа частиц различного размера для решений классических уравнений Смолуховского малоранговым методом и NFVS, по сравнению с аналитическими решениями (7.12). Число уравнений $N = 100$. См. также таблицы 7.3 и 7.4.

прямых преобразований Фурье в (7.9) полностью независимы. Поскольку на больших временах ($t > 100000$) $r \approx 20$, получаем ускорение в 20 раз.

Кроме того, благодаря разбиению матриц на части, соответствующие различным физическим процессам, эти части (как и сами матрицы C, B и D) также могут быть аппроксимированы полностью параллельно.

Наконец, крестовый метод аппроксимации также может быть существенно ускорен, как описано, например, в [126], где для $N = 100000$ получено ускорение в 70 раз на 256 ядрах.

Сравнение малорангового метода с прямым решением уравнений Смолуховского приведено в таблице 7.7. Для $N \geqslant 200$ малоранговый метод уже дает существенное ускорение. И преимущество растет с ростом требуемого числа уравнений. Для тысяч уравнений малоранговый метод оказывается в 60 раз быстрее прямого решения без использования аппроксимации.

На рисунке 7.2 показаны распределение температур частиц небольшого размера и эволюция

средней температуры.

Таблица 7.7: Время решения температурно-зависимых уравнений Смолуховского с параметрами из [54].

Время t	$N =$	Потери массы, %	Масса в хвосте, %	Малоранговый метод, сек	Классический метод, сек
1000	20	0,012%	0,83%	0,27	0,103
10000	200	0,021%	4,96%	6,0	16,6
100000	6400	0,060%	3,52%	530	30862

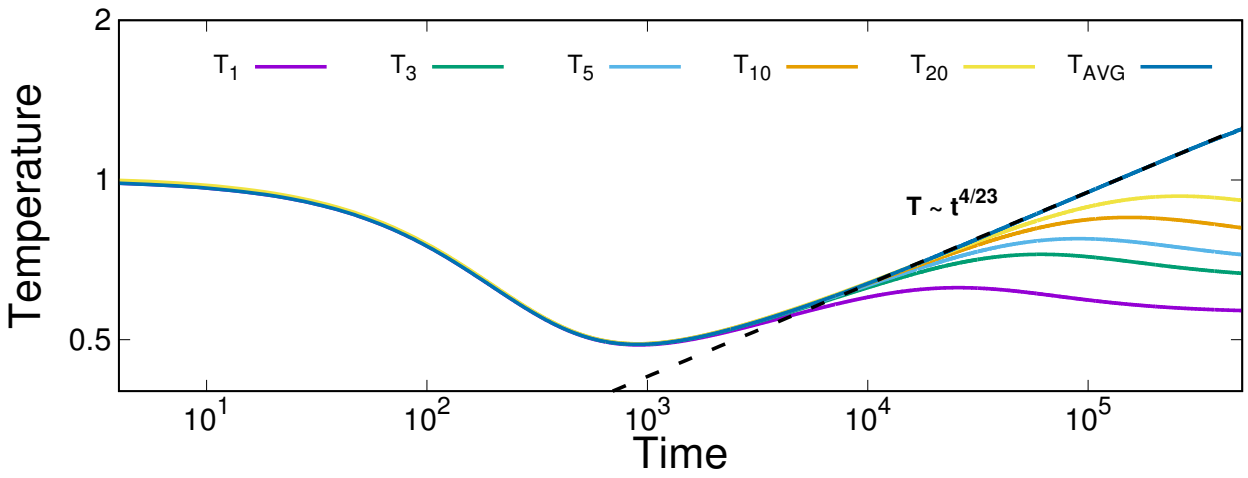
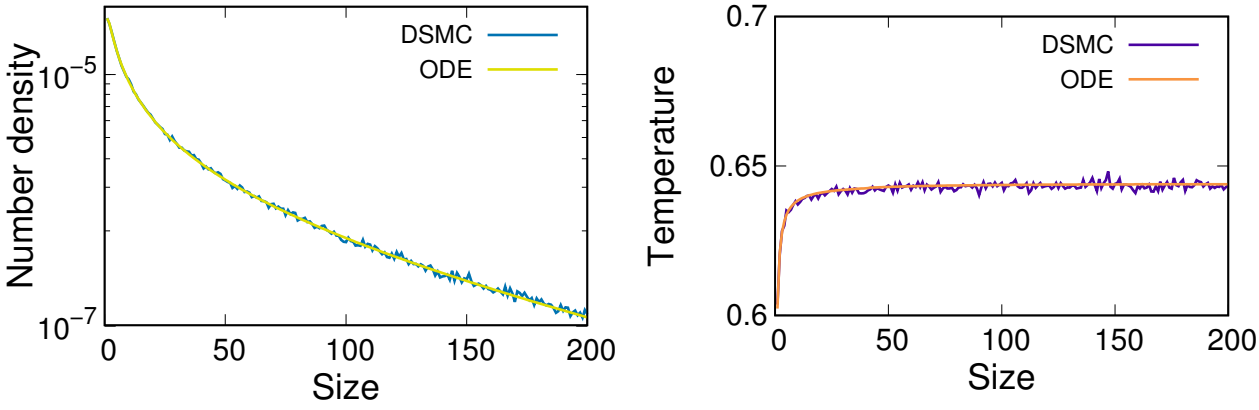


Рис. 7.2: Распределение температур частиц для модели [54]. T_i – температуры частиц размера i. T_{AVG} – средняя температура. Пунктирная линия показывает предсказание теории масштабирования.

Особенностью модели [54] является то, что средняя температура продолжает расти, даже когда индивидуальные температуры (и полная кинетическая энергия системы) падают. Рост соответствует степенному закону с показателем степени около $\approx 0{,}17$, что согласуется с предсказанием $\frac{2\Lambda}{5-\Lambda} = \frac{4}{23}$ теории масштабирования.

Заметим, что использование методов Монте-Карло в данном случае приводит к меньшей точности и большим вычислительным затратам. На рисунках 7.3a и 7.3b виден заметный стохастический шум, ассоциированный с Монте-Карло методами. Время вычисления в методе Монте-Карло при выбранных параметрах также оказывается существенно выше. Преимущество решения уравнений по сравнению с Монте-Карло моделированием состоит в том, что в методах Монте-Карло размер стохастического шума пропорционален $1/\sqrt{N_p}$, где N_p – используемое число частиц. Поскольку время симуляции линейно зависит от числа частиц, получаем сложность $O(1/\varepsilon^2)$, если требуется достичь относительной точности ε. С другой стороны, при решении дифференциальных уравнений требуется время порядка $1/\varepsilon^{1/K}$, где K – порядок разностной схемы, что существенно выгоднее.

Заметим также, что на рисунке 7.3b температуры частиц большого размера совпадают, в то время как температуры частиц малого размера существенно ниже. Таким образом, частицы малого размера не являются репрезентативной выборкой. В [54] были рассмотрены только частицы малого размера, что привело к ложному заключению о том, что средняя температура системы (в модели температурно-зависимых уравнений) рано или поздно начнет падать.



(a) Плотность числа частиц.

(b) Распределение температур.

Рис. 7.3: Плотность числа частиц и распределение температур в момент $t = 10000$ для решения уравнений Смолуховского и симуляций методом Монте-Карло с $N_p = 10^6$ частицами. Время Монте-Карло симуляций составило 284 секунд по сравнению с 6 секундами для малорангового метода.

Опишем теперь, как выгоднее всего применить алгоритм maxvol для построения аппроксимации на каждом шаге по времени. Заметим, что maxvol требует некоторых r заранее выбранных столбцов и стартовую подматрицу $\hat{A} \in \mathbb{R}^{r\times r}$. И то, и другое можно использовать с предыдущего шага по времени, однако со временем ранг r нужно будет обновлять. Это можно сделать с помощью неполного исключения Гаусса (алгоритм 7.2). В качестве критерия остановки мы используем норму Чебышева погрешности из-за наличия гарантий на её величину для крестовых аппроксимаций на основе подматриц локально максимального объема [10].

Мы используем множества индексов $\mathcal{I}$ и $\mathcal{J}$ для обозначения подматрицы $\hat{A} = A_{\mathcal{I},\mathcal{J}}$ со строками из множества $\mathcal{I}$ и столбцами из множества $\mathcal{J}$.

Заметим, что неполное исключение Гаусса само по себе уже задает крестовую аппроксимацию. Однако, применять его напрямую ко всей матрице было бы слишком дорого. Кроме того, погрешность можно далее уменьшить, если множество столбцов $\mathcal{I}$ и соответствующая им матрица V обновляются с помощью алгоритма maxvol в столбцах C. Поэтому исключение Гаусса будет нами использовано только чтобы ограничить ранг r. В численных экспериментах использовалось $\varepsilon = 10^{-7}$ (на практике имеет смысл применять $\varepsilon \sim \tau^2/N$, чтобы общая погрешность точно не превосходила погрешности дискретизации по времени).

Целиком алгоритм малоранговой аппроксимации записан как алгоритм 7.3. Мы начинаем с

Алгоритм 7.2 Неполное исключение Гаусса с полным выбором ведущего элемента

Вход: Матрица $A \in \mathbb{R}^{N \times k}$. Граница погрешности ε.

Выход: Столбцы $A_{:,\mathcal{J}} \in \mathbb{R}^{N \times r}$ и матрица $V \in \mathbb{R}^{r \times k}$ такие, что $\|A - A_{:,\mathcal{J}} V\|_C \leqslant \varepsilon \|A\|_C$.

$\delta := \varepsilon \|A\|_C$
$E := A$
$\mathcal{I}, \mathcal{J} := \emptyset$
for $r := 1$ **to** k **do**
    $\{i, j\} := \arg\max_{i,j} \left|E_{i,j}\right|$
    $\mathcal{I} := \mathcal{I} \cup \{i\}$
    $\mathcal{J} := \mathcal{J} \cup \{j\}$
    $E := E - E_{:,j} E_{ij}^{-1} E_{i,:}$
    if $\|E\|_C \leqslant \delta$ **then**
        break
    end if
end for
$V := A_{\mathcal{I},\mathcal{J}}^{-1} A_{\mathcal{I},:}$

поиска аппроксимации немного большего ранга $k = r + \Delta r$ ($\Delta r = 4$). Затем исключение Гаусса используется, чтобы выбрать подходящий ранг $r \leqslant k$ и, наконец, алгоритм maxvol применяется, чтобы найти подматрицу локально максимального объема в r столбцах.

Поскольку аппроксимация применяется на каждом шаге, а ядро меняется медленно, maxvol обычно делает $O(r)$ замен, что приводит к общей сложности $O(Nr^2)$. Исключение Гаусса в фиксированных столбцах также требует $O(Nkr) = O(Nr^2)$ операций. Кроме малорангового разложение наиболее дорогой операцией остаются быстрые преобразования Фурье, которые требуют $O\left(Nr \log N\right)$ операций. В итоге один шаг по времени имеет вычислительную сложность $O\left(Nr(r + \log N)\right)$.

Покажем теперь, что алгоритму maxvol обычно не требуется делать много дополнительных шагов после каждого изменения матрицы A. Допустим, например, что после шага по времени ядро изменилось, и нам нужно искать подматрицу в матрице $A + E$, где норма столбцов E достаточно мала. Пусть нам уже известна подматрица ρ-локально максимального объема в A. Означает ли это, что алгоритму maxvol потребуется существенно меньше замен, чтобы найти подматрицу ρ-локально максимального в $A + E$, если стартовать со столбцов, соответствующих уже найденной подматрице $\hat{A}$?

К сожалению, нет. Как пример можно рассмотреть матрицу

$$A = \begin{bmatrix} I_r & \rho F_r \end{bmatrix} \in \mathbb{C}^{r \times 2r},$$

Алгоритм 7.3 Малоранговая аппроксимация для уравнений Смолуховского

Вход: Матрица $A \in \mathbb{R}^{N \times N}$.

Граница погрешности ε.

Максимальный рост ранга Δr.

Начальная подматрица $\hat{A} = A_{\mathcal{I},\mathcal{J}} \in \mathbb{R}^{r \times r}$, где множества индексов $\mathcal{I}$ и $\mathcal{J}$ заданы с предыдущего шага (недостающие добираются случайно).

Выход: Факторы $U, V \in \mathbb{R}^{N \times r}$ малорангового разложения A: $A \approx UV^T$.

1: $k := r + \Delta r$

2: Δr индексов добавляются в $\mathcal{I}$ и $\mathcal{J}$

3: $A^T_{\mathcal{I},\mathcal{J}} := maxvol(A^T_{\mathcal{I},:}, \hat{A} = A^T_{\mathcal{I},\mathcal{J}})$
{Сейчас $|\mathcal{I}| = |\mathcal{J}| = k$}

4: К столбцам $A_{:,\mathcal{J}}$ применяется неполное разложение Гаусса и определяется r
{Теперь $|\mathcal{I}| = |\mathcal{J}| = r$}

5: $A_{\mathcal{I},\mathcal{J}} := maxvol(A_{:,\mathcal{J}}, \hat{A} = A_{\mathcal{I},\mathcal{J}})$

6: $U := A_{:,\mathcal{J}}$

7: $V := A^{-1}_{\mathcal{I},\mathcal{J}} A_{\mathcal{I},:}$

где F_r – матрица преобразования Фурье, все элементы которой равны 1 по модулю. Единичная матрица здесь является подматрицей ρ-локально максимального объема, однако любое сколь угодно малое уменьшение нормы всех столбцов I_r или увеличение нормы всех столбцов F_r приведет к тому, что единственной подматрицей ρ-локально максимального объема окажется подматрица на месте ρF_r, что потребует r замен от алгоритма maxvol. В действительном случае F_r можно заменить на матрицу Адамара H_r, составленную из 1 и -1 и, так же как и матрицы Фурье, обладающую максимальным объемом среди всех $r \times r$ матриц, с элементами, не превосходящими по модулю 1. Хотя матрицы Адамара существуют не для всех r, их можно построить для $r = 2^m$ (таким образом, как минимум $\lceil r/2 \rceil$ шагов точно может потребоваться сделать).

Тем не менее из этого же примера видно, что если возмущение привело к большому числу шагов, алгоритм приводит нас к подматрице гораздо большего объема, а потому новое возмущение вряд ли вызовет такую же реакцию. Потому в случае, когда поиск доминантной подматрицы осуществляется на каждой итерации некоторой процедуры, между шагами которой матрица A получает возмущение, общее число шагов алгоритма возможно ограничить. При достаточно малом возмущении это ограничение действительно осуществимо. Для его доказательства нам потребуется следующая лемма.

Лемма 7.1. *Для произвольных квадратных матриц $A, E \in \mathbb{C}^{r\times r}$ справедливо неравенство*

$$|\det(A+E) - \det A| \leqslant |\det A| \left(\left(1 + \frac{\max\limits_j \left\| (E_i)_{:,j} \right\|_2}{\sigma_r(A)} \right)^r - 1 \right).$$

Доказательство. Доказательство основано на лемме 4.1, а именно на линейности объема по столбцам, оценке определителя через произведение объемов подматриц и оценке объемов подматриц через произведение норм их столбцов (для подматриц из столбцов из E), либо сверху через сингулярные числа матрицы A (для подматриц из столбцов из A):

$$\begin{aligned}
|\det(A+E) - \det A| &\leqslant \sum_{k=1}^{r} \sum_{\substack{i_1,\ldots,i_k=1 \\ i_m \neq i_n \\ \mathcal{I} = \{i_1,\ldots,i_k\} \\ \mathcal{J} = \{1,\ldots,r\}/\mathcal{I}}}^{r} \mathcal{V}\left(A_{:,\mathcal{I}}\right) \mathcal{V}\left(E_{:,\mathcal{J}}\right) \\
&\leqslant \sum_{k=1}^{r} \sum_{\substack{i_1,\ldots,i_k=1 \\ i_m \neq i_n}}^{r} \sigma_1(A)\ldots\sigma_k(A) \left(\max_j \left\| E_{:,j} \right\|_2 \right)^{r-k} \\
&\leqslant \sum_{k=1}^{r} C_r^k \frac{|\det A|}{\sigma_r^{r-k}(A)} \left(\max_j \left\| E_{:,j} \right\|_2 \right)^{r-k} \\
&\leqslant |\det A| \left(\left(1 + \frac{\max\limits_j \left\| E_{:,j} \right\|_2}{\sigma_r(A)} \right)^r - 1 \right).
\end{aligned}$$

□

Теорема 7.1. *Пусть дана последовательность матриц $A_i = A_{i-1} + E_i$, $i = \overline{1,k}$, $\max\limits_j \left\| (E_i)_{:,j} \right\|_2 \leqslant \varepsilon$. Пусть в каждой из матриц A_i происходит поиск матрицы ρ-локально максимаьного объема с помощью алгоритма* maxvol*, с использованием в качестве стартовой найденной на шаге $i-1$ подматрицы. Пусть при $i = 0$ стартовой является подматрица полученная с помощью алгоритма выбора ведущих столбцов. Пусть*

$$\alpha = \left(1 + \frac{\varepsilon}{\min\limits_i \sigma_r(A_i)} \sqrt{1 + cr(N-r)} \right)^r - 1 < 1.$$

Тогда алгоритму maxvol *понадобится не более* $\left\lceil r \left(1 + \frac{\ln r}{2\ln\rho} + \ln \frac{\ln \frac{(r!)^{2/r}}{r}}{\ln \rho} \right) + k \log_\rho \frac{1+\alpha}{1-\alpha} \right\rceil$ *замен столбцов на поиск по всем матрицам в сумме, что составляет* $O(r \log_\rho r) + O\left(kr \frac{\varepsilon\sqrt{1+\rho^2 r(N-r)}}{\min\limits_i \sigma_r(A_i) \log \rho} \right)$ *замен при достаточно малом ε.*

Доказательство. Заметим, что для любой подматрицы ρ-локально максимального объема справедливо неравенство $\sigma_r(\hat{A}) \geqslant \frac{\sigma_r(A)}{\sqrt{1+\rho^2 r(N-r)}}$ (следствие 4.1). Вместе с доказанной леммой 7.1, из определения α мы получаем, что при переходе от A_{i-1} к A_i объем подматрицы с предыдущего шага падает не более, чем в $1-\alpha$ раз, а объем матрицы максимального объема растет не более, чем в $1+\alpha$ раз. Тем самым за каждый шаг i отношение объема текущей подматрицы к подматрице максимального объема меняется не более, чем в $\frac{1+\alpha}{1-\alpha}$ раз. Так как мы ищем ρ-локально максимальный объем, то каждый шаг maxvol уменьшает отношение как минимум в ρ раз. Таким образом, за $i = \overline{1,k}$ алгоритму потребуется не более $\log_\rho \left(\frac{1+\alpha}{1-\alpha}\right)^k$ дополнительных замен, тогда как число замен, необходимое на старте, уже оценено в разделе 4.3, выражение 4.7 (в условии теоремы мы воспользовались более точной оценкой из [60, 83]). □

7.1.1. Малоранговый метод Монте-Карло

Как уже упоминалось в начале этой главы, до появления малорангового метода решения уравнений Смолуховского они чаще всего решались с помощью различных методов Монте-Карло моделирования [116, 117, 127].

