Эвристические подходы к построению эллипсоида минимального объема вокруг подмножества точек*

П. С. Щербаков [I,II,III], Я. И. Квинто [I]

[I] Институт проблем управления им. В. А. Трапезникова Российской академии наук, Москва, Россия
[II] Московский Физико-технический институт, Долгопрудный, Россия
[III] Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Россия

Аннотация. В работе рассматривается следующая существенно комбинаторная задача: даны N точек в пространстве $\mathbb{R}^n$, построить эллипсоид минимального объема, содержащий ровно $N - k$ точек, где k много меньше N. Предлагаются шесть алгоритмов приближенного решения этой задачи, основанные на тех или иных эвристических соображениях. Приводятся численные результаты сравнительной эффективности алгоритмов при различных предположениях о механизме генерирования точек и их количестве.

Ключевые слова: точечное множество, отбраковка, выпуклая оптимизация, эллипсоид минимального объема, эвристика.

DOI 10.14357/20718632240411 **EDN** JCIBCK

Введение

Во многих разделах теории оценивания и фильтрации, обработки информации, анализа данных и принятия решений, разреженного представления данных и др. часто сталкиваются с ситуацией избыточности информации при ограниченном объёме хранилища данных, так что частью данных приходится жертвовать, но, по возможности, без заметного ущерба для качества содержащейся информации. Литература по данному вопросу безгранична, упомянем лишь несколько классических и наиболее «прорывных» работ [1-4]. В частности, в [2] приведена обширная библиография и подробная дискуссия по этой тематике.

В данной работе мы изучаем известную модельную задачу, формулировка которой представляется достаточно общей и важной для понимания природы рассматриваемого круга проблем. А именно, рассмотрим следующую задачу:

Из данного набора N точек $x_i \in S \subset \mathbb{R}^n$, $i = 1,\ldots, N$, удалить ровно k штук (где k много меньше N) так, чтобы эллипсоид, содержащий оставшиеся точки, имел наименьший объем.

Эллипсоидальная форма покрывающего множества выбрана по ряду причин. Во-первых, эллипсоидальные ограничения достаточно богаты (например, по сравнению с покомпонентными интервальными) и в то же время компактны в

* Результаты исследований П.С. Щербакова, представленные в разделе 2, получены за счет средств Российского научного фонда (проект №21-71-30005, https://rscf.ru/project/21-71-30005/).

описании (например, по сравнению с многогранниками); они определяются лишь матрицей, задающей форму эллипсоида, и его центром. Еще важнее, эллипсоидальная формулировка охватывает много типичных постановок задач из статистики и обработки информации при наличии паразитных помех и выбросов в измерениях, которые должны отфильтровываться. Например, если известно, что наблюдения имеют гауссовское распределение с неизвестными средним и ковариационной матрицей и содержат небольшой шум и малое количество выбросов (ошибок измерений), то естественной целью является восстановление исходного неизвестного гауссовского распределения. Это определяет выбор формы покрывающих множеств, связывая изучаемую задачу с построением доверительных эллипсоидов, с высокой вероятностью описывающих имеющиеся данные.

В специальной литературе, например [5], интересующая нас задача и ее варианты называется k-MVE, где k — количество отбрасываемых точек, а MVE означает эллипсоид минимального объёма (Minimum Volume Ellipsoid). Эта задача имеет два основных компонента: а) отыскание (суб)оптимального подмножества точек мощности $(N - k)$ и б) точное или приближенное построение эллипсоида минимального объёма вокруг данного точечного множества.

Задача б), часто именуемая MVEP, может быть численно точно решена многими способами. Общепринятым подходом к её решению является сведение проблемы к задаче выпуклой оптимизации с последующим применением методов внутренней точки [6; 7] и использованием общедоступных пакетов программ, таких, как, например, cvx [8]. Однако применимость такого способа решения ограничена задачами сравнительно невысоких размерностей и мощностей наборов данных. С другой стороны, имеются и специализированные методы, основанные на иных идеях и пригодные для решения больших задач, например, [9; 10]; утверждается, что такие методы работоспособны вплоть до $N = 30\,000$ и $n = 30$.

Задача а) существенно комбинаторна; в литературе можно найти несколько подходов к ее решению.

Первый подход предполагает отыскание точного решения с *уменьшенной вычислительной сложностью*. Примерами таких методов могут служить специализированные варианты метода ветвей и границ [11] или использование того факта, что данная задача относится к классу *LP-подобных* задач [12-14], которые могут пониматься как некоторое обобщение задачи линейного программирования. Подчеркнем, что предлагаемые подходы, конечно, имеют комбинаторную сложность, но общие затраты на вычисления существенно ниже.

Второй подход основан на рандомизации исходной задачи в духе монографии [15]: и данные, и результат решения предполагаются случайными, и при этом часто удается получить вероятностные оценки точности и вычислительной сложности. Соответствующая литература довольно богата, отметим работу [16], где приводится обширная библиография и предложен новый метод повторного генерирования выборки (resampling) с доказанными статистическими свойствами получаемого решения. В очень глубокой работе [14] подробно описан метод со случайной выборкой и удалением (sampling-and-removal) для решения задач, схожих с k-MVE; особое внимание уделяется LP-подобным задачам, упомянутым выше. Этот подход во многом основан на идеях оптимизации со случайными ограничениями (chance-constrained optimization) [17]. Упомянем также работу [18], где изучалась в некотором смысле близкая задача построения рандомизированных аппроксимаций невыпуклых множеств множествами регулярной структуры (прямоугольники, многогранники, эллипсоиды).

