Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики

На правах рукописи

Рябенко Евгений Алексеевич

Выбор функций потерь в задачах неотрицательного матричного разложения

05.13.18 — математическое моделирование, численные методы и комплексы программ

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Научный руководитель
д.ф.-м.н. К. В. Воронцов

Москва – 2014

Работа выполнена на кафедре математических методов прогнозирования факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова.

Научный руководитель:	доктор физико-математических наук **Воронцов Константин Вячеславович.**
Официальные оппоненты:	**Горнов Александр Юрьевич,** доктор технических наук, Федеральное государственное бюджетное учреждение науки Институт динамики систем и теории управления Сибирского отделения Российской академии наук, лаборатория оптимального управления, главный научный сотрудник; **Трушкин Евгений Владиславович,** кандидат биологических наук, общество с ограниченной ответственностью научно-технический центр «БиоКлиникум», главный инженер
Ведущая организация:	Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В. А. Трапезникова Российской академии наук

Защита состоится «23» октября 2014 г. в _____ на заседании диссертационного совета Д 002.017.04 при Федеральном государственном бюджетном учреждении науки Вычислительный центр им. А. А. Дородницына Российской академии наук по адресу: 119333, Москва, ул. Вавилова, д. 40, конференц-зал.

С диссертацией и авторефератом можно ознакомиться в библиотеке и на официальном сайте (http://www.ccas.ru/) ВЦ РАН.

Автореферат разослан «__» _____ 2014 г.

Ученый секретарь диссертационного совета,
доктор физико-математических наук,
профессор Н. М. Новикова

Общая характеристика работы

Диссертационная работа посвящена проблеме выбора функции потерь в задаче неотрицательного матричного разложения. Предложен способ адаптивного выбора функции потерь из семейства АБ-дивергенций, основанный на методе согласования вклада, а также мультипликативный алгоритм получения неотрицательного матричного разложения с гарантией сходимости. Полученные теоретические результаты применены к задаче анализа данных ДНК-микрочипов, для которой предложены новые модели, и на их основе создан программный комплекс, позволяющий получать более точные оценки экспрессии генов.

Актуальность темы. Развитие технологий сбора и хранения данных в последние десятилетия привело к увеличению объёмов данных и возникновению затруднений при использовании классических средств их обработки. Перед использованием индуктивных методов анализа к таким данным часто применяются сжимающие преобразования, которые позволяют уменьшить вычислительные затраты на обработку, выявить структурные особенности данных, уменьшить влияние погрешности. Одним из наиболее распространённых способов такого преобразования является переход к аппроксимации данных в некотором подпространстве. Формально, если исходные данные можно записать в виде матрицы (где, например, строки — это различные сенсоры, а столбцы — различные объекты измерения), то их аппроксимация представляет собой произведение двух матриц меньшей размерности, одна из которых задаёт подпространство, а вторая — коэффициенты разложения по нему. Такое представление данных называют факторизованным, а задачу его получения — задачей матричной факторизации.

Приложения, связанные с получением и анализом факторизованных представлений матриц, могут различаться ограничениями, накладываемыми на факторы. Так, в методе главных компонент факторы являются ортогональными (Pearson, 1901), в методе независимых компонент — независимыми (Hyvärinen, Karhunen, Oja, 2001). В задаче неотрицательного матричного разложения (non-negative matrix factorization, NMF), рассматриваемой в данной работе, ключевую роль играют ограничения на знак матриц-компонент. Впервые подобная задача была рассмотрена в работе (Paatero, Tapper, 1994) в приложении к задаче византийских генералов из теории отказоустойчивости, однако основной интерес к этой теме

возник после работ (Lee, Seung, 1999, 2001), авторы которых обобщили постановку задачи и предложили простой алгоритм получения её приближённого решения. Неотрицательные матричные разложения используются при анализе изображений, текстов, аудиозаписей, финансовых показателей, в вычислительной биологии, медицине и многих других прикладных областях. Подробный обзор применений можно найти в работе (Cichocki, Zdunek, Phan, 2009).

Задача неотрицательного матричного разложения ставится как оптимизационная: необходимо найти неотрицательные факторы, доставляющие минимум некоторому функционалу потерь. Выбор этого функционала оказывает существенное влияние на получаемое решение (Wang, Zhang, 2012). В разных прикладных областях для построения неотрицательного матричного разложения используются разные функции потерь: так, в тематическом моделировании используется дивергенция Кульбака-Лейблера (Hofmann, 1999), во многих биологических приложениях — норма Фробениуса (Pascual-Montano et al., 2006), в анализе аудиозаписей — дивергенция Итакура-Саито (Ozerov, Févotte, 2010)), в некоторых задачах машинного зрения — метрика EMD (Sandler, Lindenbaum, 2009). Ясно, что оптимальность той или иной функции потерь в конкретной прикладной задаче зависит от структуры шума, содержащегося в данных, однако часто модель шума в явном виде не задана.