Метод принятия-отклонения, используемый в кинетическом методе Монте-Карло, использует постоянный мажорант $C_{ij} \leqslant C$, где C зависит только от максимального размера частиц в системе. Это позволяет выбирать пары сталкивающихся частиц с помощью равномерного дискретного распределения по их номерам, а затем принимать выбранные кластеры с размерами i и j с вероятностью C_{ij}/C. Однако, значение вероятности принятия C_{ij}/C может быть существенно меньше 1. Так как значение ядра C_{ij} обычно растет полиномиально с размерами, получаем среднюю стоимость порядка $O\left(M^\alpha\right)$, где M – максимальный размер частиц в системе, а α зависит от вида ядра и может быть определен с помощью теории масштабирования [128]. Она предполагает, что ядро имеет вид

$$C_{ai,aj} = a^\lambda \tilde{C}(i,j) + o\left(a^\lambda\right), \quad a \to \infty,$$

где $\tilde{C}(x,y)$ является однородной функцией

$$\tilde{C}(ax,ay) = a^\lambda \tilde{C}(x,y).$$

Кроме того, если у функции $\tilde{C}(1,a)$ при $a \to 0$ доминирует полиномиальный член, можно дополнительно определить параметр μ:

$$\tilde{C}(1,a) = C_0 a^\mu + o\left(a^\mu\right), \quad a \to 0.$$

Для ядер такого вида существует «гипотеза масштабирования» [128], позволяющая сделать выводы о виде решений уравнений Смолуховского и предсказывающая следующую асимптоти-

ческую сложность классического метода принятия-отклонения:

$$\begin{cases} O\left(M^{|\mu|}\right), & \mu < 0,\ \lambda - \mu \leqslant 1, \\ O\left(M^{|\lambda|}\right), & \mu \geqslant 0,\ \lambda \leqslant 1, \end{cases}$$

где константа в сложности зависит от конкретного вида ядра. Другие значения λ и μ приводят к геляции [129], а потому должны рассматриваться отдельно. Таким образом, сложность каждого столкновения оказывается степенной, что приводит к существенному времени вычислений для типичного случая, когда кластеры содержат в себе миллионы мономеров (частиц размера 1).

Однако, в общем случае, в методе принятия-отклонения не обязательно использовать постоянную мажоранту. Достаточно использовать распределение, выбор пар в котором осуществлять было бы проще. В [109] автором было предложено использовать малоранговую мажоранту:

$$C_{ij} = \sum_{k=1}^{r} f_k(i) g_k(j) = \overline{C}_{ij}.$$

В случае, когда близкая к C_{ij} малоранговая мажоранта выписывается аналитически, можно существенно ускорить Монте-Карло метод решения уравнений Смолуховского и соответствующих им уравнений Больцмана. На практике такая ситуация встречается часто. В качестве примеров достаточно привести баллистическое и броуновское ядра, описывающие основные виды движения, приводящие к коагуляции (разреженная система с почти прямыми траекториями между соударениями и плотная система с броуновским движением соответственно). Оба этих ядра, как мы увидим, легко аналитически приближаются сверху с точностью до постоянного множителя.

Выбор сталкивающихся кластеров в ядре $\overline{C}_{ij}$ осуществляется следующим образом. Сначала выбирается одна из компонент ранга 1 с соответствующей ей вероятностью. При условии, что данная компонента k выбрана, размеры кластеров i и j теперь могут быть выбраны независимо на основе $f_k(i)$ и $g_k(j)$. После этого столкновение принимается с вероятностью $C_{ij}/\overline{C}_{ij}$.

Чтобы гарантировать быстрый выбор размеров, остается лишь гарантировать быстрый выбор на основе распределений f_k и g_k. Это можно сделать, построив для $u_k = f_k n$ и $v_k = g_k n$ деревья отрезков (см. рисунок 7.4). В этом случае выбор размеров будет осуществляться за логарифмическое время $O(\log M)$. Кроме того, после каждого столкновения требуется обновить лишь листья, соответствующие размерам i, j и $i + j$, что займет в общей сложности $O(r \log M)$ операций, что и определяет вычислительную сложность одного столкновения. Полученный алгоритм существенно быстрее, чем классический метод принятия-отклонения, требующий $O(M^{\alpha})$ операций. При этом вместо

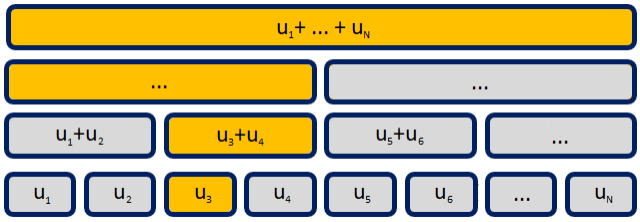


Рис. 7.4: Дерево отрезков для $u_i = f_k(i) n_i$.

бинарных деревьев имеет смысл строить деревья отрезков на основе 4-деревьев, что в 1,5 раза увеличит число сравнений при выборе сталкивающихся кластеров, но в 2 раза уменьшит время обновления.

Имеет смысл сравнить предложенный метод с другими существующими улучшениями метода принятия-отклонения. Например, в мажорантном методе [130] было предложено объединять размеры в «корзины» (bins) из нескольких размеров вместе. Так как значения ядер обычно растут полиномиально с размером, при выборе размеров бинов в геометрической прогрессии (на практике эффективно использовать знаменатель $q = 2$), мы получаем логарифмическое число корзин $O(\log M)$. В каждой корзине размеры отличаются не более, чем в q раз, а потому если использовать нашу идею с мажорированием ядра малоранговым ядром (в оригинальной статье предлагались мажоранты лишь определенного вида), и у этого ядра рост элементов полиномиальный с размером, и при этом в каждой корзине заранее известен максимум, то общая стоимоть алгоритма также будет $O(r \log M)$. Если дополнительно на массиве корзин построить деревья отрезков, то стоимость упадет до $O(r \log \log M)$. На практике, однако, такой метод оказывается менее выгоден, чем предложенный нами выше.

Еще одним улучшением принятия-отклонения является предложенный в [131] стратифицированный метод Монте-Карло на основе метода Волкера. Если ядро позволяет делать обновления не слишком часто (что зависит от вида решения). то данный метод позволяет достичь амортизированной сложности $O(1)$. На практике, однако, это требует эффективного подбора нескольких параметров, а потому сами авторы метода отказались от него в своих дальнейших работах [132] в пользу классического метода принятия-отклонения. Еще одним его недостатком является трудность в применении к ядрам, существенно меняющимся по времени. Тем не менее, при правильной реализации, его скорость на классических ядрах сравнима со скоростью малорангового метода. Сравнение скорости различных методов на линейном ядре $C_{ij} = i + j$ приведено в таблице 7.8. Для постоянного ядра (когда все пары всегда принимаются) одно столкновение требует около 60 нс. Таким образом, на практике логарифмический фактор в сложности малорангового и других быстрых методов отсутствует: выбор пары сталкивающихся кластеров на основе произвольного распределения вероятностей лишь в несколько раз дороже равновероятного выбора.

Таблица 7.8: Наносекунд на одно столкновение, в среднем до достижения $t = 6$ для ядра $C_{ij} = i + j$.

Частиц, N_p	Принятие-отклонение [116], нс	Мажорантный [130], нс	Стратифицированный [131], нс	Малоранговый, нс
10^5	8400	130	100	150
10^7	57000	250	170	150

Все изложенные здесь методы (в том числе усовершенствованные версии методов из [130],

[131], а также малоранговый метод) доступны в GitHub https://github.com/RodniO/Projective-volume-low-rank. Сами методы находятся в модуле ModMonte, а примеры их использования есть в файле ExampleM.f90 папки incfiles.

Стоит еще раз подчеркнуть, что, в отличие от других улучшений метода принятия-отклонения, использование деревьев отрезков позволяет также без введения дополнительных операций обобщить данный метод на случай температурно-зависимых уравнений Смолуховского с ядром, зависящим от времени, а также на уравнения Больцмана. Таким образом, малоранговый метод предпочтителен не только благодаря высокой скорости, но и возможности его применения для широкого класса систем, моделируемых кинетическим методом Монте-Карло.

В частности, с помощью малорангового метода Монте-Карло удалось подтвердить устойчивость точных решений температурно-зависимых уравнений Смолуховского. Мы не будем выписывать здесь вывод точных решений, так как основная идея вывода уже была описана в кандидатской диссертации [109]. Однако, приведем здесь Монте-Карло метод, с помощью которого и удалось проверить устойчивость решения для конечных систем частиц.

Пусть система объема V содержит N_i частиц размера i, так что $n_i = N_i/V$. Пусть T_i – температуры частиц размера i. Столкновения в температурно-зависимом методе Монте-Карло осуществляются следующим образом:

1. Пара размеров (i, j) выбирается с вероятностью $p_{ij} = \frac{C_{ij}N_iN_j}{\sum\limits_{k,l} C_{kl}N_kN_l}$.
2. Обновляется время $t := t + \Delta t$, согласно экспоненциальному распределению с матожиданием $\langle \Delta t \rangle = \frac{V}{\frac{1}{2}\sum\limits_{k,l} C_{kl}N_kN_l}$.
3. Обновляются температуры T_i, T_j, T_{i+j} согласно формулам ниже.
4. Обновляется число частиц участвующих в столкновении размеров:

$$N_i := N_i - 1, \quad N_j := N_j - 1, \quad N_{i+j} := N_{i+j} + 1.$$

Кроме шага обновления температур, остальные шаги те же, что и при решении классических уравнений Смолуховского. Выбор частиц с вероятностями $p_{ij} \sim C_{ij}N_iN_j$ осуществляется с помощью температурно-зависимого малорангового метода.

Обновление температур приводит также к изменению частот столкновений $C_{ij} = C_{ij}(T_i, T_j)$. Для этого используются следующие формулы:

$$T_i := \frac{N_iT_i - D_{ij}/C_{ij}}{N_i - 1},$$
$$T_j := \frac{N_jT_j - D_{ji}/C_{ji}}{N_j - 1},$$
$$T_{i+j} := \frac{N_{i+j}T_i + B_{ij}/C_{ij}}{N_{i+j} + 1}.$$

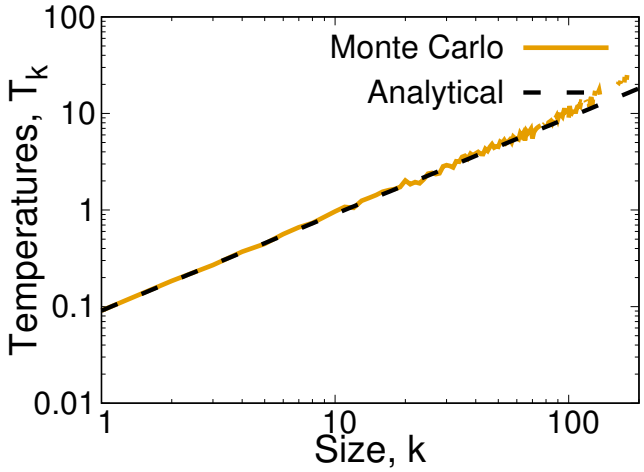


Рис. 7.5: Распределение температур кластеров размера k в момент времени $t = 10$. Решение температурно-зависимых уравнений с ядрами (7.14) получено с помощью температурно-зависимого метода Монте-Карло с 10^4 частицами. Для сравнения показано также точное аналитическое решение.

Одним из основных преимуществ температурно-зависимого метода Монте-Карло является тот факт, что стохастический шум возникает лишь от распределения количества частиц, но не их скоростей. Обновления температур при этом зависят только от текущих температур и размеров сталкивающихся частиц, в то время как в реальности столкновение частиц одного и того же размера, но с разными скоростями могут приводить к изменению температуры (средней кинетической энергии) выше или ниже, чем при усреднении по всем возможным парам частиц и их скоростей (что и происходит при выводе ядер B_{ij} и D_{ij}).

В качестве примера точного решения приведем температурно-зависимые уравнения Смолуховского с ядрами

$$C_{ij} = T_i + T_j, \quad B_{ij} = \left(T_i + T_j\right)^2, \quad D_{ij} = \left(T_i + T_j + 1\right) T_j \tag{7.14}$$

и начальными данными $n_1 = 1, T_1 = 1$ (других частиц в системе изначально нет). В этом случае решением является

$$T_k\left(t\right) = \frac{k}{1+t}, \quad n_k\left(t\right) = \frac{k^{k-1}}{k!\left(t+1\right)} \left(\frac{t}{t+1}\right)^{k-1} e^{-kt/(t+1)}$$

и соответствующее распределение температур легко увидеть при решении методом Монте-Карло даже для небольшого числа частиц в системе (рисунок 7.5).

7.2. Восстановление матриц

Восстановление матриц по небольшому числу известных элементов является важной задачей, имеющей применение в рекомендательных системах [133], кристаллографии [134], интернете вещей (построение карты по измеренным расстояниям между сенсорами) [135], при восстановлении информации о состоянии канала связи на передатчике (CSIT) [136].

Пусть необходимо найти неизвестную матрицу $X \in \mathbb{C}^{N\times N}$ (ограничимся случаем квадратной матрицы, чтобы не возникло путаницы обозначений), если известно её значение $A = M \odot X$ (символ $\odot$ обозначает поэлементное произведение) на некоторой маске $M \in \mathbb{C}^{N\times N}$:

$$M = \begin{cases} 1, & \text{для известных элементов,} \\ 0, & \text{для неизвестных элементов.} \end{cases}$$

Задачу восстановления матрицы сформулируем в виде поиска на элементах M ближайшей (в евклидовой норме) к матрице A матрицы X:

$$\|A - M \odot X\|_F \to \inf_{X,\ \mathrm{rank}\, X \leqslant r}.$$

Если значение на элементах известно точно, то инфимум будет нулевой. Однако, на практике исходная матрица не является в точности матрицей ранга r, а потому значение на заданных элементах наилучшего приближения ранга r может отличаться, и инфимум будет ненулевой. Здесь мы для простоты рассмотрим случай, когда инфимум нулевой. Более общий случай рассмотрен, в частности, в работе [55]. Отметим, что к задаче восстановления можно также свести задачу разделения матрицы на малоранговую и разреженную компоненты [137].

Для восстановления матриц разработано множество методов, различающейся вычислительной сложностью и вероятностью точного восстановления: основанных на прямом решении задачи выпуклого программирования [138], минимизации ядерной нормы [139, 140, 141], атомарном разложении [142], методе чередующихся наименьших квадратов или на других методах попеременной минимизации факторов [143, 144, 145] и оптимизации вдоль многообразий [146, 147]. Рассматриваемый здесь метод ASVP (approximate singular value projection, проектирование приближенным сингулярным разложением) был предложен совместно с Сергеем Владимировичем Петровым и является обобщением метода SVP (проектирования сингулярным разложением) [55]. Здесь будет рассмотрен вариант метода на основе крестового разложения, разработанный автором. В общем случае метод ASVP записывается следующим образом:

$$X_{(s+1)} = P_r\left(X_{(s)} + \tau\left(A - M \odot X_{(s)}\right)\right), \tag{7.15}$$

где $\tau > 0$ – размер шага итерации, а P_r – оператор (приближенного) проектирования на пространство матриц ранга r, $\forall X:\ \ \mathrm{rank}\,(P_r X) = r$. В качестве начального приближения можно использовать $X_{(0)} = P_r A$.

Если $\tau = 1$, получаем метод переменных проекций между пространством матриц ранга r и пространством матриц, совпадающих на маске M с матрицей A. Легко показать [55], что если P_r – точный проектор, то в этом случае невязка на маске M на каждом шаге убывает. Если существует точное решение, и начиная с какого-то шага удалось получить матрицу, достаточно близкую к матрице ранга r, то из результатов [148] следует геометрическая сходимость метода.

На практике, однако, самыми тяжелыми оказываются первые шаги, когда матрица $X_{(0)}$ далека от X, а в этом случае достижение матрицы близкой к матрице ранга r нельзя гарантировать.

Чтобы получить геометрическую сходимость уже начиная с первого шага, в [55] было предложено использовать $\tau > 1$. Для этого случая была доказана следующая теорема.

