ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2020, том 60, № 3, с. 369–378

УДК 519.26

ПРОБЛЕМЫ УСТОЙЧИВОСТИ И ЕДИНСТВЕННОСТИ СТОХАСТИЧЕСКОГО МАТРИЧНОГО РАЗЛОЖЕНИЯ[1)]

© 2020 г. Р. Ю. Дербаносов[1,*], И. А. Ирхин[2,]**

[1] *125319 Москва, ул. Кочновский проезд, 3, НИУ ВШЭ, Россия*

[2] *141701 Долгопрудный, М.о., Институтский пер., 9, МФТИ, Россия*

**e-mail: derbanosov@gmail.com*

***e-mail: ilirhin@gmail.com*

Поступила в редакцию 12.09.2018 г.
Переработанный вариант 05.10.2018 г.
Принята к публикации 18.11.2019 г.

Рассматриваются две близкие проблемы: устойчивости решения задачи тематического моделирования и единственности стохастического матричного разложения. Доказана теорема, описывающая аналитический способ понять по поставленной задаче стохастического матричного разложения, будет ли ее решение устойчивым. Применимость теоремы на практике исследуется в экспериментах на реальных данных. Библ. 21. Фиг. 1.

Ключевые слова: тематическое моделирование, неотрицательное матричное разложение, единственность матричного разложения.

DOI: 10.31857/S0044466920030084

1. ВВЕДЕНИЕ

Тематическое моделирование – один из популярных статистических методов анализа текстов. Результатом построения модели является набор тем, а также описание каждого документа в виде распределения над множеством тем. Тематическое моделирование применяют в задачах анализа аудио [1], текстов [2]–[5], изображений и видео [6]–[8], биоинформатике [9], [10], в задачах информационного поиска [11]–[13]. Тематическое моделирование может быть использовано для получения интерпретируемых векторных представлений слов, демонстрирующих сравнимое с векторными представлениями модели Skip-Gram Negative Sampling [14] качество на задачах сравнения семантически близких слов [15].

Сформулируем основную задачу тематического моделирования. *Темой* будем называть дискретное распределение над фиксированным множеством слов. Пусть известны параметры модели: *матрица слова-темы* Φ размера число слов на число тем, в которой по столбцам записаны распределения слов в темах, j-му столбцу матрицы Φ отвечает j-я тема, *матрица темы-документы* Θ размера число тем на число документов, в j-м столбце которой записано распределение тем для j-го документа, элемент Θ_{ij} равен весу i-й темы для j-го документа, а также для d-го документа известно числов слов n_d в нем. Предположим, что текстовая коллекция была получена в результате следующего процесса. Для d-го документа для каждого из n_d слов выбирается тема t согласно распределению, записанному в d-м столбце матрицы Θ, после чего выбирается очередное слово из распределения, записанного в t-м столбце матрицы Φ. Задача тематического моделирования заключается в восстановлении матриц Φ и Θ по таким образом сгенерированной текстовой коллекции. Обычно текстовая коллекция представляется в виде *матрицы слова-документы* F, элемент которой F_{ij} равен числу раз, которое i-е слово входит в j-й документ.

Современные методы тематического моделирования основаны на поиске представления матрицы слова-документы, нормированной по столбцам, в виде произведения стохастических мат-

[1)]Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 17-07-01536).

риц Φ и Θ. Наиболее популярным подходом к построению тематических моделей является латентное размещение Дирихле (LDA) [2]. Модель LDA основана на предположении о том, что вектора распределений слов в темах и тем в документах порождены распределением Дирихле. Другим подходом является вероятностный латентный семантический анализ (PLSA) [16]. Основной гипотезой данной модели являются гипотеза условной независимости: вероятность слова в теме не зависит от документа. Развитием модели PLSA является аддитивная регуляризация тематических моделей (ARTM) [3], [5], [17]. Модель ARTM расширяет постановку задачи PLSA, добавляя к функции потерь различные регуляризаторы, формализующие требования к тематической модели.

Одной из проблем тематического моделирования является то, что точного разложения $F = \Phi\Theta$, как правило, не существует, поэтому строится разложение некоторого приближения изначальной матрицы $\tilde{F}$, которое является локальным экстремумом задачи максимизации правдоподобия. Таким образом, неединственность решения задачи тематического моделирования может возникать как из-за неоднозначности выбора $\tilde{F}$, так и из-за неединственности точного разложения $\tilde{F}$ на Φ и Θ.

Проблема единственности неотрицательного матричного разложения исследовалась в работах [18]–[20], где, в частности, представлены различные достаточные либо необходимые условия единственности разложения. Однако до сих пор не сформулированы необходимые и достаточные условия единственности стохастического матричного разложения.

В первой части данной работы представлен результат, описывающий достаточные условия на матрицу F, гарантирующие единственность точного стохастического разложения полного ранга $F = \Phi\Theta$, $F \in \mathbb{R}^{n\times m}$, $\operatorname{rank} F = k$, $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$ в терминах элементов матриц Φ и Θ, а также простое следствие этого результата, дающее критерий существования и единственности в терминах матрицы F, без заданного разложения.

Во второй части дается интерпретация теоремы из первой части работы в терминах тематического моделирования, а также исследуется выполнение условий теоремы в экспериментах на реальных данных.