Вопрос оптимального выбора функционала потерь в литературе практически не рассматривается: как правило, функция потерь считается заданной наперёд. В немногих работах, где поднимается этот вопрос, выбор между различными функциями потерь делается на основе некоторой дополнительной информации, имеющейся о структуре модели. Например, в работе (Févotte, Bertin, Durrieu, 2009) сравниваются результаты использования нормы Фробениуса и дивергенций Кульбака-Лейблера и Итакура-Саито в применении неотрицательного матричного разложения к анализу музыкальных последовательностей. Разложения анализируются с точки зрения интерпретируемости получаемых матриц (ожидается, что восстанавливаемые компоненты будут соответствовать нотам); лучшие результаты показывает дивергенция Итакура-Саито. В работе (Choi, Choi, 2010) выбора оптимального функционала потерь делается в параметрическом семействе α-дивергенций, однако и там выбор делается на основе априорной информации. Применение критериев такого рода, как правило, невозможно в большинстве приложений, поскольку информация об ожидаемой структуре модели недоступна. Универсальных методов выбора функционала потерь в задаче неотрицательного

матричного разложения, не требующих дополнительной информации о структуре истинной модели, на настоящий момент не существует.

В данной работе рассматриваются неотрицательные матричные разложения с использованием в качестве функции потерь семейства АБ-дивергенций, являющегося одним из наиболее обширных известных на сегодняшний день параметрических семейств функционалов потерь и включающего многие широко применяемые меры близости, оптимальные в условиях шума самой разной структуры. Данное семейство вместе с мультипликативным алгоритмом получения разложения были предложены в работе (Cichocki, Cruces, Amari, 2011). Однако предложенный алгоритм не имеет теоретических гарантий сходимости; более того, нетрудно показать, что он может сходиться к нестационарным точкам на границе области неотрицательности параметров. В то же время для нормы Фробениуса были получены более сильные результаты: предложен ε-модифицированный мультипликативный алгоритм, любая предельная точка которого является стационарной точкой отделённой от нуля задачи, и показано, что эта точка близка к стационарной точке исходной задачи (Gillis, 2011). Для других функций потерь аналогичные результаты отсутствуют.

Одна из интересных прикладных задач неотрицательного матричного разложения — задача оценивания экспрессии генов по данным ДНК-микрочипов. Используя неотрицательное матричное разложение, можно построить новые модели таких данных, учитывающие не рассматриваемые в стандартных моделях эффекты альтернативного сплайсинга и кроссгибридизации. В то же время структура экспериментов с ДНК-микрочипами достаточно сложна, что не позволяет явно задать модель шума; в связи с этим вопрос оптимального выбора функционала потерь, с помощью которого будет оцениваться качество моделей, остаётся открытым.

Цель диссертационной работы — разработка метода неотрицательного матричного разложения с адаптивным выбором функции потерь и гарантией сходимости, а также создание на его основе новых моделей и методов оценивания экспрессии генов по данным ДНК-микрочипов.

Методы исследования. Задача выбора функции потерь была сведена к задаче подбора параметров АБ-дивергенций — обширного семейства, включающего многие широко при-

меняемые функционалы. Для решения последней применялся метод согласования вклада. Для получения неотрицательного матричного разложения с фиксированным функционалом потерь использовался мультипликативный блочно-покоординатный алгоритм.

Основные положения, выносимые на защиту:

1. Метод адаптивного выбора функционала потерь в задаче неотрицательного матричного разложения, основанный на согласовании вклада.

2. Метод получения неотрицательного матричного разложения с АБ-дивергенцией в качестве функции потерь, доказательство его глобальной сходимости к точке, сколь угодно близкой к стационарной.

3. Модели данных экспериментов с ДНК-микрочипами, учитывающие коэффициенты сродства, эффекты альтернативного сплайсинга и кросс-гибридизации, настроенные с помощью метода адаптивного выбора функционала потерь, а также комплекс программ, получающий оценки экспрессии генов на основе этих моделей.

Научная новизна настоящей диссертации заключается в разработке нового подхода к задаче неотрицательного матричного разложения, основанного на адаптивном выборе функции потерь из семейства АБ-дивергенций; разработке мультипликативного алгоритма неотрицательного матричного разложения и получении ряда теоретических результатов о его сходимости; применении предложенного подхода к задаче анализа данных ДНК-микрочипов, в рамках которой рассматриваются три новых модели, учитывающие ряд не рассматривавшихся ранее особенностей данных.