Третий подход можно назвать приближенным детерминированным. Решение основано на той или иной «эффективной эвристике», точность решения вряд ли может быть строго оценена, но численное моделирование говорит о разумном или хорошем качестве решения. Примером методов такого типа является так называемая итеративная 2-exchange процедура, предложенная в [5]; она основана на сравнении двух эллипсоидов, покрывающих подмножества, отличающиеся одной точкой. В принципе, эта процедура довольно эффективна, но резуль-

тат зависит от выбираемого начального приближения, так что на практике приходится прибегать к множественным рестартам с разными начальными условиями.

В настоящей работе мы следуем третьему направлению и предлагаем несколько простых приближенных детерминированных методов. Они основаны либо на общем здравом смысле, либо на тех или иных эвристических соображениях, почерпнутых из статистики, оптимизации и др. Далее проводится численное тестирование алгоритмов на простых наборах случайно сгенерированных данных. Для малых значений N и k точное решение задачи k-MVE (оптимальный эллипсоид $\mathbf{E}^*$) может быть получено прямым перебором всех возможных комбинаций из k удаляемых точек, вычислением соответствующего эллипсоида минимального объёма вокруг остальных точек и выбора наилучшего из них. Результаты работы обсуждаемых алгоритмов затем сравниваются друг с другом и с оптимальным решением по некоторому набору показателей качества: для больших значений N, k, полного перебора не проводится.

Статья организована следующим образом. В разделе 1 приводятся основные способы задания эллипсоидов и формулировка задачи MVEP в виде *линейных матричных неравенств* (linear matrix inequality, LMI). Предлагаемые алгоритмы приближенного решения задачи k-MVE и некоторые их модификации формулируются в разделе 2. Построение тестовых наборов данных и результаты численных экспериментов на этих данных описаны в разделе 3.

1. Эллипсоид минимального объема, содержащий заданное точечное множество

В этом разделе вводятся необходимые обозначения и некоторые стандартные определения, например, [6; 7].

Стандартное задание эллипсоида $\mathbf{E} \subset \mathbb{R}^n$ следующее:

$$\mathbf{E} = \{x \in \mathbb{R}^n : (x - x_c)^{\mathrm{T}} P^{-1} (x - x_c) \le 1\}, \qquad (1)$$

где $P \succ 0$ — положительно определенная матрица, определяющая форму эллипсоида, а $x_c \in \mathbb{R}^n$ — его центр. Существуют иные способы описания, которые удобны при формулировках различных задач минимизации размера эллипсоида. Под размером эллипсоида понимается либо (а) радиус наименьшего шара, содержащего эллипсоид, либо (б) сумма квадратов его полуосей, либо (в) объем.

Для рассматриваемой задачи первый критерий, очевидно, является очень грубым и не дает полной информации об эллипсоиде. Второй показатель достаточно информативен, и построение эллипсоида по этому критерию (равно как и по первому) сводится к стандартной задаче полуопределенного программирования (semidefinite programming, SDP). Мы, однако, примем наиболее естественную характеристику размера – объём, который дается формулой

$$\mathrm{Vol}(\mathbf{E}) = v_n (\det P)^{1/2},$$

где через v_n обозначен объём n-мерного единичного евклидового шара, а $\det(\cdot)$ обозначает определитель матрицы. Для оптимизации такого критерия нам будет удобнее пользоваться следующим описанием эллипсоида, эквивалентным определению (1):

$$\mathbf{E} = \{x \in \mathbb{R}^n : \|Qx - a\| \le 1\}, \qquad (2)$$

где: $Q = P^{-1/2}$, $a = P^{-1/2} x_c$, а $\|\cdot\|$ — евклидова векторная норма. Используя дополнение по Шуру, неравенство в (2) может быть переписано в эквивалентной форме в виде линейного матричного неравенства

$$\begin{pmatrix} 1 & (Qx - a)^{\mathrm{T}} \\ Qx - a & I \end{pmatrix} \succcurlyeq 0,$$

где I — единичная матрица подходящего размера. Поскольку функция $-\log\det X$ выпукла для матриц $X \succ 0$, то, имея в виду $P = Q^{-2}$, получаем $\log\det P = -2\log\det Q$. Таким образом, для данных точек $x_1, \ldots, x_N$ сформулируем следующую *задачу выпуклой оптимизации* с переменными $Q = Q^{\mathrm{T}} \in \mathbb{R}^{n \times n}$ и $a \in \mathbb{R}^n$:

$$\text{MVE Problem: } -\log\det Q \to \min \qquad (3)$$

при LMI ограничениях

$$\begin{pmatrix} 1 & (Qx_i - a)^{\mathrm{T}} \\ Qx_i - a & I \end{pmatrix} \succcurlyeq 0,\ i = 1, \ldots, N,\ Q \succ 0.$$

Решение этой задачи MVEP (Minimum Volume Ellipsoid Problem) определяет матрицу $P = Q^{-2}$ и центр $x_c = Q^{-1} a$ искомого эллипсоида (1) минимального объёма, покрывающего заданные точки $x_1, \ldots, x_N$.

Для простоты предполагается, что точечное множество $\{x_1, \ldots, x_N\}$ полномерно, т.е. не принадлежит никакому подпространству $\mathbb{R}^n$ меньшей размерности.