Теорема 7.2 ([55])**.** *Пусть оператор* $M(X) = \frac{1}{\sqrt{\delta}} M \odot X$ *(δ – доля известных элементов, $\delta = \|M\|_F^2 / N^2$) обладает свойством ограниченной изометрии на множестве матриц X,* $\operatorname{rank} X = 2r$*:*

$$(1-\kappa)\|X\|_F^2 \leqslant \|M(X)\|_F^2 \leqslant (1+\kappa)\|X\|_F^2,$$

где $\kappa < 1/3$. Пусть шаг $\tau = \frac{1}{\delta(1+\kappa)}$. Тогда алгоритм, заданный (7.15)*, при использовании точного проектора $P_r A = A_r$ $\forall A$ достигает на маске M точности ϵ за $\left\lceil \frac{2\log_2\|A\|_F - 2\log_2\epsilon}{\log_2(1/\kappa - 1) - 1} \right\rceil$ итераций.*

Замечание 7.1. Для геометрической сходимости для шага $\frac{1}{2\delta} < \tau < \frac{1}{\delta(1+\kappa)}$ достаточно $\kappa < 1 - \frac{1}{2\delta\tau}$.

Замечание 7.2. Для геометрической сходимости также достаточно $\tau = 1$ и $\kappa < 1 - \frac{1}{2\delta}$ (что более выгодно при $\delta > 3/4$). Кроме того, $\tau = 1$ всегда гарантирует отсутствие роста ошибки на маске.

При этом достаточно, чтобы свойство ограниченной изометрии выполнялось для разности матриц ранга r, возникающих на последовательных шагах алгоритма. Таким образом, хотя, в целом, выполнимость RIP (restricted isometry property, ограниченная изометрия) проверить тяжело, здесь она требуется лишь для матриц ранга не выше $2r$. К тому же не для всех, а только возникающих в процессе итераций. Да и в этом случае достаточно выполнения этого свойства для большинства матриц, ведь небольшое число неверных итераций вряд ли испортит сходимость в целом. Таким образом, RIP ограничение в такой слабой форме часто выполнено на практике (это ограничение, в частности, проверено экспериментально в [55]), что приводит к сходимости алгоритма.

Обобщим теперь результат теоремы 7.2 на случай приближенного проектора.

Теорема 7.3. *В условиях теоремы 7.2 можно заменить точный проектор на приближенный, $\forall X : \|X - P_r X\|_F^2 \leqslant (1+\varepsilon)\|X - X_r\|_F^2$. В этом случае при $1 - 2(1+\varepsilon) + \frac{1+\kappa}{1-\kappa}(1+\varepsilon) + \frac{\varepsilon}{\delta(1+\kappa)} = D < 1$ алгоритм, заданный* (7.15)*, сойдется до погрешности ϵ за $\left\lceil \frac{2\log_2\|A\|_F - 2\log_2\epsilon}{\log_2(1/D)} \right\rceil$ итераций.*

Доказательство здесь во многом повторяет доказательство теоремы 7.2.

Доказательство. Выразим погрешность на $s+1$-й итерации через погрешность на s-й итерации:

$$\begin{aligned}
&\left\|M \odot X_{(s+1)} - A\right\|_F^2 = \left\|M \odot \left(X_{(s+1)} - X_{(s)}\right) + \left(M \odot X_{(s)} - A\right)\right\|_F^2 \\
&= \left\|M \odot \left(X_{(s+1)} - X_{(s)}\right)\right\|_F^2 + \left\|M \odot X_{(s)} - A\right\|_F^2 + 2\operatorname{Re}\left(M \odot X_t - A\right) \odot \overline{M \odot \left(X_{(s+1)} - X_{(s)}\right)} \\
&= \left\|M \odot \left(X_{(s+1)} - X_{(s)}\right)\right\|_F^2 + \left\|M \odot X_{(s)} - A\right\|_F^2 + 2\operatorname{Re}\left(M \odot X_t - A\right) \odot \overline{\left(X_{(s+1)} - X_{(s)}\right)} \\
&\leqslant \delta(1+\kappa)\left\|X_{(s+1)} - X_{(s)}\right\|_F^2 + \left\|M \odot X_{(s)} - A\right\|_F^2 + 2\operatorname{Re}\left(M \odot X_t - A\right) \odot \overline{\left(X_{(s+1)} - X_{(s)}\right)},
\end{aligned} \tag{7.16}$$

где в последнем неравенстве мы воспользовались RIP.

Обозначим

$$f_s(Z) = 2\mathrm{Re}\,(M \odot X_t - A) \odot \overline{(Z - X_{(s)})} + \delta(1+\kappa)\left\|Z - X_{(s)}\right\|_F^2. \tag{7.17}$$

Обозначим матрицу на следующем шаге (до проектирования на множество матриц ранга r) через $Y_{(s+1)}$:

$$Y_{(s+1)} = X_{(s)} + \tau\left(A - M \odot X_{(t)}\right), \quad X_{(s+1)} = P_r Y_{(s+1)}.$$

Тогда разницу между s-м шагом и произвольной матрицей Z можно выразить как

$$Z - X_{(s)} = Z - Y_{(s)} - \tau\left(M \odot X_{(s)} - A\right) = Z - Y_{(s)} - \frac{1}{\delta(1+\kappa)}\left(M \odot X_{(s)} - A\right).$$

Подставим $Z - X_{(s)}$ в выражение (7.17) для f_s:

$$\begin{aligned}
f_s(Z) &= 2\mathrm{Re}\,(M \odot X_t - A) \odot \overline{(Z - Y_{(s+1)})} - \frac{2}{\delta(1+\kappa)}\mathrm{Re}\,(M \odot X_t - A) \odot \overline{(M \odot X_t - A)} \\
&\quad + \delta(1+\kappa)\left\|Z - Y_{(s+1)} - \frac{1}{\delta(1+\kappa)}\left(M \odot X_{(s)} - A\right)\right\|_F^2 \\
&= \delta(1+\kappa)\left\|Z - Y_{(s+1)}\right\|_F^2 - \frac{1}{\delta(1+\kappa)}\left\|M \odot X_{(s)} - A\right\|_F^2.
\end{aligned}$$

Норму $\left\|Z - Y_{(s+1)}\right\|_F^2$ для любого Z ранга r (в том числе для $Z = X$) можно оценить как

$$\left\|X - Y_{(s+1)}\right\|_F^2 \geqslant \left\|Y_{(s+1),r} - Y_{(s+1)}\right\|_F^2 \geqslant \frac{1}{1+\varepsilon}\left\|P_r Y_{(s+1)} - Y_{(s+1)}\right\|_F^2 = \frac{1}{1+\varepsilon}\left\|X_{(s+1)} - Y_{(s+1)}\right\|_F^2.$$

Таким образом,

$$\begin{aligned}
f_s\left(X_{(s+1)}\right) &= \delta(1+\kappa)\left\|X_{(s+1)} - Y_{(s+1)}\right\|_F^2 - \frac{1}{\delta(1+\kappa)}\left\|M \odot X_{(s)} - A\right\|_F^2 \\
&\leqslant \delta(1+\kappa)(1+\varepsilon)\left\|X - Y_{(s+1)}\right\|_F^2 - \frac{1}{\delta(1+\kappa)}\left\|M \odot X_{(s)} - A\right\|_F^2 \\
&= \delta(1+\kappa)(1+\varepsilon)\left\|X - X_{(s)} - \tau\left(A - M \odot X_{(s)}\right)\right\|_F^2 - \frac{1}{\delta(1+\kappa)}\left\|M \odot X_{(s)} - A\right\|_F^2 \\
&= \delta(1+\kappa)(1+\varepsilon)\left\|X - X_{(s)}\right\|_F^2 + \frac{\varepsilon}{\delta(1+\kappa)}\left\|M \odot X_{(s)} - A\right\|_F^2 \\
&\quad - 2(1+\varepsilon)\,\mathrm{Re}\left(X - X_{(s)}\right) \odot \overline{\left(M \odot X_{(s)} - A\right)} \\
&= \delta(1+\kappa)(1+\varepsilon)\left\|X - X_{(s)}\right\|_F^2 + \frac{\varepsilon}{\delta(1+\kappa)}\left\|M \odot X_{(s)} - A\right\|_F^2 \\
&\quad - 2(1+\varepsilon)\,\mathrm{Re}\left(M \odot X - M \odot X_{(s)}\right) \odot \overline{\left(A - M \odot X_{(s)}\right)} \\
&\leqslant \frac{1+\kappa}{1-\kappa}(1+\varepsilon)\left\|M \odot X - M \odot X_{(s)}\right\|_F^2 + \frac{\varepsilon}{\delta(1+\kappa)}\left\|M \odot X_{(s)} - A\right\|_F^2 \\
&\quad - 2(1+\varepsilon)\,\mathrm{Re}\left(M \odot X - M \odot X_{(s)}\right) \odot \overline{\left(A - M \odot X_{(s)}\right)}.
\end{aligned}$$

Используя тот факт, что $M \odot X = A$ для точного решения X, получаем

$$\begin{aligned} f_s\left(X_{(s+1)}\right) \leqslant \frac{1+\kappa}{1-\kappa}(1+\varepsilon)\left\|M \odot X_{(s)} - A\right\|_F^2 + \frac{\varepsilon}{\delta(1+\kappa)}\left\|M \odot X_{(s)} - A\right\|_F^2 \\ - 2(1+\varepsilon)\left\|M \odot X_{(s)} - A\right\|_F^2. \end{aligned} \tag{7.18}$$

Подставляя оценку (7.18) функции f_s (7.17) в правую часть (7.16), получаем

$$\begin{aligned} \left\|M \odot X_{(s+1)} - A\right\|_F^2 &\leqslant \left\|M \odot X_{(s)} - A\right\|_F^2 + f_s\left(X_{(s+1)}\right) \\ &\leqslant \left(1 - 2(1+\varepsilon) + \frac{1+\kappa}{1-\kappa}(1+\varepsilon) + \frac{\varepsilon}{\delta(1+\kappa)}\right)\left\|M \odot X_{(s)} - A\right\|_F^2 \\ &= D\left\|M \odot X_{(s)} - A\right\|_F^2. \end{aligned}$$

То есть погрешность убывает со скоростью не меньше геометрической прогрессии с параметром D (отсюда требование $D < 1$), а всего шагов до достижения погрешности ϵ потребуется не более $\left\lceil \frac{\log\left(\left\|X_{(0)}\right\|_F^2/\epsilon^2\right)}{\log D^{-1}} \right\rceil = \left\lceil \frac{\log\left(\|A\|_F^2/\epsilon^2\right)}{\log D^{-1}} \right\rceil$. □

Замечание 7.3. RIP также автоматически гарантирует близость полученного решения к точному, поскольку на финальном шаге s

$$\left\|X_{(s)} - X\right\|_F^2 \leqslant \frac{1}{\delta(1-\kappa)}\left\|M \odot X_{(s)} - A\right\|_F^2 \leqslant \frac{\epsilon^2}{\delta(1-\kappa)}.$$

Условие $D < 1$ означает, что для сходимости должно выполняться соотношение

$$\frac{1+\kappa}{1-\kappa} - 1 + O(\varepsilon) + O(\varepsilon/\delta) < 1.$$

Если κ – некоторая постоянная, меньшая $1/3$, то получаем условие

$$\varepsilon = O(\delta).$$

Согласно теореме 3.4, для крестовой аппроксимации на основе подматрицы локально максимального проективного объема следует в среднем ожидать погрешность $\varepsilon = \frac{r}{n-r+1}$. Отсюда получаем, что для построения аппроксимации нам понадобится $n = O(r/\varepsilon)$ строк и столбцов, что в случае использования алгоритма 4.10 потребует $O\left(Nr^2/\varepsilon + r^3/\varepsilon^3\right)$ операций.

К сожалению, напрямую применить этот алгоритм эффективно не представляется возможным, так как алгоритм maxvol ранга r не позволит найти оптимальные строки и столбцы, если матрица является сильно разреженной. В частности, подматрица $\hat{A} \in \mathbb{C}^{r\times r}$ не будет меняться, строки и столбцы будут оставаться теми же самыми, а потому ошибка может неконтролируемо расти в других строках и столбцах, поскольку информация в них полностью игнорируется. В связи с этим предлагается использовать maxvol для размера $2r$, а затем сокращать ранг до r с

помощью сокращенного сингулярного разложения, как это было предложено в конце раздела 1.1. В этом случае крестовая аппроксимация будет иметь вид

$$CGR = \left(C\hat{A}_{2r}^{+} R \right)_r, \; C \in \mathbb{N} \times \ltimes, \; \hat{A} \in \mathbb{C}^{n \times n}, \; R \in \mathbb{C}^{n \times N}.$$

Численные эксперименты на случайных матрицах различного размера и с различным произведением сингулярных чисел показали, что для наилучшей скорости сходимости имеет смысл брать $n = 2r + \left\lceil 0{,}7\frac{r}{\delta} \right\rceil$ строк и столбцов. За время порядка 10 сингулярных разложений алгоритм ASVP позволяет восстановить матрицы размера $N = 1000$ почти с машинной точностью, за исключением случаев, когда $r/\varepsilon \sim N$, см. рисунок 7.6.

При малых $\delta \sim r/N$ крестовый метод будет набирать почти все столбцы и строки в матрице, а потому не дает существенного ускорения. С другой стороны, в случае $\delta \gg r/N$ (точнее, при $\delta = \omega\left(\sqrt{r/N}\right)$) такой подход оказывается асимптотически быстрее всех известных методов восстановления матриц. Как видно в таблице 7.9, предложенный крестовый Cross ASVP метод опережает самый быстрый (согласно обзору [149]) из известных методов восстановления матриц, LMaFit [145], при $\delta \gtrsim 3\sqrt{r/N}$. Также видна предсказанная асимптотика времени выполнения алгоритмов: для LMaFit наблюдается рост времени выполнения с увеличением δ, тогда как крестовый метод показывает асимптотику времени выполнения порядка δ^{-1}. Сравнение проводилось на персональном компьютере с процессором Intel Core i5-8265U с частотой 1.6GHz, оба кода взяты из соответствующих GitHub репозиториев.

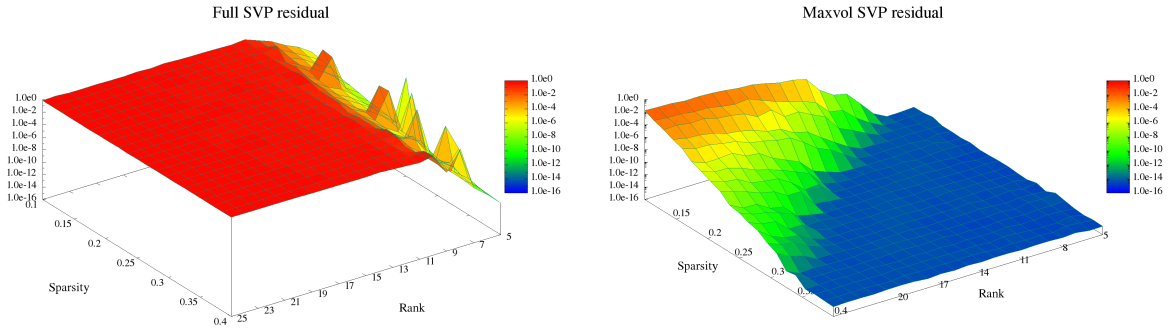
Таблица 7.9: Время достижения относительной погрешности решения по норме Фробениуса 10^{-4} для матрицы $A = UV \in \mathbb{R}^{10000 \times 10000}$, где факторы $U \in \mathbb{R}^{10000 \times 10}$ и $V \in \mathbb{R}^{10 \times 10000}$ – случайные Гауссовы матрицы.

δ	0,04	0,08	0,15	0,3
LMaFit, сек	5,28	9,27	12,8	19,2
Cross ASVP, сек	27,5	11,2	7,31	3,57

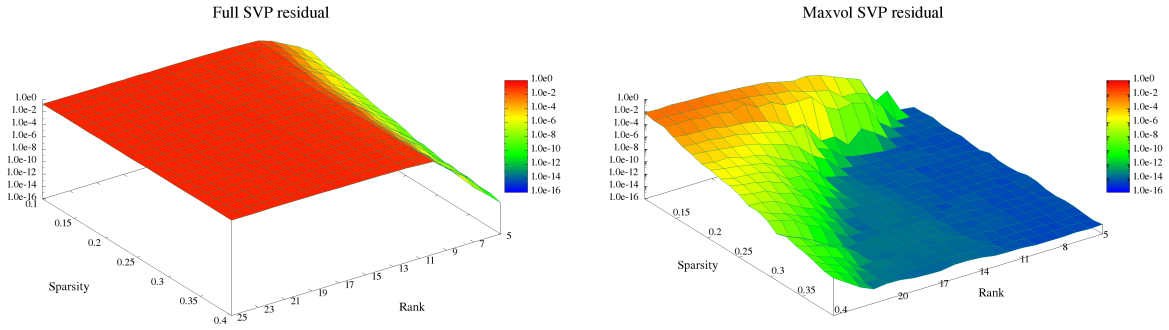
В файле $\mathrm{ExampleR.f90}$ в https://github.com/RodniO/Projective-volume-low-rank есть пример использования крестовой аппроксимации для восстановления матриц. Сам метод находится в модуле $\mathrm{ModRecon}$.