Теоретическая значимость. В работе впервые предложен универсальный метод выбора функции потерь в задаче неотрицательного матричного разложения; предложен алгоритм разложения и показано, что гарантирована его глобальная сходимость к точке, близкой к стационарной.

Практическая значимость. Полученные результаты позволяют при решении прикладных задач неотрицательного матричного разложения адаптивно определять функцио-

нал потерь, оптимальный для имеющихся данных. Предложенные модели оценивания экспрессии генов по данным ДНК-микрочипов позволяют учитывать эффекты альтернативного сплайсинга и кросс-гибридизации, ранее в литературе не рассматривавшиеся. Реализованный программный комплекс позволяет использовать результаты настройки моделей для получения более точных оценок экспрессии.

Степень достоверности. Достоверность результатов обеспечивается доказательствами теорем и описаниями проведённых экспериментов, допускающими их воспроизводимость.

Апробация работы. Результаты работы докладывались на научных семинарах и конференциях:

- всероссийская конференция «Математические методы распознавания образов» ММРО-15, Петрозаводск, 11–17 сентября 2011 г. [9, 8];
- международная конференция «International Conference on Bioinformatics and Biomedical Engineering» ICBBE, Шанхай, 17-20 мая 2012 г. [6];
- совместный семинар Независимого Московского университета и Московского физико-технического института «Стохастический анализ в задачах»;
- семинары отделов интеллектуальных систем и прикладных проблем оптимизации Вычислительного центра им. А. А. Дородницына Российской академии наук.

Публикации по теме диссертации в изданиях списка ВАК: [2, 3, 4, 5, 7]. Другие публикации по теме диссертации: [1, 6, 8, 9] Отдельные результаты включались в отчёты по проектам РФФИ №12-07-31200, №11-07-00480, министерства образования и науки (ГК № 16.522.11.2004) и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Личный вклад диссертанта в работы, выполненные с соавторами, заключается в следующем:

- в работе [6] предложены критерии качества моделей ДНК-микрочипов, основанные на данных эксперимента со смесями РНК;
- в работах [8, 9] проведены вычислительные эксперименты для определения минимальной значимой комплементарности нуклеотидных последовательностей в модели данных ДНК-микрочипов, учитывающей кросс-гибридизацию;
- в работах [2, 4, 5, 7] модели ДНК-микрочипов применены для получения оценок экспрессии в проводимых экспериментах.

Структура и объём работы. Работа состоит из оглавления, введения, трёх глав, заключения, списка иллюстраций, списка таблиц и списка литературы. Общий объём работы составляет 101 стр.

Краткое содержание работы

В автореферате сохранена нумерация основных утверждений (определений, лемм, теорем), принятая в тексте работы. Нумерация формул сквозная.

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируется цель работы, её теоретическая и практическая значимость, приводится список положений, выносимых на защиту.

В **первой главе** рассматривается задача неотрицательного матричного разложения. В **разделе 1.1** приводится общая постановка задачи и описываются её особенности, в частности, неединственность решения. Рассматриваются условия единственности с точностью до перестановок столбцов и строк матриц-факторов и их нормировки. В **разделе 1.2** приводится постановка оптимизационной задачи неотрицательного матричного разложения. Дана матрица P размера $m \times n$ с неотрицательными элементами и некоторое натуральное число $r < \min(m, n)$. Требуется найти матрицы A^*, X^* размеров $m \times r$ и $r \times n$ соответственно, такие, что

$$(A^*, X^*) = \underset{A \geqslant 0,\, X \geqslant 0}{\operatorname{argmin}} D(P, AX). \tag{1}$$

Далее рассматриваются функционалы потерь $D(P, AX)$, используемые в данной оптимизационной задаче. Вводится класс АБ-дивергенций (Cichocki, Cruces, Amari, 2011), задаваемый