Оптимизационная задача MVEP корректно определена и может быть эффективно решена численно с помощью общедоступных пакетов программ; мы будем пользоваться Matlab-совместимым пакетом cvx [8]. Из структуры задачи видно, что количество переменных и ограничений растет медленно с ростом n и N, и требования к памяти компьютера тоже невысоки, так что задачи разумной размерности легко могут решаться с помощью пакета cvx. Тем не менее даже при невысоких размерностях и небольших N решение требует определённого времени, например, в очень простой ситуации $n = 2$, $N = 10$ вычисления занимают около 0,1 с на стандартном ноутбуке, при $N = 100$ для решения требуется около 1 с, и порядка 9 с – при $n = 10$, $N = 200$. В экспериментах мы ограничивались невысокими значениями n и N; см. обсуждение в разделе 3.

Задача MVEP лежит в основе всех методов, предлагаемых в работе; ее нужно решать неоднократно в процессе работы алгоритмов, поэтому число обращений к этой процедуре может рассматриваться как мера вычислительной сложности методов.

2. Шесть эвристических алгоритмов

Общая структура всех методов, рассматриваемых в работе, очень проста: на каждой итерации отбрасывается одна точка, пока количество оставшихся точек не окажется равным $N - k$, а выбор отбрасываемой точки происходит на основе той или иной эвристики, определяющей локально оптимальный выбор. Таким образом, все методы являются «жадными» (greedy).

В этом разделе предложим шесть методов такого типа и обсудим их возможные модификации.

2.1. «Наивные» подходы

Алгоритмы этой группы очень просты и основаны на общих соображениях здравого смысла.

Метод I: «Шелушение» сферы[1] *(Spherical Peeling, SP)*. На очередной итерации, имея M оставшихся точек x_i, вычисляем среднее значение

$$\overline{x} = \frac{1}{M}\sum_{i=1}^{M} x_i \qquad (4)$$

и строим с центром в этой точке содержащий их шар минимального радиуса. Удаляем любую из точек, лежащих на границе шара (иными словами, удаляем точку, максимизирующую расстояние до среднего). После k итераций решаем задачу MVEP (3) для оставшихся $N - k$ точек.

По-видимому, это самый простой и самый быстрый метод, который только может быть предложен. Он требует лишь одного обращения к процедуре MVEP, но, очевидно, точность такого метода ожидается быть довольно низкой, что и подтверждается экспериментами.

Метод II: «Шелушение» эллипсоида (Ellipsoidal Peeling, EP). Этот метод по существу заимствован из [15] (алгоритмы 12.2 и 12.3) и является естественным обобщением предыдущего подхода. На каждой итерации решается задача MVEP (3) для имеющихся M точек и выбрасывается любая точка, лежащая на границе получаемого эллипсоида. Итерации продолжаются до тех пор, пока остается ровно $(N - k)$ точек. Метод требует $(k + 1)$ обращений к процедуре MVEP.

В «упрощенной» версии алгоритма на каждой итерации удаляется сразу k точек, наиболее близких к границе эллипсоида (это в равной степени относится к предыдущему методу SP), или все точки, лежащие на границе и т.д. Эксперименты в целом свидетельствуют в пользу основной версии алгоритма, когда на каждой итерации удаляется одна точка.

Метод III: «Шелушение» выпуклой оболочки (Convex Hull Peeling, CHP). В основе метода лежит следующее соображение. Имея M точек на текущей итерации, можно рассмотреть все M комбинаций из $M - 1$ точки, строить минимальный эллипсоид вокруг каждой из комбинаций и выбирать наилучшую (минимизирующую объём). Вместо этого, чтобы понизить вычислительную сложность, предлагается рассматри-

[1] Калька с английского peeling (отшелушивание, очистка) от peel (кожура, очищать).

вать лишь наборы мощности $(M-1)$, получаемые удалением из общего набора одной точки из *выпуклой оболочки*. Известно [19], что мощность выпуклой оболочки множества из N точек, равномерно распределённых в кубе в $\mathbb{R}^n$, имеет порядок $O(\log^{n-1} N)$. Поэтому такой алгоритм требует порядка $k\,O(\log^{n-1} N)$ обращений к процедуре MVEP. Существуют аналогичные оценки мощности выпуклой оболочки для точек, имеющих гауссовское распределение [20].

Однако следует заметить, что вычисление выпуклой оболочки в пространствах высоких размерностей является довольно трудоёмкой операцией.

Далее обсудим более продвинутые подходы, основанные на эвристиках, известных из статистики, оценивания и разреженного представления и восстановления данных.

2.2. Выборочная ковариационная матрица

Методы, рассмотренные выше, не брали во внимание происхождение имеющихся данных. В этом разделе изучим подход в основе которого лежит предположение о стохастической природе располагаемой информации, а именно, считаем, что точки сгенерированы случайным механизмом и имеют некоторое вероятностное распределение.

Метод IV: Ковариационная матрица (Cov). Имея M точек на очередной итерации, вычислим их среднее (4) и выборочную ковариационную матрицу

$$H = \frac{1}{M}\sum_{i=1}^{M}(x_i - \bar{x})(x_i - \bar{x})^{\mathrm{T}}. \qquad (5)$$

Отбрасываем точку x_{out}, наиболее удалённую от $\bar{x}$ в метрике, задаваемой матрицей H:

$$x_{out} = \arg\max_i (x_i - \bar{x})^{\mathrm{T}} H^{-1} (x_i - \bar{x}).$$

Продолжаем итерации до тех пор, пока не будут удалены ровно k точек.

Как и в методе *SP*, здесь требуется всего одно обращение к процедуре MVEP, но качество получаемого решения оказывается значительно выше. Общее время исполнения чуть больше, чем у алгоритма *SP* из-за дополнительных операций по вычислению матрицы (5). Для пересчета этой матрицы можно предложить простую рекуррентную процедуру, которая, впрочем, не сильно влияет на скорость метода.