7.3. Неотрицательные аппроксимации матриц

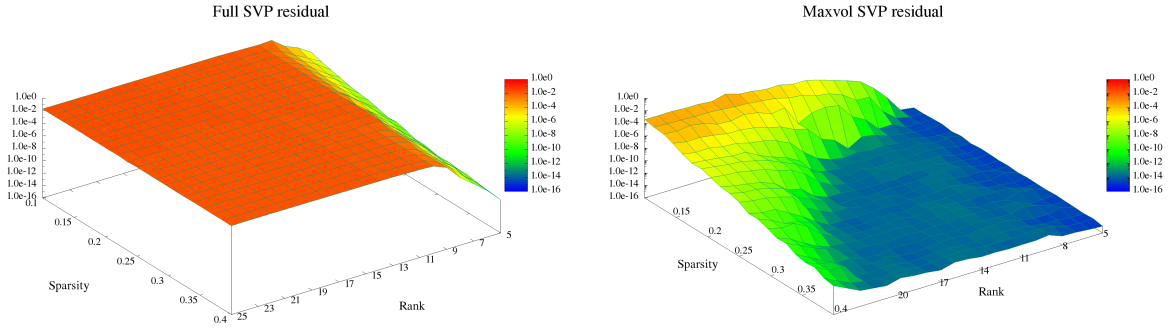
Наконец, рассмотрим задачу построения неотрицательной аппроксимации ранга r, $\tilde{A} \in \mathbb{R}^{M \times N}$, $\operatorname{rank} \tilde{A} = r$, $\tilde{A} \geqslant 0$ произвольной матрицы $A \in \mathbb{R}^{M \times N}$. Требование неотрицательности данных может возникать, например, при обработке изображений и аудио, а также при решении физических задач, требующих неотрицательности решения. В частности, при решении уравнений



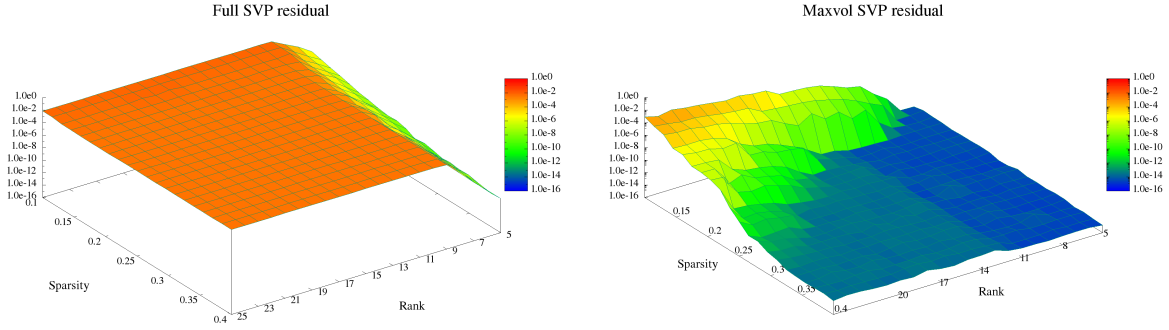
(a) $\sigma_k\ (X) = 1$.



(b) $\sigma_k\ (X) = 1/k$.



(c) $\sigma_k\ (X) = 1/k^2$.



(d) $\sigma_k\ (X) = 2^{-k}$.

Рис. 7.6: Слева – относительная погрешность (на маске) после SVP с точным проектором P_r, 50 итераций ($t = 25$ секунд). Справа – относительная погрешность (на маске) после ASVP с помощью крестового разложения за $t = 5$ секунд. Под каждой парой графиков указано распределение сингулярных чисел искомой матрицы $X \in \mathbb{R}^{1000\times1000}$ ранга Rank. Sparsity – плотность известных элементов.

Смолуховского наличие отрицательных компонент решения или матрицы коагуляции приводит к неустойчивости [150]. При этом использование алгоритмов построения неотрицательных аппроксимаций вида $\tilde{A} = UV$, $U \in \mathbb{R}^{M\times r}$, $V \in \mathbb{R}^{r\times N}$, где требуется неотрицательность факторов $U, V \geqslant 0$ [151] требует слишком высоких вычислительных затрат. С другой стороны, неотрицательной аппроксимации с $\tilde{A} \geqslant 0$ можно получить за разумное время с помощью метода переменных проекций [152].

Как и раньше, будем обозначать через A_r приближение ранга r произвольной матрицы A на основе сокращенного сингулярного разложения. Пусть $A_{(0)} = A_r$, а $A_{(s)}$ – аппроксимация на s-м шаге метода. Пусть $P_{\geqslant 0}$ – ортогональный проектор на множество неотрицательных матриц. Тогда метод переменных проекций можно записать в следующем виде:

$$A_{(s+1)} = \left(P_{\geqslant 0} A_{(s)}\right)_r . \tag{7.19}$$

В случае, если начальная матрица A близка к некасательной точке пересечения многообразий неотрицательных матриц и матриц ранга r, такой алгоритм сходится [148]. Однако, на практике он требует большого числа итераций, а построение точной проекции ранга r с помощью сингулярного разложения требует существенных вычислительных затрат. В связи с этим в [57] было предложено использовать различные приближенные методы проекции на множество матриц ранга r и было численно показано, что они также приводят к сходимости метода вида (7.19).

Здесь предлагается использовать крестовое разложение в качестве приближенного проектора на множество матриц ранга r. Предлагаемый метод построения крестовых аппроксимаций уже был описан в предыдущем подразделе. А именно, в качестве начального приближения предлагается использовать матрицу вида

$$A_{(0)} = \left(C\hat{A}_k^{+}R\right)_r ,$$

где $C \in \mathbb{R}^{M\times n}$ – столбцы матрицы A, $R \in \mathbb{R}^{n\times N}$ – её строки, $\hat{A}_k \in \mathbb{R}^{n\times n}$ – приближение ранга k с помощью сокращенного сингулярного разложения подматрицы $\hat{A}$ на пересечении столбцов C и строк R. Для достижения высокого качества аппроксимации выбирается подматрица $\hat{A}$ большого k-проективного объема. При этом, как показывают численные эксперименты, достаточно использовать $n \gtrsim k \gtrsim r$ строк и столбцов, что приводит к общей сложности алгоритма $O\left((M+N)\,r^2\right)$ (см. подраздел 4.5). Это наблюдение согласуется с недавними оценками точности метода переменных проекций [153], где показано, что требуемая относительная точность приближенного проектирования обычно порядка константы.

Чтобы компенсировать возникающую при таком приближении погрешность аппроксимации предлагается заменить проектор $P_{\geqslant 0}$ на $P_{\geqslant \varepsilon}$ – ортогональный проектор на множество матриц с элементами не меньше ε. Значение ε при этом может быть оценено исходя из средней величины

изменения элементов на первом шаге (на котором предлагается использовать $P_{\geqslant 0}$):

$$\varepsilon = \frac{\left\| A_{(1)} - A_{(0)} \right\|_F}{\sqrt{MN}}.$$

Использование проектора $P_{\geqslant \varepsilon}$ позволяет сократить число итераций метода в десятки раз, не внося при этом существенной погрешности в точность полученной в итоге неотрицательной аппроксимации, что было проверено на примерах из [57].

Один шаг метода переменных проекций с использованием крестовой аппроксимации можно записать следующим образом:

1. С помощью алгоритма maxvol ищется $r \times r$ доминантная подматрица в матрице $A_{(s)}$.
2. К ней добавляются $n - r$ случайных строк, в них ищется подматрица $n \times k$ большого объема с помощью выбора ведущих столбцов (алгоритм 4.1).
3. К k найденным столбцам добавляются $n - k$ случайных столбцов, их строки также переставляются с помощью выбора ведущих столбцов. На пересечении полученных n строк и столбцов будет подматрица $\hat{A}_{(s)} \in \mathbb{R}^{n \times n}$.
4. Строится крестовое приближение ранга r:

$$A_{(s+1)} = \left(P_{\geqslant \varepsilon} C_{(s)} \left(P_{\geqslant \varepsilon} \hat{A}_{(s)} \right)_k^+ P_{\geqslant \varepsilon} R_{(s)} \right)_r,$$

   где $C_{(s)} \in \mathbb{R}^{M \times n}$ и $R_{(s)} \in \mathbb{R}^{n \times N}$ – столбцы и строки, соответствующие подматрице $\hat{A}_{(s)} \in \mathbb{R}^{n \times n}$.

Сходимость метода можно проверять по наличию отрицательных элементов в C и R. В случае их отсутствия можно проверить всю матрицу $A_{(s+1)}$ или сделать еще несколько дополнительных шагов, чтобы с большей вероятностью гарантировать отсутствие отрицательных элементов построенного приближения.

В качестве примера матрицы A рассмотрим численное решение двухкомпонентных уравнений Смолуховского (автор выражает благодарность Сергею Александровичу Матвееву за предоставленные данные). Само решение и его сингулярные числа показаны на рисунке 7.7.

Его решения для рангов 10, 20 и 50 показаны на рисунке 7.8.

В таблице 7.10 сложность и точность крестового метода сравнивается с решением, полученным путем точного проектирования с использованием сингулярного разложения. Видно, что предложенный метод обладает существенно меньшей вычислительной сложностью и почти не теряет в точности, несмотря на использование $\varepsilon > 0$.

В файле $\mathrm{ExampleP.f90}$ в https://github.com/RodniO/Projective-volume-low-rank есть пример построения неотрицательной аппроксимации для $A_{ij} = e^{-0{,}01\,ij}$. Сам метод находится в модуле $\mathrm{ModAppr}$.

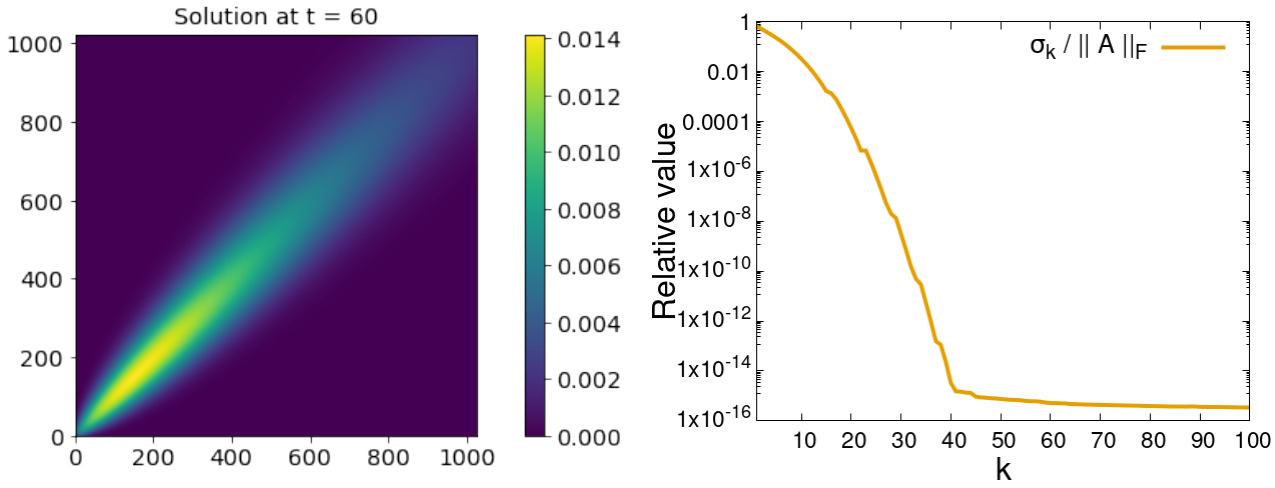
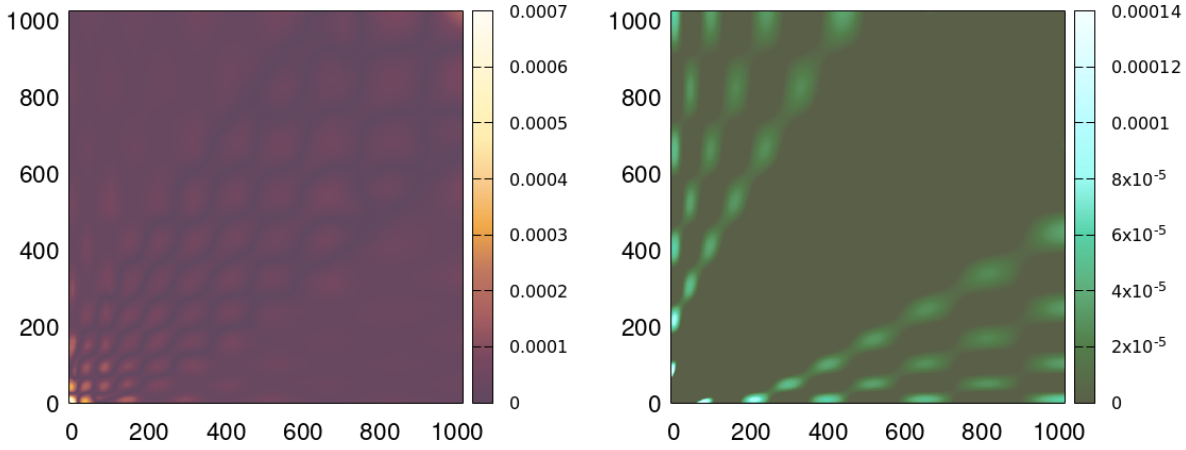


Рис. 7.7: Слева – приближаемое решение двухкомпонентных уравнений Смолуховского. Справа – его сингулярные числа.

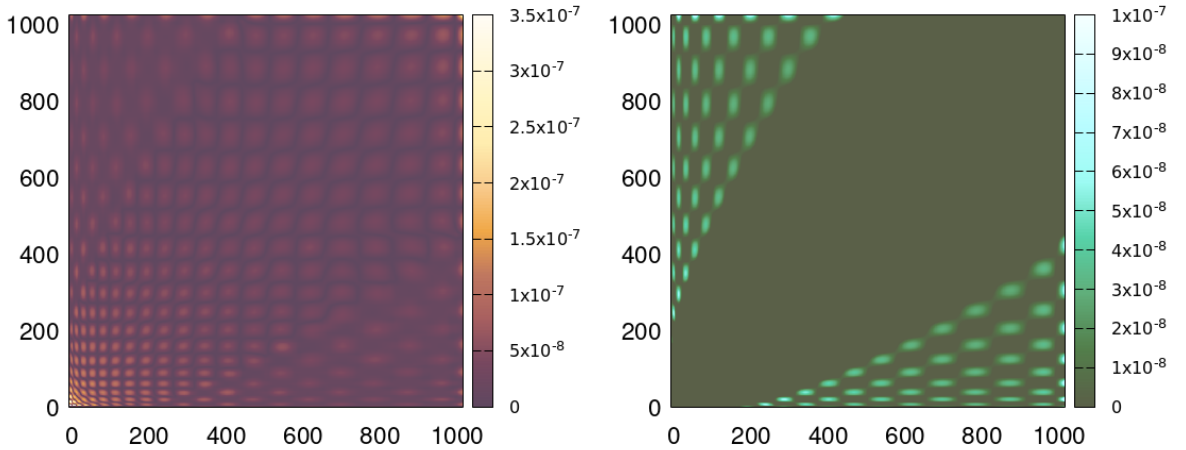
Таблица 7.10: Сравнение вычислительной сложности и погрешности построения неотрицательных аппроксимаций с помощью точного и приближенного проектирования. $N = 1024$, $r = 10$, элементы матрицы A и её сингулярные числа показаны на рисунке 7.7. Число итераций для точного проектирования соответствует достижению относительной погрешности отрицательных элементов $\frac{\|P_{\leqslant 0} A_{(s)}\|_F}{\|A_{(s)}\|_F} < 10^{-9}$.

Метод	Flop на итерацию	Итераций	$\frac{\|A - A_{(s)}\|_F}{\|A\|_F}$	$\frac{\|A - A_{(s)}\|_C}{\|A\|_C}$
$\tilde{A} = U\Sigma_r V$, $\varepsilon = 0$	$2{,}3 \cdot 10^{10}$	≈ 700	$2{,}70 \cdot 10^{-2}$	$1{,}48 \cdot 10^{-1}$
$\tilde{A} = CGR$, $\varepsilon > 0$	$6{,}7 \cdot 10^{6}$	15	$2{,}91 \cdot 10^{-2}$	$1{,}49 \cdot 10^{-1}$

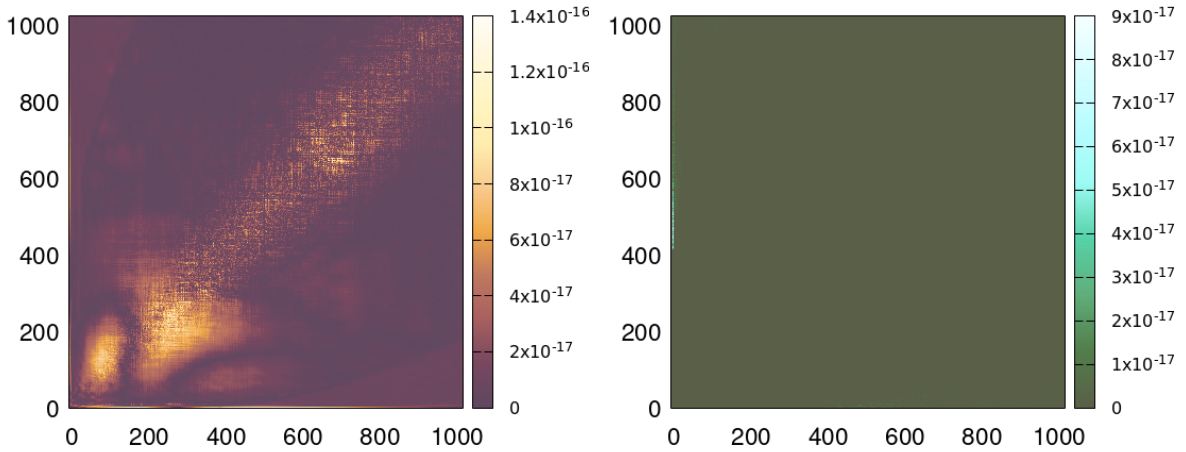
Примером, где реализация быстрого крестового метода не помогла, является использование переменных проекций для построения аппроксимации по норме Чебышева [153]. Это может быть связано с отсутствием теоретических гарантий сходимости при использовании приближенных методов проектирования с гарантиями относительной точности по норме Фробениуса. Даже использование точного проектора (сингулярного разложения) приводит (на единичных матрицах) к погрешностям в разы выше (с коэффициентом порядка $\log^2 N$ при фиксированном r вместо $\log N$ при больших рангах или $\sqrt{\log N}$ при малых), чем метод из [154], быстрая версия которого показывает на практике сложность $O\left(N^2 r^2\right)$ и погрешность, близкую к полученным нижним оценкам.