в виде двупараметрического семейства функционалов следующего вида:

$$D_{AB}^{(\alpha,\beta)}(P,Q) = \sum_{i,j} d_{AB}^{(\alpha,\beta)}(p_{ij}, q_{ij}),$$

$$d_{AB}^{(\alpha,\beta)}(p,q) = \begin{cases} \frac{1}{\alpha\beta}\left(\frac{\alpha}{\alpha+\beta}p^{\alpha+\beta} + \frac{\beta}{\alpha+\beta}q^{\alpha+\beta} - p^{\alpha}q^{\beta}\right), & \alpha,\beta,\alpha+\beta \neq 0, \\ \frac{1}{\alpha^2}\left(p^{\alpha}\ln\frac{p^{\alpha}}{q^{\alpha}} - p^{\alpha} + q^{\alpha}\right), & \alpha \neq 0,\ \beta = 0, \\ \frac{1}{\alpha^2}\left(\ln\frac{q^{\alpha}}{p^{\alpha}} + \left(\frac{q^{\alpha}}{p^{\alpha}}\right)^{-1} - 1\right), & \alpha = -\beta \neq 0, \\ \frac{1}{\beta^2}\left(q^{\beta}\ln\frac{q^{\beta}}{p^{\beta}} - q^{\beta} + p^{\beta}\right), & \alpha = 0,\ \beta \neq 0, \\ \frac{1}{2}\left(\ln p - \ln q\right)^2, & \alpha = \beta = 0. \end{cases}$$

Рассматриваются свойства АБ-дивергенций: влияние параметров α и β на получаемые оценки, условия выпуклости функции $d_{AB}^{(\alpha,\beta)}(p,q)$ по q. Приводятся значения параметров, при которых АБ-дивергенция задаёт некоторые широко используемые функционалы потерь.

В **разделе 1.3** задача оптимального выбора функции потерь сводится к выбору параметров α и β АБ-дивергенции. Для всех α и β множеству АБ-дивергенций ставится в соответствие обобщённое семейство распределений, задаваемых следующим образом:

$$\begin{aligned} p(P,\alpha,\beta) &= \frac{1}{Z(\alpha,\beta)} p_0(P,\alpha,\beta), \\ p_0(P,\alpha,\beta) &= e^{-D_{AB}^{(\alpha,\beta)}(P,Q)}, \\ Z(\alpha,\beta) &= \int_X p_0(X,\alpha,\beta)\, dX. \end{aligned} \tag{2}$$

Поскольку нормировочный множитель $Z(\alpha,\beta)$ неизвестен и может даже не существовать, для оценки оптимальных значений параметров α и β нельзя применить метод максимального правдоподобия. Вместо этого предлагается использовать метод согласования вклада (Hyvärinen, 2006), для которого достаточно знать только $p_0(P,\alpha,\beta)$. Максимизация логарифма правдоподобия соответствует минимизации дивергенции Кульбака-Лейблера между истинной и модельной плотностями распределения данных. Аналогично, согласование вклада эквивалентно минимизации дивергенции Фишера между ними. Следующая теорема приводит к методу получения оптимальных значений параметров АБ-дивергенции в задаче неотрицательного матричного разложения.

Теорема 1. *Оценка согласования вклада в модели* (2) *определяется следующим выражени-*

ем:

$$(\alpha^*, \beta^*) = \underset{\alpha,\beta}{argmin}\, J(P, \alpha, \beta),$$

$$J(P, \alpha, \beta) = \begin{cases} \frac{1}{\beta} \sum\limits_{i,j} p_{ij}^{\alpha} \left(\frac{1}{2\beta} p_{ij}^{\alpha} \left(p_{ij}^{\beta} - q_{ij}^{\beta} \right)^2 - p_{ij}^{\beta} (\alpha + \beta + 1) + q_{ij}^{\beta} (\alpha + 1) \right), & \beta \neq 0, \\ \sum\limits_{i,j} p_{ij}^{\alpha} \left(\ln \frac{q_{ij}}{p_{ij}} \left(\frac{p_{ij}^{\alpha}}{2} \ln \frac{q_{ij}}{p_{ij}} + \alpha + 1 \right) - 1 \right), & \beta = 0. \end{cases}$$

В **разделе 1.4** рассматривается метод получения неотрицательного матричного разложения с АБ-дивергенцией при фиксированных значениях параметров α и β. Поскольку оптимизационная задача (1) не является выпуклой по совокупности аргументов A и X, как правило, используются методы поочерёдной минимизации $D(P, AX)$ по X при фиксированном A и наоборот. При использовании блочно-покоординатного спуска ограничения неотрицательности задачи (1) можно сохранять естественным образом за счёт такого выбора шага в направлении градиента, чтобы обновления стали мультипликативными. Для АБ-дивергенции при $\alpha \neq 0$ обновления мультипликативного алгоритма записываются следующим образом:

$$\begin{aligned} X &\leftarrow X \otimes \left(\left(A^T Z \right) \oslash \left(A^T Q^{[\alpha+\beta-1]} \right) \right)^{[\frac{1}{\alpha}]}, \\ A &\leftarrow A \otimes \left(\left(Z X^T \right) \oslash \left(Q^{[\alpha+\beta-1]} X^T \right) \right)^{[\frac{1}{\alpha}]}, \end{aligned} \tag{3}$$

где символом $\otimes$ обозначается поэлементное произведение (произведение Адамара) двух матриц, символом $\oslash$ — операция поэлементного деления матриц, $^{[.]}$ — поэлементное возведение в степень, а $Z = P^{[\alpha]} \otimes Q^{[\beta-1]}$.