2.3. Метод главных компонент

Второй «продвинутый» подход к решению задачи основан на идеях метода главных компонент [21; 22], нацеленного на отбраковку выбросов в данных.

Метод V: Метод главных компонент (Principal Component Analysis, PCA). Простейшая версия метода выглядит следующим образом. Строим эллипсоид минимального объёма вокруг всех имеющихся на данной итерации точек, проектируем точки на малую ось эллипсоида и отбрасываем ту точку, чья проекция наиболее удалена от центра.

Пусть пара (Q, a) определяет текущий эллипсоид (2) и пусть e — собственный вектор матрицы Q, отвечающий минимальному собственному значению. Тогда проекция точки x_i на малую ось эллипсоида вычисляется как

$$\pi_e(x_i) = e^{\mathrm{T}}(x_i - x_c)e + x_c,$$

а расстояние до центра x_c равно

$$\operatorname{dist}(\pi_e(x_i), x_c) = \|\pi_e(x_i) - x_c\| = \\ = \|e^{\mathrm{T}}(x_i - x_c)e\| = |e^{\mathrm{T}}(x_i - x_c)|.$$

Удаляем ту точку, которая максимизирует это расстояние, обозначим ее через x_{small}.

В экспериментах также рассматривались проекции на *большую* ось, определялась соответствующая точка x_{large}, максимизирующая расстояние до центра, и отбрасывалась наихудшая точка из пары (x_{small}, x_{large}). Эксперименты свидетельствуют о большей эффективности такой модификации, она требует $2k+1$ вызовов процедуры MVEP.

2.4. Подход на основе ℓ_1-оптимизации

Еще один содержательный подход основан на идеях ℓ_1-оптимизации и разреженности [7; 3].

Метод VI: Идеи ℓ_1-оптимизации (ℓ_1). Схема алгоритма следующая. Пусть пара (Q, a) определяет эллипсоид на текущей итерации. Вместо условия $\|Qx - a\|^2 \le 1$ в (2) рассмотрим ограничения

$$\|Qx_i - a\|^2 \le 1 + d_i,\ d_i \ge 0;\ i = 1, \ldots, M,$$

где число d_i имеет смысл штрафа для x_i за нахождение вне эллипсоида. Введем векторную переменную $d = (d_1, \ldots, d_M)^{\mathrm{T}}$ и решим следующую задачу выпуклой оптимизации с переменными Q, a, d:

$$-\log\det Q + \mu\|d\|_1 \to \min$$

при ограничениях

$\|Qx_i - a\| \le 1 + d_i,\ d_i \ge 0;\ i = 1, \ldots, M,\ Q \succ 0$, где $\|d\|_1 = \Sigma_i |d_1|$ — векторная ℓ_1-норма, а $\mu > 0$ — скалярный параметр. Понятно, что введенные ограничения могут быть переписаны в форме линейных матричных неравенств подобно ограничениям в MVEP (3).

Пусть $\mathbf{E}_\mu$ обозначает эллипсоид, определяемый решением (Q, a) сформулированной оптимизационной задачи. Ясно, что при больших значениях параметра μ все точки окажутся внутри эллипсоида $\mathbf{E}_\mu$, а с уменьшением μ получающийся эллипсоид сжимается и не содержит точек. Поэтому на каждой итерации начинаем с некоторого большого значения μ и уменьшаем его (например, дихотомией) до тех пор, пока вне соответствующего эллипсоида $\mathbf{E}_\mu$ не окажется ровно одна точка. Отбрасываем ее и переходим к следующей итерации с оставшимися точками; останавливаемся, когда ровно k точек отброшены.

В экспериментах пользовались следующей модификацией метода. На каждой итерации величина параметра μ уменьшается до тех пор, пока вне эллипсоида $\mathbf{E}_\mu$ не окажется ровно *половина* точек (точнее $\lfloor M/2 \rfloor$), после чего отбрасывается точка x_i с наибольшим значением штрафа d_i. Эвристика, лежащая в основе такой модификации, не вполне ясна, но по результатам экспериментов такая версия алгоритма оказалась много более эффективной. В обоих случаях оценить количество обращений к процедуре MVEP представляется затруднительным, но очевидно, что оно «велико».

2.5. Еще о возможных модификациях

Возможны следующие более или менее очевидные модификации, ускоряющие или упрощающие решение, или делающие его чуть более точным за счет небольшого усложнения алгоритма.

1. Методы допускают гораздо более быструю, но, разумеется, и более грубую модификацию, в которой на одной единственной итерации сразу удаляются ровно k наихудших точек.

2. На каждой итерации центр эллипсоида не является оптимизационной переменной, а фиксируется как среднее значение текущих оставшихся точек.

3. Осуществление полного перебора на последней итерации: имея оставшиеся $N - k + 1$ точек, проверить все комбинации из $N - k$ точек и выбрать ту, которая минимизирует объём (ср. с методом *CHP*).

4. В специальном случае $n = 2$ эллипсоид задается небольшим числом параметров (два для центра и три для матрицы); в этом случае можно организовать «недорогую» прямую оптимизационную процедуру поиска минимального эллипса.