(a) Ранг 10. Время аппроксимации 0,15 сек.



(b) Ранг 20. Время аппроксимации 0,3 сек.



(c) Ранг 50. Время аппроксимации 0,75 сек.

Рис. 7.8: Слева – итоговая относительная (к норме Фробениуса всей матрицы) погрешность. Справа – начальная погрешность отрицательных элементов. Для $r = 10$ и $r = 20$ погрешность A_r совпадает с погрешностью начального CGR приближения в первых трех знаках. При ранге $r = 50$ (что выше численного ранга исходной матрицы) получаем погрешность порядка машинной (двойной) точности.

Заключение

В диссертационной работе получен ряд новых оценок точности столбцовых и крестовых аппроксимаций, а также построены алгоритмы, позволяющие достигнуть полученных оценок на практике за время, линейное по размерам матрицы. Основной результат работы состоит в общем подходе к разработке алгоритмов крестовой аппроксимации, анализу их эффективности, достигаемой точности и других свойств полученных малоранговых структур. Данный подход основан на использовании подматриц локально максимального объема и проективного объема. Связь между свойствами таких подматриц и свойствами столбцовых и крестовых аппроксимаций не только позволяет достичь и улучшить теоретические оценки точности, но также говорит о важности их поиска при построении соответствующих аппроксимаций на практике.

Полученные оценки снизу описывают наилучшие возможные гарантии точности, и позволяют определить, в каких случаях ни один метод крестовой или столбцовой аппроксимации не позволит достичь меньшей погрешности. Полученные оценки сверху позволяют оценить, какая точность является достижимой, и что точно можно гарантировать при использовании соответствующих видов аппроксимации. Наконец, особое внимание уделено вероятностным оценкам, которые позволяют определить, какую точность аппроксимации следует ожидать на практике. В частности, при построении аппроксимаций на основе подматриц локально максимального объема и проективного объема.

Теоретическая значимость полученных результатов заключается в использовании новых подходов к построению аппроксимаций, близости верхних и нижних оценок, а также доказательстве возможности достижения аппроксимаций по норме Фробениуса с произвольной относительной точностью (сколь угодно близкой к сингулярному разложению), используя небольшое число дополнительных строк и столбцов. Полученные теоретические результаты согласуются с результатами численных экспериментов.

Предложенные алгоритмы построения крестовых аппроксимаций обладают одновременно высокой наблюдаемой на практике точностью и очень низкой вычислительной сложностью. В связи с этим они имеют широкий круг применений и особенно важны в тех задачах, где аппроксимацию требуется строить несколько раз. Примерами таких задач служат, в частности, температурно-зависимые уравнения Смолуховского (где ядро меняется со временем), задачи восстановление матриц и построение неотрицательных аппроксимаций с помощью переменных проекций. Успешное применение разработанных алгоритмов в вышеуказанных задачах говорит об их высокой эффективности в различных условиях. Возможность построения аппроксимаций за меньшее время, чем требуется для вычисления приближаемой матрицы целиком, позволяет строить и обновлять малоранговую структуру быстрее, чем прямое использование полного набора данных, что приводит к ускорению различных алгоритмов даже в тех случаях, когда данные

постоянно меняются.

Список сокращений и условных обозначений

Сокращения

RRQR – выявляющее ранг QR разложение.

RRLU – выявляющее ранг LU разложение.

Символы

C_n^r – число сочетаний из n по r.

$k = \overline{1, n}$ обозначает, что k принимает целые значения начиная с 1 и заканчивая n.

$O\left(f\left(N, \ldots\right)\right)$ – «O» большое. В контексте определенного алгоритма означает, что число элементарных операций (сложений, умножений, сравнений и т.д.) ограничено $Cf\left(N, \ldots\right)$ для некоторого $C > 0$.

$g(x) = O\left(f\left(x\right)\right)$ – «O» большое для функции g. Означает, что $g(x) \leqslant Cf(x)$ для всех достаточно больших (или достаточно малых, в зависимости от контекста) значений x.

$g(x) = o\left(f\left(x\right)\right)$ – «o» малое для функции g. Означает $\lim\limits_x \frac{g(x)}{f(x)} = 0$. Предельное значение x зависит от контекста.

$g(x) = \omega\left(f\left(x\right)\right)$ – «омега» малое для функции g. Означает $\lim\limits_x \frac{g(x)}{f(x)} = +\infty$. Предельное значение x зависит от контекста.

$\mathbb{R}$ – множество всех действительных чисел.

$\mathbb{C}$ – множество всех комплексных чисел.

Векторы

Векторы обозначаются стандартным шрифтом. По умолчанию любой вектор есть вектор-столбец.

x_i – i-й элемент вектора x.

$x_{i:j}$ – вектор, состоящий из элементов x начиная с i-го и заканчивая j-м.

x^* – вектор-строка с комплексно-сопряженными элементами $x_i^* = \bar{x}_i$.

Матрицы

Матрицы обозначаются большими буквами.

I – единичная матрица.

A_{ij} – элемент матрицы A на пересечении строки i и столбца j.

$A_{i,:}$ – вся i-я строка матрицы A.

$A_{:,j}$ – весь j-й столбец матрицы A (иногда используется обозначение A_j).

$A_{\mathcal{I},:}$ – подматрица строк с индексами $i \in \mathcal{I}$.

$A_{:,\mathcal{J}}$ – подматрица столбцов с индексами $j \in \mathcal{J}$ (иногда используется обозначение $A_{\mathcal{J}}$).

$A_{\mathcal{I},\mathcal{J}}$ – подматрица матрицы A на пересечении строк с индексами $i \in \mathcal{I}$ и столбцов с индексами $j \in \mathcal{J}$.

$\hat{A}$ – некоторая подматрица матрицы A.

$A \subset B$ – A является подматрицей матрицы B.

A^T – транспонированная матрица.

$\overline{A}$ – матрица с комплексно-сопряженными элементами.

$A^* = \overline{A}^T$ – эрмитово-сопряженная матрица.

A^{-1} – обратная матрица.

A^+ – псевдообратная матрица (Мура-Пенроуза).

A_r – сокращенное до ранга r сингулярное разложение матрицы A.

$\det A$ – определитель.

$\operatorname{tr} A$ – след.

$\operatorname{rank} A$ – ранг.

$\mathcal{V}(A)$ – объем матрицы, см. определение 1.5.

$\mathcal{V}_r(A)$ – (r-)проективный объем матрицы, см. определение 1.5.

$\|A\|$ – произвольная матричная норма. $\|A\|_2$ – спектральная норма. $\|A\|_F$ – норма Фробениуса. $\|A\|_C = \max\limits_{i,j} \left|A_{ij}\right|$ – норма Чебышева.

Использование сразу нескольких обозначений, например, $\|A\|_{2,F}$ говорит о том, что соответствующее выражение верно при одновременной подстановке как спектральной нормы, так и нормы Фробениуса (но одной и той же для всех норм внутри данного выражения).

Операторы

$x := y$ – присвоение значения переменной.

$C = A \odot B$ – поэлементное произведение, $C_{ij} = A_{ij} B_{ij}$.

Публикации автора по теме диссертации

Научные статьи, опубликованные в журналах Scopus, WoS, RSCI, а также в изданиях, рекомендованных для защиты в диссертационном совете МГУ по специальности 1.1.6 — «Вычислительная математика»

A.1 Osinsky A.I. Low-rank Monte Carlo for Smoluchowski-class equations // Journal of Computational Physics. — 2024. — V. 506. — P. 112942. (WoS, JIF impact factor: 3.8) [2.1 п. л.]

A.2 Osinsky A. Volume-based subset selection // Numerical Linear Algebra with Applications. — 2024. — V. 31, no. 2. — P. e2525. (WoS, JIF impact factor: 1.8) [0.9 п. л.]

A.3 Осинский А.И. Нижние оценки точности столбцовых аппроксимаций матриц // Журнал вычислительной математики и математической физики. — 2023. — Т. 63, № 11. — С. 1816. (RSCI, двухлетний импакт-фактор РИНЦ: 1.118) [0.1 п. л.] Перевод:
Osinsky A. Lower Bounds for Column Matrix Approximations // Computational Mathematics and Mathematical Physics. — 2023. — V. 63, no. 11. — P. 2024-2037. (WoS, JIF impact factor: 0.7) [1.3 п. л.]

A.4 Osinsky A. Polynomial time ρ-locally maximum volume search // Calcolo. — 2023. — V. 60, no. 42. (WoS, JIF impact factor: 1.4) [1.9 п. л.]

A.5 Osinsky A.I., Brilliantov N.V. Exact solutions of temperature-dependent Smoluchowski equations // Journal of Physics A: Mathematical and Theoretical. — 2022. — V. 55, no. 42. — P. 425003. (WoS, JIF impact factor: 2) [1.5/1.3 п. л.]

A.6 Kalinov A., Osinsky A.I., Matveev S.A., Otieno W., Brilliantov N.V. Direct simulation Monte Carlo for new regimes in aggregation-fragmentation kinetics // Journal of Computational Physics. — 2022. — V. 467. — P. 111439. (WoS, JIF impact factor: 3.8) [1.4/1.1 п. л.]

A.7 Osinsky, A.I., Brilliantov, N.V. Anomalous aggregation regimes of temperature-dependent Smoluchowski equations // Physical Review E. — 2022. — V. 105, no. 3. — P. 034119. (WoS, JIF impact factor: 2.2) [0.6/0.5 п. л.]

A.8 Лебедева О.С., Осинский А.И., Петров С.В. Приближенные алгоритмы малоранговой аппроксимации в задаче восполнения матрицы на случайном шаблоне // Журнал вычислительной математики и математической физики. — 2021. — Т. 61, № 5. — С. 827–844. (RSCI, двухлетний импакт-фактор РИНЦ: 1.118) [1.3/1.0 п. л.] Перевод:
Lebedeva O.S., Osinsky A.I., Petrov S.V. Low-Rank Approximation Algorithms for Matrix Completion with Random Sampling // Computational Mathematics and Mathematical Physics. — 2021. — V. 61, no. 5. — P. 799–815. (WoS, JIF impact factor: 0.7) [1.3/1.0 п. л.]

A.9 Замарашкин Н.Л., Осинский А.И. О точности крестовых и столбцовых малоранговых maxvol-приближений в среднем // Журнал вычислительной математики и математической

физики. — 2021. — Т. 61, № 5. — С. 813–826. (RSCI, двухлетний импакт-фактор РИНЦ: 1.118) [1.0/0.9 п. л.] Перевод:

Zamarashkin N.L., Osinsky A.I. On the Accuracy of Cross and Column Low-Rank Maxvol Approximations in Average // Computational Mathematics and Mathematical Physics. — 2021. — V. 61, no. 5. — P. 786–798. (WoS, JIF impact factor: 0.7) [1.0/0.9 п. л.]

A.10 Osinsky A.I. Low-rank method for fast solution of generalized Smoluchowski equations // Journal of Computational Physics. — 2020. — Vol. 422. — P. 109764. (WoS, JIF impact factor: 3.8) [1.1 п. л.]

A.11 Zheltkov D.A., Osinsky A.I. Global Optimization Algorithms Using Tensor Trains. In: Lirkov, I., Margenov, S. (eds) Large-Scale Scientific Computing. LSSC 2019. Lecture Notes in Computer Science. — 2020. — V. 11958. Springer, Cham. (WoS, JIF impact factor: 0.4) [0.4/0.35 п. л.]

A.12 Осинский А.И., Оценки аппроксимации тензорных поездов по норме Чебышёва // Журнал вычислительной математики и математической физики. — 2019. — Т. 59, № 2. — С. 211–216. (RSCI, двухлетний импакт-фактор РИНЦ: 1.118) [0.4 п. л.] Перевод:

Osinsky A.I. Tensor Trains Approximation Estimates in the Chebyshev Norm // Computational Mathematics and Mathematical Physics. — 2019. — V. 59, no. 2. — P. 201–206. (WoS, JIF impact factor: 0.7) [0.4 п. л.]

A.13 Замарашкин Н.Л., Осинский А.И. О существовании близкой к оптимальной скелетной аппроксимации матрицы во фробениусовой норме // Доклады Академии наук. — 2018. — Т. 479, № 5. — С. 489–492. (RSCI, двухлетний импакт-фактор РИНЦ: 0.859) [0.3/0.25 п. л.] Перевод:

Zamarashkin N.L., Osinsky A.I. On the Existence of a Nearly Optimal Skeleton Approximation of a Matrix in the Frobenius Norm // Doklady Mathematics. — 2018. — V. 97, no. 2. — P. 164–166. (WoS, JIF impact factor: 0.5) [0.3/0.25 п. л.]

A.14 Osinsky A.I., Zamarashkin N.L. Pseudo-skeleton approximations with better accuracy estimates // Linear Algebra and its Applications. – 2018. – V. 537, no. 4. – P. 221-249. (WoS, JIF impact factor: 1) [2.1/1.9 п. л.]

A.15 Замарашкин Н.Л., Осинский А.И. Новые оценки точности псевдоскелетных аппроксимаций матриц // Доклады академии наук. — 2016. — Т. 471, № 3. — С. 263-266. (RSCI, двухлетний импакт-фактор РИНЦ: 0.859) [0.3/0.25 п. л.] Перевод:

Zamarashkin N.L., Osinsky A.I. New accuracy estimates for pseudoskeleton approximations of matrices // Doklady Mathematics. — 2016. — V. 94, no. 3. — P. 643-645. (WoS, JIF impact factor: 0.5) [0.3/0.25 п. л.]

Список литературы

[1] Tyrtyshnikov E. E. Mosaic-Skeleton approximations // Calcolo. — 1996. — Vol. 33. — P. 47–57.

[2] Wilkinson J. H. Error Analysis of Direct Methods of Matrix Inversion // Journal of the ACM. — 1961. — Vol. 8, no. 3. — P. 281–330.

[3] Foster L. V. Gaussian Elimination with Partial Pivoting Can Fail in Practice // SIAM Journal on Matrix Analysis and Applications. — 1994. — Vol. 15, no. 4. — P. 1354–1362.

[4] Савостьянов Д. В. Мозаично-скелетонные аппроксимации : Квалификационная работа бакалавра / Д. В. Савостьянов ; ИВМ РАН. — М., 2001.

[5] Bebendorf M., Rjasanow S. Adaptive low-rank approximation of collocation matrices // Computing. — 2003. — Vol. 70. — P. 1–24.

[6] Foster L. V. The Growth Factor and Efficiency of Gaussian Elimination with Rook Pivoting // Journal of Computational and Applied Mathematics. — 1997. — Vol. 86, no. 1. — P. 177–194.

[7] Pan C.-T. On the existence and computation of rank revealing LU factorizations // Linear Algebra and its Applications. — 2000. — Vol. 316. — P. 199–222.

[8] Goreinov S. A., Tyrtyshnikov E. E., Zamarashkin N. L. A theory of pseudo-skeleton approximations // Linear Algebra and Its Applications. — 1997. — Vol. 261. — P. 1–21.

[9] Goreinov S. A., Tyrtyshnikov E. E. The maximal-volume concept in approximation by low-rank matrices // Contemporary Mathematics. — 2001. — Vol. 268. — P. 47–51.

[10] Горейнов С. А., Тыртышников Е. Е. Квазиоптимальность скелетного приближения матрицы в чебышевской норме // Доклады Академии наук. — 2011. — Т. 438, № 5. — С. 593–594.

[11] Goreinov S. A., Oseledets I. V., Savostyanov D. V. et al. How to find a good submatrix // Matrix Methods: Theory, Algorithms, Applications. — World Scientific Publishing, 2010. — P. 247–256.

[12] Sorensen D. C., Embree M. A DEIM Induced CUR Factorization // arXiv 1407.5516v2 (Submitted on 21 Jul 2014). — 2014.

[13] Boutsidis C., Woodruff D. P. Optimal CUR matrix decompositions // Proceedings of the 46th Annual ACM Symposium on Theory of Computing, ACM. — 2014. — P. 353–362.

[14] Anderson D. G., Gu M. An Efficient, Sparsity-Preserving, Online Algorithm for Low-Rank Approximation // Proceedings of the 34th International Conference on Machine Learning. — 2017.

[15] de Hoog F., Markus Hegland M. A note on error bounds for pseudo skeleton approximations of matrices // Linear Algebra and its Applications. — 2023. — Vol. 669. — P. 102–117.

[16] Bebendorf M. Efficient inversion of the Galerkin matrix of general second-order elliptic operators with nonsmooth coefficients // Mathematics of Computation. — 2004. — Vol. 74. — P. 1179–1199.