В **разделе 1.5** рассматриваются вопросы сходимости данного мультипликативного алгоритма. Поскольку решаемая задача не является выпуклой, лучшее, что можно гарантировать — это сходимость к стационарной точке, задаваемой условиями Каруша-Куна-Таккера. Оказывается, мультипликативный алгоритм (3) не обеспечивает сходимости к стационарной точке и может останавливаться вблизи границы области неотрицательности. В **разделе 1.5.1** рассматриваются известные результаты для нормы Фробениуса, соответствующей случаю $\alpha = \beta = 1$:

$$(A^*, X^*) = \underset{A \geqslant 0,\, X \geqslant 0}{\operatorname{argmin}} \|P - AX\|_F^2. \tag{4}$$

Определение 2. *Для любого $\varepsilon > 0$ ε-модификацией итерационного алгоритма с обновлениями $x \leftarrow f(x)$ назовём алгоритм с обновлениями $x \leftarrow \max(\varepsilon, f(x))$.*

Для нормы Фробениуса ε-модификация мультипликативного алгоритма имеет вид

$$\begin{aligned} X &\leftarrow \max\left(\varepsilon, X \otimes \left(\left(A^T P\right) \oslash \left(A^T A X\right)\right)\right), \\ A &\leftarrow \max\left(\varepsilon, A \otimes \left(\left(P X^T\right) \oslash \left(A X X^T\right)\right)\right); \end{aligned} \tag{5}$$

и обладает следующими свойствами (Gillis, 2011): в ходе обновлений функционал потерь монотонно невозрастает; любая предельная точка является стационарной точкой оптимизационной задачи, переформулированной в отделённой от нуля области $A \geqslant \varepsilon$, $X \geqslant \varepsilon$.

Определение 3. *Назовём ε-прореживанием операцию обнуления элементов матрицы, в точности равных ε.*

Показано, что матрицы, полученные ε-прореживанием предельной точки алгоритма (5), близки к стационарной точке исходной оптимизационной задачи (4): условия Каруша-Куна-Таккера для них выполняются с точностью до $\mathcal{O}(\varepsilon)$. В диссертационной работе получено следующее усиление утверждения о сходимости ε-модифицированного алгоритма (5).

Лемма 1. *Мультипликативный ε-модифицированный алгоритм с обновлениями* (5) *сходится к стационарной точке отделённой от нуля задачи*

$$(A_\varepsilon^*, X_\varepsilon^*) = \underset{A \geqslant \varepsilon,\, X \geqslant \varepsilon}{argmin} \|P - AX\|_F^2.$$

В **разделе 1.5.2** рассматривается сходимость мультипликативного алгоритма неотрицательного матричного разложения с АБ-дивергенцией, решающего задачу

$$(A^*, X^*) = \underset{A \geqslant 0,\, X \geqslant 0}{\operatorname{argmin}} D_{AB}^{(\alpha,\beta)}(P, AX) \tag{6}$$

при произвольных α и β. Для монотонного убывания функции потерь необходимо модифицировать показатель степени обновлений (3):

$$\begin{aligned} X &\leftarrow X \otimes \left(\left(A^T Z\right) \oslash \left(A^T Q^{[\alpha+\beta-1]}\right)\right)^{[\omega'(\alpha,\beta)]}, \\ A &\leftarrow A \otimes \left(\left(Z X^T\right) \oslash \left(Q^{[\alpha+\beta-1]} X^T\right)\right)^{[\omega'(\alpha,\beta)]}, \\ \omega'(\alpha,\beta) &= \begin{cases} \frac{1}{1-\beta}, & \frac{\beta}{\alpha} < \frac{1}{\alpha} - 1, \\ \frac{1}{\alpha}, & \frac{\beta}{\alpha} \in \left[\frac{1}{\alpha} - 1, \frac{1}{\alpha}\right], \\ \frac{1}{\alpha+\beta-1}, & \frac{\beta}{\alpha} > \frac{1}{\alpha}. \end{cases} \end{aligned}$$