3. Численные эксперименты

В экспериментах мы ограничились маломерными (n) данными малого и небольшого объемов (N). Причина — в использовании пакета cvx, в котором оптимизационные процедуры основаны на методах внутренней точки. При высоких размерностях принятый в этом пакете способ представления данных может потребовать очень большой памяти компьютера и серьезных временны*х* затрат. Как упоминалось во введении, в ситуации с большими объемами данных можно пользоваться более мощными средствами; например такими, как в [9; 10]. В настоящей работе мы прежде всего интересовались выяснением работоспособности и сравнительной эффективности методов, основанных на различных эвристиках.

3.1 Генерирование тестовых данных

В каждом эксперименте качество решений, полученных по алгоритмам, сравнивалось на $N_{set} = 500$ наборах данных, случайно сгенерированных из разных распределений. Первое из распределений — наиболее естественное гауссовское $x \sim \mathcal{N}(0, \Sigma)$, которое моделировалось как $x = F * \mathrm{randn}(n, 1)$, где $F \in \mathbb{R}^{n \times n}$, так что $\Sigma = FF^{\mathrm{T}}$. Матрица F либо фиксировалась раз и навсегда, либо также генерировалась случайно как $F = 2 * \mathrm{rand}(n) - 1$. Второе распределение — равномерное на единичном кубе $\mathbf{B} = \{x \in \mathbb{R}^n : \|x\|_\infty \le 1\}$.

Имеем три свободных параметра для набора данных: размерность n пространства, количество N точек и число k удаляемых точек. Соответствующие данные будем обозначать (n, N, k).

Табл. 1. Результаты первого эксперимента

	volume	wins	calls	time	std
SP	1,4011	29%	1	0,2912	0,5640
EP	1,1203	62%	4	1,1811	0,2317
CHP	1,1131	64%	19,47	4,7216	0,2271
Cov	1,0312	76%	1	0,3054	0,0882
PCA	1,1189	57%	7	1,7439	0,2047
ℓ_1	1,0292	78%	20,02	8,1572	0,0854
true	1	100%	120	32,5175	0

Далее, обозначим через $v_i(\mathrm{A})$ объем эллипсоида, полученного алгоритмом A для набора данных i, а через $\mathbf{v}_i$ — наилучший среди алгоритмов результат, полученный для набора i. В экспериментах с малыми N доступно точное решение, получаемое полным перебором. Соответствующая (последняя) строка в Табл. 1 и 2 с результатами экспериментов имеет обозначение *true*.

В каждом эксперименте качество алгоритма A характеризуем следующими показателями:

- средний относительный объём:

$$\mathrm{volume}(\mathrm{A}) = \frac{1}{N_{set}} \sum_{i=1}^{N_{set}} \frac{v_i(\mathrm{A})}{\mathbf{v}_i};$$

- стандартное уклонение случайной величины $\xi = \dfrac{v_i(\mathrm{A})}{\mathbf{v}_i}$:

$$\mathrm{std}(\mathrm{A}) = \left(\frac{1}{N_{set} - 1} \sum_{i=1} (\xi_i - \bar{\xi}) \right)^{1/2};$$

- среднее число «побед» в эксперименте (победа для набора i означает, что алгоритм A дает наименьший объём) либо среднее количество точных (оптимальных) решений:

$$\mathrm{wins}(\mathrm{A}) = \#\{v_i(\mathrm{A}) = \mathbf{v}_i\};$$

- среднее число calls обращений к процедуре MVEP;
- среднее время исполнения time(A), с.

3.2 Моделирование

Первый эксперимент: (2, 10, 3). В этом простом маломерном примере легко может быть получено точное решение $\mathbf{v}_i$. Он призван продемонстрировать, что часто некоторые методы дают точное решение задачи и что эллипсы, полученные по разным алгоритмам, могут значительно отличаться.

В этом эксперименте данные генерировались по гауссовскому распределению $\mathcal{N}(0, \Sigma)$ с $\Sigma = FF^{\mathrm{T}}$ и фиксированной матрицей F:

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

Результаты приведены в Табл. 1.

Первое наблюдение: все алгоритмы дают эллипсы очень схожего объёма (кроме метода *SP*, который даёт много худшие результаты), что, вероятно, объясняется малой размерностью точек и мощностью наборов. Всё же при этом методы *Cov* и ℓ_1 выглядят предпочтительнее.

Второе наблюдение: очень часто методы дают оптимальное решение; особенно это относится к *Cov* и ℓ_1. Вероятно, причина та же – малые размерности и мощности.

Третье наблюдение: методы *Cov* и ℓ_1 в 95% дают один и тот же эллипс, но при этом второй метод гораздо более трудоёмок.

Наконец, в целом, результаты свидетельствует о действительно разной природе рассматриваемых методов, дающих значительно отличающиеся формы эллипсов, например, Рис. 1.

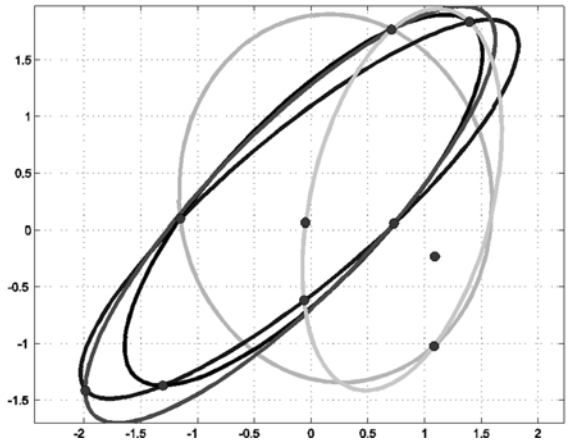

Рис. 1: Результаты работы разных алгоритмов для случайного набора точек (2, 10, 3)

Второй эксперимент: (2, 10, 3). В отличие от первого эксперимента точки генерировались равномерно на единичном квадрате. Результаты приведены в Табл. 2.