[17] Börm S. Construction of Data-Sparse $\mathcal{H}^2$-Matrices by Hierarchical Compression // SIAM Journal on Scientific Computing. — 2009. — Vol. 31, no. 3. — P. 1820–1839.

[18] Aparinov A., Setukha A., Stavtsev S. Low Rank Methods of Approximation in an Electromagnetic Problem // Lobachevskii Journal of Mathematics. — 2019. — Vol. 40. — P. 1771–1780.

[19] Mach T., Reichel L., Van Barel M., Vandebril R. Adaptive cross approximation for ill-posed problems // Journal of Computational and Applied Mathematics. — 2016. — Vol. 303. — P. 206–217.

[20] Матвеев С. А., Тыртышников Е. Е., Смирнов А. П., Бриллиантов Н. В. Быстрый метод решения уравнений агрегационно-фрагментационной кинетики типа уравнений Смолуховского // Вычислительные методы и программирование. — 2014. — Т. 15, № 1. — С. 1–8.

[21] Gidisu P. Y., Hochstenbach M. E. A Generalized CUR Decomposition for Matrix Pairs // SIAM Journal on Mathematics of Data Science. — 2022. — Vol. 4, no. 1. — P. 386–409.

[22] Arioli M., Duff I. S. Preconditioning Linear Least-Squares Problems by Identifying a Basis Matrix // SIAM Journal on Scientific Computing. — 2015. — Vol. 37, no. 5. — P. S544–S561.

[23] Shcherbakova E., Tyrtyshnikov E. Nonnegative Tensor Train Factorizations and Some Applications // Large-Scale Scientific Computing / Ed. by I. Lirkov, S. Margenov. — Cham : Springer International Publishing, 2020. — P. 156–164.

[24] Espig M., Hackbusch W., Litvinenko A. et al. Iterative algorithms for the post-processing of high-dimensional data // Journal of Computational Physics. — 2020. — Vol. 410. — P. 109396.

[25] Ahmadi-Asl S., Caiafa C. F., Cichocki A. et al. Cross Tensor Approximation Methods for Compression and Dimensionality Reduction // IEEE Access. — 2021. — Vol. 9. — P. 150809–150838.

[26] Oferkin I. V., Zheltkov D. A., Tyrtyshnikov E. E. et al. Evaluation of the docking algorithm based on tensor train global pptimization // Bulletin of the South Ural State University. Series "Mathematical Modelling, Programming and Computer Software. — 2015. — Vol. 8, no. 4. — P. 83–99.

[27] Townsend A., Trefethen L. N. An Extension of Chebfun to Two Dimensions // SIAM Journal on Scientific Computing. — 2013. — Vol. 35, no. 6. — P. C495–C518.

[28] Liu N. N., Meng X., Liu C., Yang Q. Wisdom of the Better Few: Cold Start Recommendation via Representative Based Rating Elicitation // Proceedings of the Fifth ACM Conference on Recommender Systems. — RecSys '11. — New York, NY, USA : Association for Computing Machinery, 2011. — P. 37–44.

[29] Çivril A., Magdon-Ismail M. On selecting a maximum volume sub-matrix of a matrix and related problems // Theoretical Computer Science. — 2009. — Vol. 410, no. 47-49. — P. 4801–4811.

[30] Lukas S., Jacek G. Implementation of an interior point method with basis preconditioning // Mathematical Programming Computation. — 2020. — Vol. 12. — P. 603–635.

[31] Brannick J., Cao F., Kahl K. et al. Optimal Interpolation and Compatible Relaxation in Classical Algebraic Multigrid // SIAM Journal on Scientific Computing. — 2018. — Vol. 40, no. 3. — P. A1473–A1493.

[32] Vanaret C., Seufert P., Schwientek J. et al. Two-phase approaches to optimal model-based design of experiments: how many experiments and which ones? // Computers & Chemical Engineering. — 2021. — Vol. 146. — P. 107218.

[33] Gubaev K., Podryabinkin E. V., Shapeev A. V. Machine learning of molecular properties: Locality and active learning // The Journal of Chemical Physics. — 2018. — Vol. 148, no. 24. — P. 241727.

[34] Yeh C.-Y., Chu T.-C., Chen C.-E., Yang C.-H. A Hardware-Scalable DSP Architecture for Beam Selection in mm-Wave MU-MIMO Systems // IEEE Transactions on Circuits and Systems I: Regular Papers. — 2018. — Vol. 65, no. 11. — P. 3918–3928.

[35] Fang B., Qian Z., Shao W., Zhong W. RAISE: A New Fast Transmit Antenna Selection Algorithm for Massive MIMO Systems // Wireless Personal Communications. — 2015. — 02. — Vol. 80. — P. 1147–1157.

[36] Горейнов С. А. О крестовой аппроксимации многоиндексного массива // Доклады Академии наук. — 2008. — Т. 420, № 4. — С. 439–441.

[37] Oseledets I. V., Tyrtyshnikov E. E. TT-cross approximation for multidimensional arrays // Linear Algebra and Its Applications. — 2010. — Vol. 432. — P. 70–88.

[38] Goreinov S. A., Oseledets I. V., Savostyanov D. V. Wedderburn Rank Reduction and Krylov Subspace Method for Tensor Approximation. Part 1: Tucker Case // SIAM Journal on Scientific Computing. — 2012. — Vol. 34, no. 1. — P. A1–A27.

[39] Gu M., Eisenstat S. C. Efficient algorithms for computing a strong rank-revealing qr factorization // SIAM Journal on Scientific Computing. — 1996. — Vol. 17, no. 4. — P. 848–869.

[40] Halko N., Martinsson P. G., Tropp J. A. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions // SIAM Review. — 2011. — Vol. 53, no. 2. — P. 217–288.

[41] Gu M., Miranian L. Strong rank revealing cholesky factorization // Electronic Transactions on Numerical Analysis. — 2004. — Vol. 17. — P. 76–92. — Access mode: https://etna.math.kent.edu/vol.17.2004/pp76-92.dir/pp76-92.pdf.

[42] Михалев А. Ю., Оселедец И. В. Прямоугольные подматрицы максимального объема и их вычисление // Доклады академии наук. — 2015. — Т. 462, № 1. — С. 19–20.

[43] Deshpande A., Rademacher L., Vempala S., Wang G. Matrix Approximation and Projective Clustering via Volume Sampling // Theory of Computing. — 2006. — Vol. 2. — P. 225–247.

[44] Avron H., Boutsidis C. Faster Subset Selection for Matrices and Applications // SIAM Journal on Matrix Analysis and Applications. — 2011. — Vol. 34, no. 4.

[45] Welch W. J. Algorithmic complexity: three NP- hard problems in computational statistics // Journal of Statistical Computation and Simulation. — 1982. — Vol. 15, no. 1. — P. 17–25.

[46] Guruswami V., Sinop A. K. Optimal Column-Based Low-Rank Matrix Reconstruction // Proceedings of the 2012 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). — 2012. — P. 1207–1214.

[47] Cortinovis A., Kressner D. Low-Rank Approximation in the Frobenius Norm by Column and Row Subset Selection // SIAM Journal on Matrix Analysis and Applications. — 2020. — Vol. 41, no. 4. — P. 1651–1673.

[48] Drineas P., Kannan R. Pass Efficient Algorithms for Approximating Large Matrices // Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms. — SODA '03. — USA : Society for Industrial and Applied Mathematics, 2003. — P. 223–232.

[49] Wang S., Zhang Z. Improving CUR Matrix Decomposition and the Nyström Approximation via Adaptive Sampling // The Journal of Machine Learning Research. — 2013. — Vol. 14, no. 1. — P. 2729–2769.

[50] Nikolov A. Randomized Rounding for the Largest Simplex Problem // Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC / Ed. by Rocco A. Servedio, Ronitt Rubinfeld. — USA : ACM, 2015. — P. 861–870.

[51] de Hoog F. R., Mattheij R. M. M. Subset selection for matrices // Linear Algebra and its Applications. — 2007. — Vol. 422, no. 2. — P. 349–359.

[52] Boutsidis C. Topics in matrix sampling algorithms : Ph. D. thesis / C. Boutsidis ; Rensselaer Polytechnic Institute. — 2011. — May.

[53] Tyrtyshnikov E. E. Incomplete Cross Approximation in the Mosaic-Skeleton Method // Computing. — 2000. — Vol. 64. — P. 367–380.

[54] Brilliantov N. V., Formella A., Pöschel T. Increasing temperature of cooling granular gases // Nature Communications. — 2018. — Vol. 9, no. 1. — P. 797.

[55] Jain P., Meka R., Dhillon I. Guaranteed Rank Minimization via Singular Value Projection // Advances in Neural Information Processing Systems / Ed. by J. Lafferty, C. Williams, J. Shawe-Taylor et al. — Vol. 23. — Curran Associates, Inc., 2010.

[56] Ma S., Goldfarb D., Chen L. Fixed point and Bregman iterative methods for matrix rank minimization // Mathematical Programming. — 2009. — Vol. 128. — P. 321–353.

[57] Sultonov A., Matveev S., Budzinskiy S. Low-rank nonnegative tensor approximation via alternating projections and sketching // Computational and Applied Mathematics. — 2023. — Vol. 42. — P. 68.

[58] Тыртышников Е. Е., Щербакова Е. М. Методы неотрицательной матричной фактоизации на основе крестовых малоранговых приближений // Журнал вычислительной математики и математической физики. — 2019. — Т. 59, № 8. — С. 1314–1330.

[59] Miranian L., Gu M. Strong rank revealing LU factorizations // Linear Algebra and its Applications. — 2003. — Vol. 367. — P. 1–16.

[60] Osinsky A. Rectangular maximum volume and projective volume search algorithms // arXiv 1809.02334 (Submitted on 7 Sep 2018). — 2018.

[61] Boutsidis C., Mahoney M.W., Drineas P. An Improved Approximation Algorithm for the Column Subset Selection Problem // Proceedings of the 2009 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). — 2009. — P. 968–977.

[62] Leviatan D., Temlyakov V.N. Simultaneous approximation by greedy algorithms // Advances in Computational Mathematics. — 2006. — Vol. 25. — P. 73–90.

[63] Chen J., Huo X. Theoretical Results on Sparse Representations of Multiple-Measurement Vectors // IEEE Transactions on Signal Processing. — 2006. — Vol. 54, no. 12. — P. 4634–4643.

[64] Farahat A. K., Ghodsi A., Kamel M. S. Efficient greedy feature selection for unsupervised learning // Knowledge and Information Systems. — 2012. — Vol. 35. — P. 285 – 310.

[65] Cotter S.F., Adler J., Rao B.D., Kreutz-Delgado K. Forward sequential algorithms for best basis selection // IEE Proceedings - Vision, Image and Signal Processing. — 1999. — Vol. 146. — P. 235–244(9).

[66] Drmač Z., Bujanović Z. On the Failure of Rank-Revealing QR Factorization Software – A Case Study // ACM Transactions on Mathematical Software. — 2008. — Vol. 35, no. 2.

[67] Altschuler J., Bhaskara A., Fu G. et al. Greedy column subset selection: New bounds and distributed algorithms // International Conference on Machine Learning. — 2016. — P. 2539–2548.

[68] Boutsidis C., Drineas P., Magdon-Ismail M. Near-Optimal Column-Based Matrix Reconstruction // SIAM Journal on Computing. — 2014. — Vol. 43, no. 2. — P. 687–717.

[69] Hamm K., Huang L. Perturbations of CUR Decompositions // SIAM Journal on Matrix Analysis and Applications. — 2021. — Vol. 42, no. 1. — P. 351–375.

[70] Oseledets I. V. Tensor-Train Decomposition // SIAM Journal on Scientific Computing. — 2011.

[71] Pinkus A. n-Widths in Approximation Theory. — Berlin, Heidelberg : Springer, 1985. — ISBN: 978-3-642-69896-5.

[72] Кашин Б. С. О поперечниках октаэдров // Успехи математических наук. — 1975. — Т. 30, № 4. — С. 251–252.

[73] Shannon C. E. Probability of error for optimal codes in a Gaussian channel // Bell System Technical Journal. — 1959. — Vol. 38. — P. 611–656.

[74] Welch L. Lower bounds on the maximum cross correlation of signals (Corresp.) // IEEE Transactions on Information Theory. — 1974. — Vol. 20, no. 3. — P. 397–399.

[75] Глускин Е. Д. Октаэдр плохо приближается случайными подпространствами // Функциональный анализ и его приложения. — 1986. — Т. 20, № 1. — С. 14–20.

[76] Todd M. J. Minimum-Volume Ellipsoids. — Philadelphia, PA : Society for Industrial and Applied Mathematics, 2016.

[77] Левенштейн В. И. Границы максимальной мощности кода с ограниченным модулем скалярного произведения // Доклады Академии наук СССР. — 1982. — Т. 263, № 6. — С. 1303–1308.

[78] Michalev A. Y., V. Oseledets I. Rectangular maximum-volume submatrices and their applications // Linear Algebra and its Applications. — 2018. — Vol. 538. — P. 187–211.

[79] Горейнов С. А. О крестовой аппроксимации многоиндексного массива // Доклады Академии наук. — 2008. — Т. 420, № 4. — С. 439–441.

[80] Deshpande A., Rademacher L. Efficient volume sampling for row/column subset selection // Proceedings of the 42th Annual ACM Symposium on Theory of Computing (STOC). — 2010.

[81] Shitov Y. Column subset selection is NP-complete // Linear Algebra and its Applications. — 2021. — Vol. 610. — P. 52–58.

[82] Çivril A. Column Subset Selection Problem is UG-Hard // Journal of Computer and System Sciences. — 2014. — Vol. 80, no. 4. — P. 849–859.

[83] Осинский А. И. Нелинейные малоранговые аппроксимации матриц, основанные на принципе максимального объема [Текст] : Квалификационная работа магистра : 01.03.04 / А. И. Осинский ; МФТИ. — Долгопрудный, 2018. — 106 с.

[84] Deshpande A., Vempala S. Adaptive sampling and fast low-rank matrix approximation // Approximation, randomization and combinatorial optimization. — 2006. — Vol. 4110, no. 3. — P. 292–303.

[85] Wang S., Luo L., Zhang Z. SPSD matrix approximation vis column selection: theories, algorithms, and extensions // Journal of Machine Learning Research (JMLR). — 2016. — Vol. 17. — P. 1–49. — Id/No 49. Access mode: jmlr.csail.mit.edu/papers/v17/14-199.html.

[86] Osinsky A. I. Probabilistic estimation of the rank 1 cross approximation accuracy // arXiv 1706.10285 (Submitted on 30 Jun 2017). — 2017.

[87] Laurent B., Massart P. Adaptive estimation of a quadratic functional by model selection // The Annals of Statistics. — 2000. — Vol. 28, no. 5. — P. 1302 – 1338.

[88] Kaas R., Dhaene J., Goovaerts M.J. Upper and Lower Bounds for Sums of Random Variables // Insurance Mathematics and Economics. — 2000. — Vol. 27, no. 2. — P. 151–168.

[89] Савостьянов Д. В. Быстрая полилинейная аппроксимация матриц и интегральные уравнения : дис. . . . канд. наук : 01.01.07 / Д. В. Савостьянов ; ИВМ РАН. — М., 2006. — 144 с.

[90] Kahan W. Numerical Linear Algebra // Canadian Mathematical Bulletin. — 1966. — Vol. 9, no. 5. — P. 757–801.

[91] Федоров В. В. Теория оптимального эксперимента. — М. : Наука, 1971. — С. 312.

[92] Miller A. J., Nam-Ky Nguyen N.-K. A Fedorov Exchange Algorithm for D-optimal Design // Applied statistics. — 1994. — Vol. 43, no. 4. — P. 669–677.

[93] Bebendorf M. Approximation of boundary element matrices // Numerische Mathematic. — 2000. — Vol. 86. — P. 565–589.

[94] Cortinovis A., Kressner D., Massei S. On maximum volume submatrices and cross approximation for symmetric semidefinite and diagonally dominant matrices // Linear Algebra and its Applications. — 2020. — Vol. 593. — P. 251–268.

[95] Wilkinson J. Error Analysis of Direct Methods of Matrix Inversion // Journal of the ACM. — 1961. — Vol. 8, no. 3. — P. 281–330.

[96] Massei S. Some algorithms formaximum volume and cross approximation of symmetric semidefinite matrices // BIT Numerical Mathematics. — 2022. — Vol. 62. — P. 195–220.

[97] Woodruff D., Yasuda T. New Subset Selection Algorithms for Low Rank Approximation: Offline and Online // arXiv 2304.09217 (Submitted on 18 Apr 2023). — 2023.

[98] Rudelson M., Vershynin R. Sampling from Large Matrices: An Approach through Geometric Functional Analysis // Journal of the ACM. — 2007. — Vol. 54, no. 4. — P. 21–es.

[99] Batson J. D., Spielman D. A., Srivastava N. Twice-Ramanujan Sparsifiers // Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing. — STOC '09. — 2009. — P. 255–262.

[100] Stewart G. W. Incremental condition calculation and column selection // Technical Report UMIACS-TR90-87. — College Park : Department of Computer Science and Institute for Advanced Computer Studies,University of Maryland, 1990.

[101] Madan V., Singh M., Tantipongpipat U., Xie W. Combinatorial Algorithms for Optimal Design // Proceedings of the Thirty-Second Conference on Learning Theory / Ed. by Alina Beygelzimer, Daniel Hsu. — Vol. 99 of Proceedings of Machine Learning Research. — PMLR, 2019. — P. 2210–2258.

[102] Nikolov A., Singh M. Maximizing Determinants under Partition Constraints // Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing. — STOC '16. — New York, NY, USA : Association for Computing Machinery, 2016. — P. 192–201.