Рассмотрим ε-модификацию этого алгоритма:

$$\begin{aligned} X &\leftarrow \max\left(\varepsilon, X \otimes \left(\left(A^T Z\right) \oslash \left(A^T Q^{[\alpha+\beta-1]}\right)\right)^{[\omega'(\alpha,\beta)]}\right), \\ A &\leftarrow \max\left(\varepsilon, A \otimes \left(\left(Z X^T\right) \oslash \left(Q^{[\alpha+\beta-1]} X^T\right)\right)^{[\omega'(\alpha,\beta)]}\right). \end{aligned} \tag{7}$$

Теорема 4. *Для любого $\varepsilon > 0$ функционал $D_{AB}^{(\alpha,\beta)}(P, AX)$ монотонно невозрастает при обновлениях ε-модифицированного алгоритма (7) для любого начального приближения $(A^0, X^0) \geqslant \varepsilon$.*

Теорема 5. *Любая предельная точка последовательности, порождаемой алгоритмом с обновлениями вида (7) для любого начального приближения $(A^0, X^0) \geqslant \varepsilon$, является стационарной точкой отделённой от нуля задачи*

$$(A_\varepsilon^*, X_\varepsilon^*) = \underset{A \geqslant \varepsilon,\, X \geqslant \varepsilon}{argmin}\, D_{AB}^{(\alpha,\beta)}(P, AX). \tag{8}$$

Теорема 6. *ε-модифицированный мультипликативный алгоритм с обновлениями (7) сходится к стационарной точке отделённой от нуля задачи (8).*

Поскольку норма Фробениуса — частный случай АБ-дивергенции, последняя теорема включает в себя утверждение леммы 1.

Пусть $(A_\varepsilon, X_\varepsilon)$ — предельная точка ε-модифицированного алгоритма с обновлениями (7). Проведём ε-прореживание матриц A_ε и X_ε:

$$\begin{aligned} A_0 &= A_\varepsilon \otimes [A_\varepsilon > \varepsilon], \\ X_0 &= X_\varepsilon \otimes [X_\varepsilon > \varepsilon]. \end{aligned}$$

Теорема 7. *Для матриц (A_0, X_0), полученных из $(A_\varepsilon, X_\varepsilon)$ ε-прореживанием, верно следующее:*

$$\begin{cases} \left[\begin{array}{ll} a_{0ik} = 0, & \left[\nabla_A D_{AB}^{(\alpha,\beta)}(P, A_0 X_0)\right]_{ik} \geqslant -\mathcal{O}(\varepsilon), \\ a_{0ik} > 0, & \left|\left[\nabla_A D_{AB}^{(\alpha,\beta)}(P, A_0 X_0)\right]_{ik}\right| \leqslant \mathcal{O}(\varepsilon), \end{array}\right. \\ \left[\begin{array}{ll} x_{0kj} = 0, & \left[\nabla_X D_{AB}^{(\alpha,\beta)}(P, A_0 X_0)\right]_{kj} \geqslant -\mathcal{O}(\varepsilon), \\ x_{0kj} > 0, & \left|\left[\nabla_X D_{AB}^{(\alpha,\beta)}(P, A_0 X_0)\right]_{kj}\right| \leqslant \mathcal{O}(\varepsilon), \end{array}\right. \end{cases}$$

то есть, в точке (A_0, X_0) *условия Каруша-Куна-Таккера для исходной задачи* (6) *выполняются с точностью до* $\mathcal{O}(\varepsilon)$.

Таким образом, про предложенный ε-модифицированный мультипликативный алгоритм (7) неотрицательного матричного разложения с произвольной АБ-дивергенцией в качестве функционала потерь показано, что он всегда сходится, а его предельная точка является стационарной точкой отделённой от нуля задачи (8) и лежит сколь угодно близко к некоторой стационарной точке исходной задачи (6).

В **разделе 1.6** рассматриваются практические особенности решения оптимизационной задачи: получение начального приближения, критерий останова, а также метод обработки пропусков и выбросов.