Результаты работы алгоритмов заметно отличаются от предыдущего эксперимента, указывая на то, что качество решения может существенно зависеть от априорного знания о природе набора данных.

В самом деле, в то время как число обращений к процедуре MVEP осталось тем же (что очевидно), количество точных решений заметно уменьшилось, как и средняя точность (столбцы 3 и 2 таблицы). Еще одно наблюдение — ощутимое увеличение стандартного отклонения (последний столбец) для всех методов, кроме *SP*. Однако, как и ранее, методы *Cov* и ℓ_1 превосходят остальные методы (столбцы 2 и 3).

Третий эксперимент: (2, 100, 3). В этом более размерном эксперименте точки генерировались из гауссовского распределения с различной случайно генерируемой ковариационной матрицей для каждого из N_{set} = 500 наборов данных (начало раздела 3.1), а все показатели качества методов оставлены теми же. Понятно, что отыскание точного решения слишком трудоёмко из-за большого объёма данных (для каждого набора требуется 161 700 обращений к процедуре MVEP). Результаты сведены Табл. 3.

Имеют место несколько очевидных наблюдений.

Во-первых, разумеется, время вычислений значительно возросло, поскольку на порядок увеличилось количество ограничений в оптимизационной задаче (3).

Во-вторых, метод *CHP* приводит к существенно лучшим результатам по сравнению с другими методами, хотя и за счет много более тяжелых вычислений (время исполнения алгоритма выросло в 8,5 раз при десятикратном увеличении количества N точек). Возможным объяснением является схожесть алгоритма *CHP* с переборными методами, поэтому он вряд ли может применяться к решению задач больших размерностей.

Наконец, методы *Cov*, *PCA* и ℓ_1 продемонстрировали очень близкие по точности результаты, но при этом метод *Cov* оказывается значительно более быстрым.

Подытоживая результаты проведенных экспериментов и имея в виду все показатели качества алгоритмов (прежде всего точность и время исполнения), можем сделать вывод о том, что метод *Cov*, основанный на использовании выборочной ковариационной матрицы, представляется наиболее предпочтительным.

Четвертый эксперимент: (4, 62, 2). Этот практический пример заимствован из [23], где изучалась взаимосвязь между свойствами целлюлозного волокна и получаемой из него бумаги. Данные представляют из себя N = 62

Табл. 2. Результаты второго эксперимента

	volume	wins	calls	time	std
SP	1,3313	15%	1	0,2728	0,3277
EP	1,2611	42%	4	1,1643	0,3527
CHP	1,2209	40%	18,65	4,8882	0,3414
Cov	1,1237	61%	1	0,2861	0,2317
PCA	1,2528	45%	7	1,7439	0,4403
ℓ_1	1,1287	55%	20,73	8,0745	0,2289
true	1	100%	120	30,5180	0

Табл. 3. Результаты третьего эксперимента

	volume	wins	calls	time	std
SP	1,1392	5%	1	1,4408	0,1529
EP	1,0911	31%	4	6,1843	0,0527
CHP	1,0053	85%	29,75	41,3714	0,0152
Cov	1,0393	45%	1	1,4824	0,0523
PCA	1,0474	40%	7	9,5766	0,0573
ℓ_1	1,0393	45%	49,70	100,5363	0,0523

измерения следующих четырех ($n = 4$) характеристик волокна: длина волокна, доля длинного волокна, доля тонкого волокна и прочность при растяжении. В работе [16] набор таких измерений анализировался на наличие и отбраковку выбросов в измерениях (аномальных значений четырех измеренных характеристик) с использованием разработанного авторами метода на основе решения задачи k-MVE. Получаемый эллипсоид имеет минимальный объем среди доверительных эллипсоидов с уровнем доверия $v = 0{,}975$. В нашей постановке задачи это соответствует нецелому числу $k_1 = 1{,}55$ выбрасываемых точек, поэтому в экспериментах мы положили $k = 2$. Как и ранее, при моделировании точки генерировались нами из гауссовского распределения с различной случайно генерируемой ковариационной матрицей для каждого из $N_{set} = 500$ наборов данных. Сравнение проводилось между методом из [16] и наиболее перспективным методом *Cov*, описанным выше в разделе 2.2.

Среднее (по $N_{set} = 500$ наборам данных) время вычислений по методу *Cov* оказалось равным 1,0604 с против 1,2831 с для метода из [16], а средний объем эллипсоида составил 12,5501 против 12,6171 для метода из [16]. Небольшое преимущество нашего метода по объему может объясняться тем, что формально мы выбрасываем «чуть больше» точек. Иными словами, качество методов сопоставимо, в то время как метод *Cov* продемонстрировал ощутимо более высокую скорость. Поскольку метод из [16] осуществляет пусть и неполный, но все же перебор, то с ростом мощности набора данных преимущество по скорости метода *Cov*, по-видимому, будет увеличиваться.

4. Направления дальнейших исследований

В работе предложено несколько простых методов приближенного решения задачи k-MVE и приведены результаты предварительных числовых экспериментов. Теоретические оценки эффективности методов вряд ли возможны. Желательно проведение более масштабного моделирования, которое предполагается осуществить в будущем. Сюда прежде всего относится оптимизация как используемых числовых процедур, так и реализованного программного кода – для ускорения работы методов и экономии памяти при бо́льших объемах и размерностях данных. Существенными при проведении такого моделирования являются следующие показатели тестовых точечных множеств:

- происхождение точек (из предварительных экспериментов следует, что поведение каждого отдельного метода зависит от принятого вероятностного распределения и формы носителя S);
- количество N точек и соотношение между N и k;
- наличие выбросов (использование загрязнённого гауссовского распределения);
- наличие кластеров.