[103] Çivril A., Magdon-Ismail M. Exponential Inapproximability of Selecting a Maximum Volume Sub-Matrix // Algorithmica. — 2013. — Vol. 65, no. 1. — P. 159–176.

[104] Silverman B. W., Titterington D. M. Minimum Covering Ellipses // SIAM Journal on Scientific and Statistical Computing. — 1980. — Vol. 1, no. 4. — P. 401–409.

[105] Glineur F. Pattern separation via ellipsoids and conic programming : Master's thesis / F. Glineur ; University of Mons, Belgium. — Faculté Polytechnique de Mons, 1998. — Aug.

[106] Eberly D. 3D Game Engine Design: A Practical Approach to Real-Time Computer Graphics. — CRC Press, 2006. — ISBN: 9781482267303.

[107] Kiefer J., Wolfowitz J. The Equivalence of Two Extremum Problems // Canadian Journal of Mathematics. — 1960. — Vol. 12. — P. 363–366.

[108] Zheltkov D., Tyrtyshnikov E. Global optimization based on TT-decomposition // Russian Journal of Numerical Analysis and Mathematical Modelling. — 2020. — Vol. 35, no. 4. — P. 247–261.

[109] Осинский А. И. Кинетика агрегации и фрагментации в неоднородных системах : дис. . . . канд. наук : 1.2.2 : защищена 28.09.22 : утв. 09.02.23 / А. И. Осинский ; Сколтех. — М., 2022. — 171 с.

[110] Poeschel T., Brilliantov N.V., Frommel C. Kinetics of prion growth // Biophysics Journal. — 2003. — Vol. 85, no. 6. — P. 3460–3474.

[111] Murthy C.R., Gao B., Tao A.R., Arya G. Dynamics of nanoparticle assembly from disjointed images of nanoparticle-polymer composites // Physical Review E. — 2016. — Vol. 93, no. 2. — P. 022501.

[112] Falkovich G., Fouxon A., Stepanov M.G. Acceleration of rain initiation by cloud turbulence // Nature. — 2002. — Vol. 419. — P. 151–154.

[113] Zereini F., Wiseman C. L. Urban Airborne Particulate Matter: Origin, Chemistry, Fate and Health Impacts. — Springer, 2011.

[114] Birnstiel, T., Dullemond, C. P., Brauer, F. Gas- and dust evolution in protoplanetary disks // Astronomy & Astrophysics. — 2010. — Vol. 513. — P. A79.

[115] Brilliantov N. V., Krapivsky P. L., Bodrova A. S. et al. Size distribution of particles in Saturn's rings from aggregation and fragmentation // Proceedings of the National Academy of Sciences. — 2015. — Vol. 112, no. 31. — P. 9536–9541.

[116] Meakin P. The growth of fractal aggregates // Time-Dependent Effects in Disordered Materials / Ed. by R. Pynn, T. Riste. — Boston, MA : Springer US, 1987. — Vol. 167. — P. 45–70.

[117] Garcia A. L., Alejandro L., van den Broeck C. et al. A Monte Carlo simulation of coagulation // Physica A: Statistical Mechanics and its Applications. — 1987. — Vol. 143, no. 3. — P. 535–546.

[118] Kotalczyk G., Kruis F. E. A Monte Carlo method for the simulation of coagulation and nucleation based on weighted particles and the concepts of stochastic resolution and merging // Journal of Computational Physics. — 2017. — Vol. 340. — P. 276–296.

[119] Zhao H., Maisels A., Matsoukas T., Zheng C. Analysis of four Monte Carlo methods for the solution of population balances in dispersed systems // Powder Technology. — 2007. — Vol. 173, no. 1. — P. 38–50.

[120] Matveev S. A., Zheltkov D. A., Tyrtyshnikov E. E., Smirnov A. P. Tensor train versus Monte Carlo for the multicomponent Smoluchowski coagulation equation // Journal of Computational Physics. — 2016. — Vol. 316. — P. 164–179.

[121] Matveev S. A., Krapivsky P. L., Smirnov A. P. et al. Oscillations in aggregation-shattering processes // Physical Review Letters. — 2017. — Vol. 119, no. 26. — P. 260601.

[122] Matveev S. A., Ampilogova N. V., Stadnichuk V. I. et al. Anderson acceleration method of finding steady-state particle size distribution for a wide class of aggregation–fragmentation models // Computer Physics Communications. — 2018. — Vol. 224. — P. 154–163.

[123] Krapivsky P.L., Redner S., Ben-Naim E.A. Kinetic View of Statistical Physics. — Cambridge University Press, 2010.

[124] Singh M., Kumar J., Bück A., Tsotsas E. A volume-consistent discrete formulation of aggregation population balance equations // Mathematical Methods in The Applied Sciences. — 2016. — Vol. 39. — P. 2275–2286.

[125] Intel Math Kernel Library. — Access mode: http://software.intel.com/en-us/intel-mkl (online; accessed: 13.06.2023).

[126] Д. А. Желтков Е. Е. Тыртышников. Параллельная реализация матричного крестового метода // Вычислительные методы и программирование. — 2015. — Т. 16, № 3. — С. 369–375.

[127] Zhao H., Zheng C., M.-H. Xu. Multi-monte-carlo method for general dynamic equation considering particle coagulation // Applied Mathematics and Mechanics. — 2005. — Vol. 26. — P. 953–962.

[128] Leyvraz F. Scaling theory and exactly solved models in the kinetics of irreversible aggregation // Physics Reports. — 2003. — Vol. 383, no. 2–3. — P. 95–212.

[129] Wei J. A Monte Carlo simulation for particle aggregation containing a sol–gel phase transition // Journal of Sol-Gel Science and Technology. — 2016. — Vol. 78, no. 2. — P. 270–278.

[130] Eibeck A., Wagner W. An Efficient Stochastic Algorithm for Studying Coagulation Dynamics and Gelation Phenomena // SIAM Journal on Scientific Computing. — 2000. — Vol. 22, no. 3. — P. 802–821.

[131] Sabelfeld K., Levykin A., Privalova T. A Fast Stratified Sampling Simulation of Coagulation Processes // Monte Carlo Methods and Applications. — 2007. — Vol. 13, no. 1. — P. 71–88.

[132] Sabelfeld K. K., Eremeev G. A hybrid kinetic-thermodynamic Monte Carlo model for simulation of homogeneous burst nucleation // Monte Carlo Methods and Applications. — 2018. — Vol. 24, no. 3. — P. 193–202.

[133] Netflix Prize. — Access mode: https://web.archive.org/web/20090919150646/http://www.netflixprize.com/index (online; accessed: 23.05.2023).

[134] Candès Emmanuel J., Eldar Yonina C., Strohmer Thomas, Voroninski Vladislav. Phase Retrieval via Matrix Completion // SIAM Review. — 2015. — Vol. 57, no. 2. — P. 225–251.

[135] Nguyen L. T., Kim J., Kim S., Shim B. Localization of IoT Networks via Low-Rank Matrix Completion // IEEE Transactions on Communications. — 2019. — Vol. 67, no. 8. — P. 5833–5847.

[136] Shen W., Dai L., Shim B. et al. Joint CSIT Acquisition Based on Low-Rank Matrix Completion for FDD Massive MIMO Systems // IEEE Communications Letters. — 2015. — Vol. 19, no. 12. — P. 2178–2181.

[137] Petrov S., Zamarashkin N. Matrix completion with sparse measurement errors // Calcolo. — 2023. — Vol. 60, no. 9.

[138] Candès E., Recht B. Exact Matrix Completion via Convex Optimization // Communications of the ACM. — 2012. — Vol. 55, no. 6. — P. 111–119.

[139] Cai J.-F., Candès E. J., Shen Z. A Singular Value Thresholding Algorithm for Matrix Completion // SIAM Journal on Optimization. — 2010. — Vol. 20, no. 4. — P. 1956–1982.

[140] Fornasier M., Rauhut H., Ward R. Low-rank Matrix Recovery via Iteratively Reweighted Least Squares Minimization // SIAM Journal on Optimization. — 2011. — Vol. 21, no. 4. — P. 1614–1640.

[141] Hu Y., Zhang D., Ye J. et al. Fast and Accurate Matrix Completion via Truncated Nuclear Norm Regularization // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2013. — Vol. 35, no. 9. — P. 2117–2130.

[142] Lee K., Bresler Y. ADMiRA: Atomic Decomposition for Minimum Rank Approximation // IEEE Transactions on Information Theory. — 2010. — Vol. 56, no. 9. — P. 4402–4416.

[143] Hastie T. J., Mazumder R., Lee J., Zadeh R. B. Matrix completion and low-rank SVD via fast alternating least squares // Journal of machine learning research : JMLR. — 2014. — Vol. 16. — P. 3367–3402.

[144] Tanner J., Wei K. Low rank matrix completion by alternating steepest descent methods // Applied and Computational Harmonic Analysis. — 2016. — Vol. 40, no. 2. — P. 417–429.

[145] Wen Z., Yin W., Zhang Y. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm // Mathematical Programming Computation. — 2012. — Vol. 4. — P. 333–361.

[146] Dai W., Milenkovic O. SET: An algorithm for consistent matrix completion // 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. — 2010. — P. 3646–3649.

[147] Vandereycken B. Low-Rank Matrix Completion by Riemannian Optimization // SIAM Journal on Optimization. — 2013. — Vol. 23, no. 2. — P. 1214–1236.

[148] Andersson F., Carlsson M. Alternating Projections on Nontangential Manifolds // Constructive Approximation. — 2013. — Vol. 38, no. 3. — P. 489–525.

[149] Nguyen L. T., Kim J., Shim B. Low-Rank Matrix Completion: A Contemporary Survey // IEEE Access. — 2019. — Vol. 7. — P. 94215–94237.

[150] Галкин В. А. Уравнение Смолуховского. — Физматлит, М., 2001.

[151] Gillis N. Nonnegative Matrix Factorization. — Philadelphia, PA : Society for Industrial and Applied Mathematics, 2020.

[152] Song G.-J., Ng M. K. Nonnegative low rank matrix approximation for nonnegative matrices // Applied Mathematics Letters. — 2020. — Vol. 105. — P. 106300.

[153] Budzinskiy S. Quasioptimal alternating projections and their use in low-rank approximation of matrices and tensors // arXiv 2308.16097 (Submitted on 30 Aug 2023). — 2023.

[154] Замарашкин Н. Л., Морозов С. В., Тыртышников Е. Е. Об алгоритме наилучшего приближения матрицами малого ранга в норме Чебышёва // Журнал вычислительной математики и математической физики. — 2022. — Т. 62, № 5. — С. 723–741.

Приложение А. Подробные версии алгоритмов

Поскольку в SRRQR [39], который используется для поиска $r \times k$, $k < r$ подматриц локально максимального объема в $A \in \mathbb{C}^{r \times N}$ (в оригинале был рассмотрен действительный случай; версию для комплексного случая мы выпишем ниже) для пересчета используются квадраты норм ошибок в столбцах, а также квадраты строк псевдообратной к текущим столбцам $C \in \mathbb{C}^{r \times k}$, это приводит к потере половины точности. Поэтому и при выборе начальных $k < r$ столбцов с помощью алгоритма выбора ведущих столбцов (алгоритм 4.1) достаточно ограничиться версией, сохраняющей лишь половину точности.

Алгоритм А.1 premaxvol

Вход: Матрица $R \in \mathbb{C}^{r \times N}$, требуемый ранг k.

Выход: Набор индексов столбцов $\mathcal{I}$ размера k, содержащий подматрицу большого объема.

$\mathcal{I} := \emptyset$
for $i := 1$ **to** N **do**
    $\gamma_i := \|R_{:,i}\|_2^2$
end for
$C = 0_{r \times N}$
for $i := 1$ **to** k **do**
    $j := \arg\max_j \gamma_j$
    $q := R_{:,j} - R_{:,\mathcal{I}} C_{1:i,j}$
    $q := q / \|q\|_2$
    $d := q^* R$
    $a := d_i$
    $d_{1:i} = 0$
    $\mathcal{I} := \mathcal{I} \cup \{j\}$
    {Мы сохраняем $R_{:,\mathcal{I}}^{-1}$ в первых i столбцах C}
    $C_{1:i-1,i} := -C_{1:i,j} / a$
    $C_{i,i} := 1/a$
    $\gamma := \gamma - d \odot \overline{d}$
    {Мы сохраняем $R_{:,\mathcal{I}^{-1}R}$ в оставшихся $N - i$ столбцах C}
    $C_{1:i-1,:} := C_{1:i-1,:} - C_{1:i-1,j} d / a$
    $C_{1:i,i} := d/a$
end for

Такой подход является выгодным, поскольку позволяет в 2 раза сократить общее число операций. Кроме того, в оценке общей сложности величина $O(Nkr)$ будет возникать только

в результате вычисления k произведений исходной $r \times N$ матрицы на вектор. Если матрица является разреженной, стоимость этой операции оказывается меньше, что приводит к меньшей асимптотической сложности всего алгоритма, см. теорему 4.2). Этот факт может быть важен, например, при восстановлении матриц.

Поскольку, как сказано выше, нам достаточно половинной точности, вместо пересчета $Q \in \mathbb{C}^{r\times k}$, $R \in \mathbb{C}^{k\times N}$ и $A - QR$ для матрицы A, как в случае выбора ведущих столбцов для алгоритма на основе отражений Хаусхолдера, нам достаточно лишь пересчитывать C^+A. Получаем алгоритм, который в коде назван $\mathrm{premaxvol}$. Вычисления в нем производятся таким образом, чтобы избежать существенного роста ошибки даже в случае, когда подматрица является вырожденной с точностью до машинного эпсилон. Если требуется не терять точность, то в момент, когда величина γ_j падает до машинного эпсилон от начальной, нужно вычислить все γ_i напрямую [66]. Чтобы также не терялась точность самого разложения, вместо (или вместе с) C^+A можно хранить Q и R, а каждый новый столбец q реортогонализовывать к предыдущим. Данный подход все еще будет быстрее стандартного метода через отражения Хаусхолдера, если $k \ll \min(r, N)$.

В случае $k = r$ получаем хорошие стартовые столбцы для алгоритма maxvol, который позволяет найти квадратную подматрицу локально максимального объема. Здесь мы явно записали, что набор r индексов текущей подматрицы сохраняется как первые r элементов перестановки N столбцов. Перестановка обновляется каждый раз, когда переставляются столбцы C, что мы обозначили функцией swap.

Алгоритм А.2 maxvol

Вход: Матрица $R \in \mathbb{C}^{r\times N}$, стартовый набор столбцов $\mathcal{I}$ of cardinality r.

Выход: Набор $\mathcal{I}$, содержащий доминантную подматрицу.

1: $C := R_{\mathcal{I}}^{-1} R$
2: permutation := $\{1, \ldots, N\}$
3: C.swap($\mathcal{I}, \{1, \ldots, r\}$, permutation)
4: $\{i, j\} := \arg\max_{i,j} \left|C_{ij}\right|$
5: **while** $\left|C_{ij}\right| > 1$ **do**
6: $\quad C^J := C_{:,j}$
7: $\quad C_i^J := C_i^J - 1$
8: $\quad C := C - C^J C_{i,:}/C_{ij}$
9: $\quad C$.swap(i, j, permutation)
10: $\quad \{i, j\} := \arg\max_{i,j} \left|C_{ij}\right|$
11: **end while**
12: $\mathcal{I}$:= permutation$[1..r]$

Дальнейшее добавление столбцов производится с помощью алгоритма $\mathrm{rect\text{-}maxvol}$, который подробно выписан в разделе 4.2, алгоритм 4.2.

Если затем требуется найти $r \times n$ подматрицу локально максимального объема, используется алгоритм $\mathrm{dominant}$.

Алгоритм А.3 $\mathrm{dominant}$

Вход: Матрица $R \in \mathbb{C}^{r \times N}$, стартовый набор столбцов $\mathcal{I}$ размера $n \geqslant r$.

Выход: Набор $\mathcal{I}$, содержащий $r \times n$ доминантную подматрицу.

$C := R_{\mathcal{I}}^{+} R$
permutation $:= \{1, \ldots, N\}$
$C.\mathrm{swap}(\mathcal{I}, \{1, \ldots, n\}, \mathrm{permutation})$
$l := 0_N$
for $i := 1$ **to** N **do**
    $l_i := \left\| C_{:,i} \right\|_2^2$
end for
$B := 0_{n \times N}$
for $i := 1$ **to** n **do**
    for $j := n + 1$ **to** N **do**
        $B_{ij} := \left| C_{ij} \right|^2 + (1 + l_j)(1 - l_i)$
    end for
end for
$\{i, j\} := \arg\max_{i,j} B_{ij}$
while $B_{ij} > 1$ **do**
    $C_I := \frac{C^*_{:,i}}{1 + l_i} C$
    for $k := 1$ **to** N **do**
        $\tilde{l}_k := l_k - \left| C_{I,k} \right|^2 (1 + l_i)$
    end for
    $C := C - C_{:,i} C_I$
    $C.\mathrm{swap}(i, j, \mathrm{permutation})$
    $C_I.\mathrm{swap}(i, j)$
    $\tilde{l}.\mathrm{swap}(i, j)$
    $\mathrm{swap}(C_{j,:}, C_I)$
    for $k := 1$ **to** M **do**
        $l_k := \tilde{l}_k + \frac{|C_{I,k}|^2}{1 - \tilde{l}_i}$
    end for
    $C := C + \frac{C_{:,j}}{1 - \tilde{l}_i} C_I$

dominant (продолжение)

```
    for i := 1 to n do
        for j := n + 1 to N do
            B_ij := |C_ij|^2 + (1 + l_j)(1 − l_i)
        end for
    end for
    {i, j} := arg max_{i,j} B_ij
end while
I := permutation[1..n]
```

Наконец, алгоритм SRRQR [39] также можно ускорить в несколько раз, если заменить вращения Гивенса на отражения Хаусхолдера. Это возможно сделать благодаря тому факту, что для текущей подматрицы не обязательно поддерживать верхнюю треугольную форму. Половина точности теряется за счет необходимости использования псевдообратной матрицы и квадратов длин столбцов c в процессе пересчета. Если использовать $c_i = 0$ в тех случаях, когда относительное значение c_i падает ниже машинного эпсилон, можно избежать возникновения нестабильности в алгоритме, без внесения каких-либо еще изменений (что и сделано в коде).