Во **второй главе** рассматривается применение описанных в первой главе методов неотрицательного матричного разложения к задаче оценки экспрессии генов по данным экспериментов с ДНК-микрочипами. В **разделе 2.1** приводится постановка задачи с точки зрения предметной области и краткое описание существующих методов её решения. ДНК-микрочип представляет собой многомерный сенсор для одновременного измерения экспрессии десятков тысяч генов. Разным участкам каждого гена соответствует несколько десятков проб на поверхности микрочипа, интенсивность флуоресценции которых пропорциональна уровню экспрессии данного гена в исследуемом образце. Стандартные методы анализа сводятся к неотрицательному матричному разложению ранга $r = 1$:

$$I_{pk} \approx \hat{I}_{pk} = a_p c_{g(p)k}. \tag{9}$$

где $k = 1, \ldots, K$ — номер микрочипа, $p = 1, \ldots, P$ — номер пробы, $g = 1, \ldots, G$ — номер гена, $g(p)$ — номер гена, соответствующего пробе p, I_{pk} — интенсивность флуоресценции пробы p на микрочипе k, $c_{g(p)k}$ — уровень экспрессии гена g, которому проба p комплементарна, на микрочипе k, a_p — коэффициент сродства пробы p своему гену. При этом распределение шума на микрочипах неизвестно и плохо описывается стандартными распределениями. В **разделе 2.2** описываются методы получения и предварительной обработки данных.

В **разделе 2.3** рассматривается настройка модели (9) по выборке микрочипов, описанной в 2.2, с использованием метода адаптивного неотрицательного матричного разложения. Предлагаются функционалы для оценки качества моделей. Приводятся результаты численных экспериментов. Показывается, что оптимальная функция потерь соответствует модели

шума, несколько отличающейся от традиционной логнормальной.

В **разделе 2.4** предлагается модель, учитывающая эффект альтернативного сплайсинга, в результате которого некоторые участки генов могут отсутствовать, а соответствующие им пробы могут вызывать занижение оценок экспрессии. Эффект предлагается учитывать с помощью бинарной матрицы весов $W \in \{0,1\}^{P\times K}$, заполненной следующим образом:

$$e_{pk} = \frac{\hat{I}_{pk} - I_{pk}}{I_{pk}} \cdot c_{g(p)k},$$

$$w_{pk} = \begin{cases} 1, & e_{pk} < e_{0.95}, \\ 0, & e_{pk} \geqslant e_{0.95}. \end{cases}$$

Здесь $e_{0.95}$ — 95% выборочный квантиль e_{pk}. Полученные веса встраиваются в обновления мультипликативного алгоритма неотрицательного матричного разложения:

$$X \leftarrow \max\left(\varepsilon, X \otimes \left(\left(A^T\left(Z \otimes W\right)\right) \oslash \left(A^T\left(Q^{[\alpha+\beta-1]} \otimes W\right)\right)\right)^{[\omega'(\alpha,\beta)]}\right),$$
$$A \leftarrow \max\left(\varepsilon, A \otimes \left(\left(\left(Z \otimes W\right)X^T\right) \oslash \left(\left(Q^{[\alpha+\beta-1]} \otimes W\right)X^T\right)\right)^{[\omega'(\alpha,\beta)]}\right).$$

Процесс настройки модели с учётом весов и их переопределение повторяется несколько раз. Приводятся результаты численных экспериментов.

В **разделе 2.5** предлагается модель, учитывающая эффект кросс-гибридизации, в результате которого флуоресценция пробы может быть вызвана экспрессией неспецифических генов, частично комплементарных пробе. Задача настройки такой модели сводится к получению неотрицательного матричного разложения ранга G:

$$I_{pk} \approx \hat{I}_{pk} = \sum_{g=1}^{G} a_{pg}c_{gk}.$$

Для уменьшения числа параметров перед настройкой модели производится сравнение последовательностей рассматриваемых проб и генов и создаётся бинарная матрица весов $W^A \in \{0,1\}^{P\times G}$, нули в которой соответствуют парам проба-ген, с большой вероятностью не вступающим в реакцию кросс-гибридизации. Полученная матрица была встроена в обновления A мультипликативного алгоритма неотрицательного матричного разложения:

$$A \leftarrow W^A \otimes \max\left(\varepsilon, A \otimes \left(\left(ZX^T\right) \oslash \left(Q^{[\alpha+\beta-1]}X^T\right)\right)^{[\omega'(\alpha,\beta)]}\right).$$

В результате на каждой итерации алгоритма элементы, соответствующие существенно различным пробе и гену, обнуляются. В рассматриваемой модели выполняются условия единственности неотрицательного матричного разложения, описанные в разделе 1.1: матрица A содержит диагональную подматрицу, поскольку для каждого гена существует хотя бы одна проба, комплементарная ему. Приводятся результаты численных экспериментов.