Далее, вероятно, знание числа k может быть использовано для оптимизации алгоритмов.

Заключение

Сформулируем на физическом уровне строгости три задачи, имеющие прямое отношение к рассмотренной проблеме и представляющиеся важными.

Задача 1: Дано вероятностное распределение точек x_i и размер выборки. Оценить вероятность того, что эллипсоид, полученный тем или иным методом, отличается от оптимального не более чем на заданную величину.

Задача 2: Даны N точек равномерно распределённых в единичном n-мерном кубе, и пусть v — случайная величина, имеющая смысл объёма содержащего их минимального эллипсоида. Оценить математическое ожидание Ev величины v.

Задача 3: Задан объём эллипсоида. Максимизировать (по матрице и центру) количество точек, им покрываемых. Эта задача может рассматриваться как «обратная» к задаче k-MVE.

Литература

1. Becker S., Bobin J., Candes E. NESTA: A fast and accurate first-order method for sparse recovery. SIAM J. on Imaging Sciences. 2009;4(1):1–39.
2. Burke E.K., Kendall G. (eds.) Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques. New York: Springer Science+Business Media, 2014.
3. Donoho D.L. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. Communications on Pure and Applied Mathematics. 2006;56(6):797–829.

4. Ruan D., Chen G., Kerre E.E., Wets G. (eds.) Intelligent Data Mining: Techniques and Applications. Studies in Computational Intelligence. Vol. 5. New York: Springer, 2005.
5. Ahipaşaoğlu S.D. Fast algorithms for the minimum volume estimator. Journal of Global Optimization. 2015;62(2):351–370.
6. Boyd S., El Ghaoui L., Feron E., Balakrishnan V. Linear Matrix Inequalities in System and Control Theory. Philadelphia: SIAM, 1994.
7. Boyd S., Vandenberge L. Convex Optimization. Cambridge, MA: Cambridge University Press, 2004.
8. Grant M., Boyd S. CVX: Matlab software for disciplined convex programming, ver. 2.2. Available from: http://stanford.edu/~boyd/cvx, 2020.
9. Ahipaşaoğlu S.D., Sun P., Todd M.J. Linear convergence of a modified frank-wolfe algorithm for computing minimum-volume enclosing ellipsoids. Optimization Methods and Software. 2008;23:5–19.
10. Sun P., Freund R.M. Computation of minimum-volume covering ellipsoids. Operations Research. 2004;52(5):690–706.
11. Agullo J. Exact iterative computation of the multivariate minimum volume ellipsoid estimator with a branch and bound algorithm. In: Pratt A. (ed.) Proceedings in Computational Statistics. Heidelberg; Physica-Verlag, 1996. p. 175–180.
12. Bai E.-W., Chi H., Tempo R., Ye Y. Optimization with few violated constraints for linear bounded error parameter estimation. IEEE Trans. Autom. Control. 2002;47(7):1067–1077.
13. Dabbene F., Shcherbakov P. Minimum volume ellipsoid comprising a subset of points. In: Proceedings of the 6th Int. Conf. Control, Decision Inform. Technologies (CoDiT 2019), Paris, Apr 23–26, 2019, paper WiP 20
14. Gärtner B.. Sampling with removal in LP-type problems. Journal of Computational Geometry. 2015;6(2):93–112.
15. Tempo R., Calafiore G., Dabbene F. Randomized Algorithms for Analysis and Control of Uncertain Systems: With Applications. 2nd ed. Springer, 2013.
16. Van Aelst S., Rousseeuw P. Minimum volume ellipsoid. WIREs Computational Statistics. 2009;1:71–82.
17. Campi M.C., Garatti S. A sampling-and-discarding approach to chance-constrained optimization: Feasibility and optimality. Journal of Optimization Theory and Applications.2011;148:257–280.
18. Dabbene F., Henrion D., Lagoa C., Shcherbakov P. Randomized approximations of the image set of nonlinear discrete-time systems with applications to filtering. In: Proceedings of the 8th IFAC Symposium on Robust Control Design (ROCOND 2015), Bratislava, Slovak Republic, Jul 8–11, 2015.
19. Raynaud H. Sur l'enveloppe convexe des nuages de points aleatoires dans IRn. I. Journal of Applied Probability. 1970;7(1):35–48.
20. Hueter I. Limit theorems for the convex hull of random points in higher dimensions. Trans. Amer. Math. Soc. 1999;351(11):4337–4363.
21. Jolliffe I.T. Principal Component Analysis. Springer Series in Statistics. New York: Springer, 2002.
22. Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine. 1901;2:559–572.
23. Lee J. Relationships between Properties of Pulp-Fibre and Paper. PhD Diss, University of Toronto, 1992.