Алгоритм А.4 SRRQR

Вход: Матрица $A \in \mathbb{C}^{m \times N}$, стартовый набор столбцов $\mathcal{I}$ размера $r < m$.

Выход: Набор столбцов $\mathcal{I}$, содержащий $m \times r$ доминантную подматрицу.

```
permutation := {1, ..., N}
A.swap(I, {1, ..., r}, permutation)
A_{:,I} := QR
AI := R^{-1}
AB := A_{:,I} A
for i := 1 to N do
    c_i := ||A_{:,i}||_2^2 − ||Q* A_{:,i}||_2^2
end for
for i := 1 to m do
    w_i := ||AI_{i,:}||_2^2
end for
B := AB ⊙ AB + w* c
{i, j} := arg max_{i,j} B_ij
ρ := sqrt(B_ij) e^{sqrt(−1) arg(AB)_ij}
```

SRRQR (продолжение)

15: **while** $|\rho| > 1$ **do**

16: $aii := (AI)_{i,:}$

17: $aig := AI\ (aii)^*$

18: $abi := (AB)_{i,:}$

19: $abj := (AB)_{:,j}$

20: $abij := (AB)_{ij}$

21: $cl := \left(A_{:,j} - A_{:,1:r}\,(AB)_{:,j}\right)^* A_{:,r+1:}$

22: $(AB)_{:,j} := 0$

23: $(AB)_{ij} := 1$

24: $abi_j := 1$

25: swap $\left(R_{:,i}, A_{:,j}\right)$

26: swap $\left(A_{:,i}, A_{:,j}\right)$

27: $l := c_j$

28: $c := c - c_j \mathrm{Re}\left(\left(cl \cdot \left(1 - \frac{abij}{\rho}\right)/c_j + abi/\rho\right) \odot \overline{\left(cl \cdot \left(1 + \frac{abij}{\rho}\right)/c_j - abi/\rho\right)}\right)$

29: $c_j := l/B_{ij}$

30: $v := abj/\rho + aig\left(1 - \frac{abij}{\rho}\right)/w_i$

31: $v_i := 1 - 1/\rho$

32: $AI := AI - v \cdot aii$

33: $wi := w_i$

34: $w := w + v^* \odot (wi \cdot v - 2\,aig)$

35: $AB := AB - v \cdot abi$

36: $vaig := abj/\rho - aig \cdot abij/(wi \cdot \rho)$

37: $vaig_i := -1/\rho$

38: $cv := cl \cdot wi/\overline{\rho} - abi\left(1 - \frac{abij}{\rho}\right)$

39: $cv_j := -1 + \frac{abij}{\rho}$

40: $AB := AB - vaig \cdot cv$

41: $B := AB \odot AB + w^* c$

42: $\{i, j\} := \arg\max_{i,j} B_{ij}$

43: $\rho := \sqrt{B_{ij}}\, e^{\sqrt{-1}\arg(AB)_{ij}}$

44: **end while**

45: $\mathcal{I} :=$ permutation$[1..r]$

Здесь также можно достичь большей точности, вычислив аккуратно начальное c и реортогонализуя погрешность в строке 21. В этом случае потребуется обновлять текущую матрицу Q,

что можно сделать по формуле

$$Q := Q\left(I - \left(1 - \frac{abij}{\rho}\right)\frac{aii^* \cdot aii}{wi}\right) + \frac{\sqrt{l}}{\rho} \cdot \frac{(I - QQ^*)\left(A_{:,j} - A_{:,1:r}\,(AB)_{:,j}\right)}{\left\|(I - QQ^*)\left(A_{:,j} - A_{:,1:r}\,(AB)_{:,j}\right)\right\|_2} \cdot aii.$$

Если $m \sim N$, такой пересчет Q может оказаться дорогим. В этом случае можно вообще не вычислять Q, а вместо реортогонализации после строки 21 добавить

$$cl := cl - \left(A_{:,j} - A_{:,1:r}\,(AB)_{:,j}\right)^* A_{:,1:r}\,(AB)_{:,j}\,,$$

что также позволит не потерять точность итоговой CC^+A аппроксимации.

Запишем также алгоритмы поиска сильно невырожденных подматриц из раздела 4.6. Алгоритм A.5 жадно добавляет столбцы вплоть до r штук, на каждом шаге минимизируя величину $\left\|\hat{R}^{+}R\right\|_F$ для текущей подматрицы $\hat{R} \in \mathbb{C}^{r\times k}$ матрицы $R \in \mathbb{C}^{r\times n}$. Текущая подматрица поддерживается верхней треугольной за счет использования отражений Хаусхолдера.

Алгоритм A.5 Жадный набор столбцов до r

Вход: Строки $R \in \mathbb{C}^{r\times N}$.

Выход: Подмножество столбцов $\hat{R} = R_{:,\mathcal{I}} \in \mathbb{C}^{r\times r}$, для которого $\left\|\hat{R}^{-1}R\right\|_F$ достаточно мала.

$R = LQ$

$V := Q$

$U := 0_{r\times N}$

for $k := 0$ **to** $r-1$ **do**

  $j = \arg\min_{j>k} \left(1 + \left\|U_{1:k,j}\right\|_2^2\right) / \|V_{k+1:r,j}\|_2^2$

  Перестановка $k+1$-го и j-го столбца в R, U и V

  Генерируем вектор Хаусхолдера:

  $v := V_{k+1:r,k+1}$

  $v_1 := v_1 + e^{i\arg v_1}\,\|v\|_2$

  $v := v/\|v\|_2$

  Применяем вектор Хаусхолдера:

  $V_{k+1:r,1:N} = (I - 2vv^*)\,V_{k+1:r,1:N}$

  Обновляем $U = \hat{V}^{-1}V_{1:k,:}$:

  $U_{k+1,:} := V_{k+1,:}/V_{k+1,k+1}$

  $U_{1:k,:} := U_{1:k,:} - U_{1:k,k+1}U_{k+1,:}$

end for

Алгоритм А.5 эквивалентен дерандомизации выборки по объему в случае $n = r$ столбцов. Для полноты изложения запишем также алгоритм дерандомизации в общем виде как алгоритм А.6. Заметим, что на практике его применять не стоит, так как его полная сложность составляет $O\left(Nnr^2\right)$. Поскольку дерандомизация выборки по объему [43] (см. также доказательство теоремы 2.13) гарантирует

$$\left\|A - CC^+A\right\|_F^2 \leqslant (n+1)\,\frac{c_{n+1}\,(A)}{c_n\,(A)} \tag{A.1}$$

где

$$c_k = \sum_{i_1<\ldots<i_k} \sigma_{i_1}^2 \cdot \ldots \cdot \sigma_{i_k}^2,$$

а теорема 2.21 гарантирует

$$\frac{\|A - CC^+A\|_F}{\|A - A_r\|_F} \leqslant \sqrt{1 + \frac{\left\|\hat{Q}^+\right\|_F^2 - r}{N-r}} \tag{A.2}$$

для $A = \begin{bmatrix} Q \\ \varepsilon I \end{bmatrix} \in \mathbb{C}^{(N+r)\times N}$ в пределе $\varepsilon \to 0$ (см. также замечание 2.13), то, объединяя (А.1) и (А.2), получаем, что дерандомизация выборки по объему позволяет достичь

$$\left\|\hat{Q}^+\right\|_F \leqslant \sqrt{r\frac{N-r+1}{n-r+1}},$$

где $R = LQ$.

Алгоритм А.6 Дерандомизация выборки по объему

Вход: Строки $R \in \mathbb{C}^{r\times N}$, итоговое число столбцов n.

Выход: Подмножество столбцов $\hat{R} = R_{:,\mathcal{I}} \in \mathbb{C}^{r\times n}$, для которого $\left\|\hat{R}^+R\right\|_F$ достаточно мала.

1: $R = LQ$
2: **for** $k := 1$ **to** n **do**
3: **for** $i := 1$ **to** N **do**
4: Добавляем столбец i в $Q_{:,\mathcal{I}}$
5: Задаем первые $\min(r,k)$ сингулярных чисел $\Sigma \in \mathbb{R}^{(N-k-\max(0,r-k))\times(N-k-\max(0,r-k))}$ как сингулярные числа $Q^+_{:,\mathcal{I}}$, а остальные 1
6: $\Delta_i = c_{n-k+1-\max(0,r-k)}\,(\Sigma)\,/c_{n-k-\max(0,r-k)}\,(\Sigma)$
7: Удаляем столбец i из $Q_{:,\mathcal{I}}$
8: **end for**
9: Среди i выше добавляем в $U_{:,\mathcal{I}}$ тот, что соответствует максимальному Δ_i
10: **end for**

Основная сложность в асимптотике проистекает из вычисления коэффициентов характеристического полинома для Σ^2, которые, в свою очередь, получаются рекуррентно из коэффициентов характеристического полинома $\left(Q_{:,\mathcal{I}}Q^*_{:,\mathcal{I}}\right)^{-1}$ или, соответственно, $\left(Q^*_{:,\mathcal{I}}Q_{:,\mathcal{I}}\right)^{-1}$. В первом

случае можно вычислить спектральное разложение матрицы без добавления i-го столбца. После этого все возможные добавления нового столбца будут соответствовать изменениям ранга 1 диагональной матрицы. В этом случае можно напрямую вычислить, как изменятся собственные числа, а можно вычислить характеристический полином обновления диагональной D через uu^* напрямую как

$$\det\left(D + uu^* - \lambda I\right) = \det\left(D - \lambda I\right) + \sum_{i=1}^{r} |u_i|^2 \frac{\det\left(D - \lambda I\right)}{D_{ii} - \lambda}.$$

Так как каждое деление выполняется за $O(r)$, а $\det\left(D - \lambda I\right)$ вычисляется рекуррентно за $O\left(r^2\right)$, получаем сложность $O\left(r^2\right)$ внутри внутреннего цикла. Во втором случае все аналогично сводится к изменению ранга 1: единственной разницей будет то, что в D последнее диагональное значение будет нулевым, а сама матрица будет размера $k \times k$.

Алгоритм А.7 продолжает набор до $n > r$ столбцов. Запишем его, используя обновление матрицы $Z = \left(\hat{A}\hat{A}^*\right)^{-1}$, что сократит общее число операций. Если требуется минимизировать $\left\|\hat{R}^+\right\|_F$ (что можно сделать только при удалении столбцов) меняется только определение $d_i = \left\|\Sigma^{-1}\left(\hat{V}\hat{V}^*\right)^{-1} V_i\right\|_2^2$, где $V \in \mathbb{C}^{r\times N}$ – правые сингулярные векторы R, а $\Sigma \in \mathbb{C}^{r\times r}$ – матрица сингулярных чисел R. Если R плохо обусловлена, можно пересчитывать матрицу $\left(\hat{V}\hat{V}^*\right)^{-1} V$ (такой пересчет всегда устойчив), и вычислять нормы её взвешенных столбцов напрямую.

Алгоритм А.7 Жадный набор столбцов до $n > r$

Вход: Строки $R \in \mathbb{C}^{r\times N}$, число столбцов n.

Выход: Подмножество столбцов $\hat{R} = R_{:,\mathcal{I}} \in \mathbb{C}^{r\times n}$, для которого $\left\|\hat{R}^+ R\right\|_F$ достаточно мала.

Производим жадный набор r столбцов $\hat{Q}$ в Q из $R = LQ$ с помощью алгоритма А.5.
$C := Q$
$Z := \left(\hat{Q}\hat{Q}^*\right)^{-1}$
for $j := 1$ **to** N **do**
    $l_j := \left\|ZC_{:,j}\right\|_2^2$
    $d_j := \left\|Z^2 C_{:,j}\right\|_2^2$
end for
for $k := r+1$ **to** n **do**
    $i := \arg\max_{i, i\notin\mathcal{I}} \frac{d_i}{1+l_i}$
    Индекс i добавляется в множество $\mathcal{I}$
    $Z' := C_{:,i}^* Z$
    $C' := Z'C$
    $Z'' := Z'Z$
    $C'' := Z''C$
    for $j := 1$ **to** N **do**
        $d_j := d_j + \frac{d_i}{(1+l_i)^2}\left|C'_j\right|^2 - \frac{2}{1+l_i}\mathrm{Re}C''_j C'_j$
    end for
    $Z := Z - \frac{1}{1+l_i} Z'^* Z'$
    for $j := 1$ **to** N **do**
        $l_j := l_j - \frac{1}{1+l_i}\left|C'_j\right|^2$
    end for
end for

Алгоритм А.8 удаляет столбцы, начиная с $\hat{R} = R$ и заканчивая $\hat{R} \in \mathbb{C}^{r\times n}$, $n \geqslant r$.

Алгоритм А.8 Жадное удаление столбцов до $n \geqslant r$

Вход: Строки $R \in \mathbb{C}^{r\times N}$, число столбцов n.

Выход: Подмножество столбцов $\hat{R} = R_{:,\mathcal{I}} \in \mathbb{C}^{r\times n}$, для которого $\left\|\hat{R}^{+}R\right\|_F$ достаточно мала.

$R = LQ$
$C := Q$
$Z := I_{r\times r}$
for $j := 1$ **to** N **do**
    $l_j := \left\|C_{:,j}\right\|_2^2$
    $d_j := \left\|C_{:,j}\right\|_2^2$
end for
$\mathcal{I} := \{1, \ldots, N\}$
for $k := 1$ **to** $N - n$ **do**
    $i := \arg\min\limits_{i,i\in\mathcal{I}} \frac{d_i}{1-l_i}$
    Индекс i удаляется из множества $\mathcal{I}$
    $Z' := C^{*}_{:,i}Z$
    $C' := Z'C$
    $Z'' := Z'Z$
    $C'' := Z''C$
    for $j := 1$ **to** N **do**
        $d_j := d_j + \frac{d_i}{(1-l_i)^2}\left|C'_j\right|^2 + \frac{2}{1-l_i}\mathrm{Re}C''_jC'_j$
    end for
    $Z := Z + \frac{1}{1-l_i}Z'^{*}Z'$
    for $j := 1$ **to** N **do**
        $l_j := l_j + \frac{1}{1-l_i}\left|C'_j\right|^2$
    end for
end for

По аналогии с алгоритмом $\mathrm{dominant}$ можно также делать замены на основе критерия замены j-го столбца на i-й

$$\frac{(1+l_i)d_j - (1-l_j)d_i - 2\mathrm{Re}\left(D_{ij}\overline{C_{ij}}\right)}{(1+l_i)(1-l_j) + \left|C_{ij}\right|^2} \to \min_{i,j}, \quad C = \hat{Q}^+ Q, \quad D = \hat{Q}^+ \left(\hat{Q}\hat{Q}^*\right)^{-1} Q, \tag{A.3}$$

который следует из 4.77, где возможен быстрый пересчет матриц C и D. Здесь мы также ввели $d_j = \left\| \left(\hat{Q}\hat{Q}^*\right)^{-1} Q_{:,j} \right\|_2^2$. В отличие от поиска локально максимального объема, неизвестно, насколько это может улучшить оценку для итоговой подматрицы, и какое число замен для этого понадобится. Критерием останова будет неотрицательность минимума в (A.3).

Алгоритм A.9 позволяет найти $r \times n$ подматрицу с ограничением на $\left\|\hat{R}^+ R\right\|_F$. Использование алгоритма из [99] в таком виде было впервые предложено в [68]; название алгоритма сохранено.

Алгоритм A.9 Single-set spectral sparsification

Вход: Строки $R \in \mathbb{C}^{r\times N}$, итоговое число столбцов n.

Выход: Подмножество столбцов $\hat{R} = R_{:,\mathcal{I}} \in \mathbb{C}^{r\times n}$, для которого $\left\|\hat{R}^+ R\right\|_2$ достаточно мала.

1: $R = LQ$
2: $l := -\frac{\sqrt{r(n+1)}}{\frac{N}{1-\sqrt{\frac{r}{n+1}}}+2}$
3: $\delta_l := \frac{1}{\frac{N}{1-\sqrt{\frac{r}{n+1}}}+2}$
4: $\mathcal{I} = \emptyset$
5: **for** $k := 0$ **to** $n-1$ **do**
6: $\quad A := Q_{:,\mathcal{I}} Q_{:,\mathcal{I}}^*$
7: $\quad A_l := (A - lI)^{-1}$
8: $\quad A_\delta := (A - (l+\delta_l)\, I)^{-1}$
9: $\quad$ Добавим в $\mathcal{I}$ столбец j, соответствующий $\frac{\left\|A_\delta Q_{:,j}\right\|_2^2}{\mathrm{tr}\, A_\delta - \mathrm{tr}\, A_l} - Q_{:,j}^* A_\delta Q_{:,j} \to \max_j$
10: $\quad l := l + \delta_l$
11: **end for**

Возможно, значение l в данном алгоритме можно увеличивать более, чем на δ_l, если удалось найти достаточно хороший столбец, однако это требует дополнительного анализа.