В **третьей главе** описывается комплекс программ, реализующий рассматривавшиеся в работе алгоритмы неотрицательного матричного разложения с функционалом потерь из семейства АБ-дивергенций, предназначенный для обработки данных экспериментов с ДНК-микрочипами с использованием предложенных в главе 2 моделей. Программный комплекс состоит из следующих частей:

- модуль неотрицательного матричного разложения с фиксированным функционалом потерь ABNMFFixed;
- модуль адаптивного неотрицательного матричного разложения ABNMFAdaptive;
- модуль чтения и предобработки данных экспериментов с ДНК-микрочипами PreprocessArrays;
- модуль настройки параметров моделей, описанных в разделах 2.3, 2.4 и 2.5 TuneModel;
- модуль оценки экспрессии генов на основании настроенных моделей EstimateExpression.

Заключение

Основные результаты данной работы заключаются в следующем.

1. Предложен метод адаптивного выбора функции потерь в задаче неотрицательного матричного разложения из семейства АБ-дивергенций, основанный на согласовании вклада.
2. Предложен ε-модифицированный мультипликативный алгоритм неотрицательного матричного разложения с АБ-дивергенцией в качестве функции потерь; доказана его глобальная сходимость к стационарной точке отделённой от нуля оптимизационной задачи.

3. Предложен метод ε-прореживания решения ε-модифицированного мультипликативного алгоритма; доказано, что в получаемой с его помощью точке условия стационарности исходной оптимизационной задачи выполняются с точностью до $\mathcal{O}(\varepsilon)$.

4. Предложен ряд моделей данных экспериментов с ДНК-микрочипами, учитывающих коэффициенты сродства, эффекты альтернативного сплайсинга и кросс-гибридизации.

5. Создан программный комплекс обработки данных экспериментов с ДНК-микрочипами, позволяющий получать более точные оценки экспрессии генов по сравнению с существующими аналогами.

Публикации автора по теме диссертации

[1] Рябенко, Е.А. (2014). Мультипликативный метод неотрицательного матричного разложения с АБ-дивергенцией и его сходимость. Машинное обучение и анализ данных, 1(7), 800–816.

[2] **Крайнова, Н.А., Хаустова, Н.А., Макеева, Д.С., Федотов, Н.Н., Гудим, Е.А., Рябенко, Е.А., Шкурников, М.Ю., Галатенко, В.В., Сахаров, Д.А., Мальцева, Д.В. (2013). Оценка потенциальных референсных генов для нормализации данных ПЦР-РВ в экспериментах с клетками линии HeLa. Биотехнология, 1, 42–50.**

[3] **Рябенко, Е.А. (2012). Настройка нелинейной модели данных экспериментов с экспрессионными ДНК-микрочипами. Математическая биология и биоинформатика, 7(2), 554–566.**

[4] **Sakharov, D.A., Maltseva, D.V, Riabenko, E.A., Shkurnikov, M.U., Northoff, H., Tonevitsky, A.G., Grigoriev, A.I. (2012). Passing the anaerobic threshold is associated with substantial changes in the gene expression profile in white blood cells. European journal of applied physiology, 112(3), 963–972.**

[5] **Мальцева, Д.В., Рябенко, Е.А., Сизова, С.В., Яшин, Д.В., Хаустова, С.А., Шкурников, М.Ю. (2012). Влияние физической нагрузки на экспрессию ге-**

нов HSPBP1, PGLYRP1 и HSPA1A в лейкоцитах человека. Бюллетень экспериментальной биологии и медицины, 153(6), 846–850.

[6] Riabenko, E.A., Kogadeeva, M., Gavrilyuk, K., Sokolov, E., Shanin, I., Tonevitsky, A.G. (2012). Comparing Affymetrix Human Gene 1.0 ST preprocessing methods on tissue mixture data. 6th International Conference on Bioinformatics and Biomedical Engineering (iCBBE) (pp. 631–634). Shanghai, China.

[7] **Riabenko, E.A., Tonevitsky, E.A., Tonevitsky, A.G., Grigoriev, A.I. (2011). Structural Peculiarities of Human Genes Which Expression Increases in Response to Stress. American Journal of Biomedical Sciences, 3(2), 90–94.**

[8] Рябенко, Е.А., Когадеева, М.С. (2011). Нижняя граница числа комплементарных нуклеотидов при моделировании кросс-гибридизации. Математические методы распознавания образов: 15-я Всероссийская конференция, г.Петрозаводск, 11–17 сентября 2011 г.: Сборник докладов. (сс. 540–542). Петрозаводск: МАКС Пресс.

[9] Когадеева, М.С., Рябенко, Е.А. (2011). Математическая модель данных микрочипов ДНК, учитывающая эффекты кросс-гибридизации и насыщения. Математические методы распознавания образов: 15-я Всероссийская конференция, г.Петрозаводск, 11–17 сентября 2011 г.: Сборник докладов. (сс. 536–539). Петрозаводск: МАКС Пресс.