Щербаков Павел Сергеевич. Институт проблем управления им. В.А. Трапезникова РАН, Москва, Россия. Главный научный сотрудник, доктор физико-математических наук. Московский Физико-технический институт, Долгопрудный, Россия. Ведущий научный сотрудник. ФИЦ ИУ РАН, Москва, Россия. Ведущий научный сотрудник. Область научных интересов: робастный анализ, робастное управление, вероятностные методы в управлении, рандомизация, всплеск, линейные матричные неравенства, инвариантные эллипсоиды. E-mail: cavour118@mail.ru

Квинто Яна Игоревна. Институт проблем управления им. В.А. Трапезникова РАН им. В.А. Трапезникова РАН, Москва, Россия. Старший научный сотрудник, кандидат технических наук. Область научных интересов: робастное управление, рандомизированные методы в управлении, линейные матричные неравенства, инвариантные эллипсоиды, управление на основе данных. E-mail: yanakvinto@mail.ru

Heuristic Approaches to Constructing a Minimum Volume Ellipsoid around a Subset of Points

P. S. Shcherbakov[I,II,III], Y. I. Kvinto[I]

[I] V. A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
[II] Moscow Institute of Physics and Technology, Dolgoprudny, Russia
[III] Federal Research Center «Informatics and Control», Russian Academy of Sciences, Moscow, Russia

Abstract. The paper deals with the following essentially combinatorial problem: Given N points in $\mathbb{R}^n$, compose the ellipsoid of minimum volume containing exactly $N - k$ points where k is much less than N. Six algorithms for an approximate solution of this problem are proposed; they are based on certain heuristic considerations. Under various assumptions on the mechanism of generating the points and their

amount, the comparative efficiency of the algorithms was conducted and the results of numerical experiments were presented.

Keywords: point sets, partial information rejection, convex optimization, minimum volume ellipsoid, heuristics.

DOI 10.14357/20718632240411 **EDN** JCIBCK

References

1. Becker S., Bobin J., Candes E. NESTA: A fast and accurate first-order method for sparse recovery. SIAM J. on Imaging Sciences. 2009;4(1):1–39.
2. Burke E.K., Kendall G. (eds.) Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques. New York: Springer Science+Business Media, 2014.
3. Donoho D.L. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. Communications on Pure and Applied Mathematics. 2006;56(6):797–829.
4. Ruan D., Chen G., Kerre E.E., Wets G. (eds.) Intelligent Data Mining: Techniques and Applications. Studies in Computational Intelligence. Vol. 5. New York: Springer, 2005.
5. Ahipaşaoğlu S.D. Fast algorithms for the minimum volume estimator. Journal of Global Optimization. 2015;62(2):351–370.
6. Boyd S., El Ghaoui L., Feron E., Balakrishnan V. Linear Matrix Inequalities in System and Control Theory. Philadelphia: SIAM, 1994.
7. Boyd S., Vandenberge L. Convex Optimization. Cambridge, MA: Cambridge University Press, 2004.
8. Grant M., Boyd S. CVX: Matlab software for disciplined convex programming, ver. 2.2. Available from: http://stanford.edu/~boyd/cvx, 2020.
9. Ahipaşaoğlu S.D., Sun P., Todd M.J. Linear convergence of a modified frank-wolfe algorithm for computing minimum-volume enclosing ellipsoids. Optimization Methods and Software. 2008;23:5–19.
10. Sun P., Freund R.M. Computation of minimum-volume covering ellipsoids. Operations Research. 2004;52(5): 690–706.
11. Agullo J. Exact iterative computation of the multivariate minimum volume ellipsoid estimator with a branch and bound algorithm. In: Pratt A. (ed.) Proceedings in Computational Statistics. Heidelberg; Physica-Verlag, 1996. p. 175–180.
12. Bai E.-W., Chi H., Tempo R., Ye Y. Optimization with few violated constraints for linear bounded error parameter estimation. IEEE Trans. Autom. Control. 2002;47(7):1067–1077.
13. Dabbene F., Shcherbakov P. Minimum volume ellipsoid comprising a subset of points. In: Proceedings of the 6th Int. Conf. Control, Decision Inform. Technologies (CoDiT 2019), Paris, Apr 23–26, 2019, paper WiP 20
14. Gärtner B. Sampling with removal in LP-type problems. Journal of Computational Geometry. 2015;6(2):93–112.
15. Tempo R., Calafiore G., Dabbene F. Randomized Algorithms for Analysis and Control of Uncertain Systems: With Applications. 2nd ed. Springer, 2013.
16. Van Aelst S., Rousseeuw P. Minimum volume ellipsoid. WIREs Computational Statistics. 2009;1:71–82.
17. Campi M.C., Garatti S. A sampling-and-discarding approach to chance-constrained optimization: Feasibility and optimality. Journal of Optimization Theory and Applications.2011;148:257–280.
18. Dabbene F., Henrion D., Lagoa C., Shcherbakov P. Randomized approximations of the image set of nonlinear discrete-time systems with applications to filtering. In: Proceedings of the 8th IFAC Symposium on Robust Control Design (ROCOND 2015), Bratislava, Slovak Republic, Jul 8–11, 2015.
19. Raynaud H. Sur l'enveloppe convexe des nuages de points aleatoires dans IRn. I. Journal of Applied Probability. 1970;7(1):35–48.
20. Hueter I. Limit theorems for the convex hull of random points in higher dimensions. Trans. Amer. Math. Soc. 1999;351(11):4337–4363.
21. Jolliffe I.T. Principal Component Analysis. Springer Series in Statistics. New York: Springer, 2002.
22. Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine. 1901;2:559–572.
23. Lee J. Relationships between Properties of Pulp-Fibre and Paper. PhD Diss, University of Toronto, 1992.

Shcherbakov Pavel S. D.Sc. V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznaya str., Moscow, 117997, Russia. E-mail: cavour118@mail.ru.

Kvinto Yana I. PhD. V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznaya str., Moscow, 117997. E-mail: yanakvinto@mail.ru