Последующий текст устроен следующим образом. В разд. 2 формулируются задача и определения, требуемые для формулировки основной теоремы, после чего дается обзор предыдущих результатов. В разд. 3 формулируется и доказывается основная теорема. В разд. 4 описываются практические эксперименты, проверяющие выполнение условий основной теоремы на реальной текстовой коллекции.

2. ПОСТАНОВКА ЗАДАЧИ МАТРИЧНОГО РАЗЛОЖЕНИЯ

В этом разделе речь идет о представлении матрицы F в виде произведения двух стохастических матриц полного ранга Φ и Θ: $F = \Phi\Theta$. Обсуждается следующая задача. Пусть дана матрица F, $F \in \mathbb{R}^{n\times m}$, $\operatorname{rank} F = k$, и некоторое ее разложение $F = \Phi\Theta$, $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$. Такое разложение всегда не единственно даже при фиксированных размерах матриц Φ и Θ. Поэтому обычно разложения рассматривают с точностью до добавления в разложение матрицы F перестановки S. В этом случае существуют матрицы F, для которых стохастическое разложение полного ранга единственно.

2.1. Стохастическое матричное разложение

Определение 1. Будем называть матрицу F *неотрицательной*, если $F_{ij} \geq 0\ \forall i, j$.

Определение 2. Будем называть матрицу F *стохастической*, если она является неотрицательной и $\sum_i F_{ij} = 1\ \forall j$.

Определение 3. *Неотрицательным (стохастическим) матричным разложением матрицы* $F \in \mathbb{R}^{n\times m}$ называют представление F в виде произведения двух неотрицательных (*стохастических*) *матриц* $F = \Phi\Theta$.

Определение 4. *Матричным разложением полного ранга матрицы* $F \in \mathbb{R}^{n\times m}$, $\operatorname{rank} F = k$ называют представление F в виде произведения двух матриц полного ранга $F = \Phi\Theta$, $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$.

Далее в этой работе для краткости *разложением матрицы* F называется стохастическое матричное разложение полного ранга, кроме тех случаев, в которых явно сказано, что это стохастическое разложение или неотрицательное разложение. Также будем предполагать, что в матрице F, для которой ищется неотрицательное, стохастическое или стохастическое полноранговое разложение, нет нулевых столбцов.

Из определения разложения матрицы сразу следует, что матрицы Φ и Θ имеют ранг k. Также можно заметить, что если у матрицы есть хотя бы одно разложение, то она является стохастической.

Заметим, что если дано разложение $\Leftrightarrow$ и у матрицы Φ есть хотя бы два различных столбца или у матрицы Θ есть хотя бы две различных строки, то можно найти новое разложение $F = \Phi SS^{-1}\Theta$, где S – матрица перестановки. В связи с этим общепринятым является следующее определение единственности разложения.

Определение 5. Разложение $F = \Phi\Theta$ называется *единственным*, когда выполнено следующее условие: если нашлось некоторое другое разложение $F = \Phi'\Theta'$, то $\Phi' = \Phi S$, $\Theta' = S^{-1}\Theta$ для некоторой матрицы перестановки S.

Введем следующие вспомогательные обозначения:

$\overline{\operatorname{supp}(v)}$ – множество позиций, на которых стоят нулевые элементы вектора v;

$\operatorname{supp}(v)$ – множество позиций, на которых стоят ненулевые элементы вектора v;

X_j – есть j-й столбец матрицы X;

$X[[i_1,\ldots,i_p],[j_1,\ldots,j_q]]$ – подматрица, состоящая из строк $i_1,\ldots,i_p$ и столбцов $j_1,\ldots,j_q$.

2.2. Связь между неотрицательными и стохастическими матричными разложениями

Утверждение 1. *Пусть* $F \in \mathbb{R}^{n\times m}$ *матрица без нулевых столбцов с неотрицательным разложением* $F = \Phi\Theta$, $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$. *Рассмотрим стохастическую матрицу* $\tilde{F}$, $\tilde{F}_{ij} = \frac{F_{ij}}{\sum_i F_{ij}}$. *Тогда* $\tilde{F} = \tilde{\Phi}\tilde{\Theta}$ *является стохастическим разложением матрицы* $\tilde{F}$, *где*

$$\tilde{\Phi} = \Phi S, \quad \tilde{\Theta} = S^{-1}\Theta',$$
$$\Theta' = \frac{\Theta_{ij}}{\sum_i F_{ij}}, \quad S = \operatorname{diag}\left(\left(\sum_i \Phi_{i1}\right)^{-1},\ldots,\left(\sum_i \Phi_{ik}\right)^{-1}\right).$$

Доказательство. Заметим, что матрицы S и Θ' определены корректно, т.е. деления на ноль произойти не может, благодаря предположению о том, что в матрице F нет нулевых столбцов.

Докажем, что $\tilde{F} = \tilde{\Phi}\tilde{\Theta}$ является стохастическим разложением матрицы $\tilde{F}$. Разделив j-й столбец левой и правой части равенства $F = \Phi SS^{-1}\Theta$ на $\sum_i F_{ij}$, мы получаем, что $\tilde{F} = \tilde{\Phi}\tilde{\Theta}$ является верным равенством. Матрица S подобрана таким образом, чтобы матрица $\tilde{\Phi}$ была стохастической, следовательно, верно равенство $e_n^{\mathrm{T}}\tilde{\Phi} = e_k^{\mathrm{T}}$, где $e_p \in \mathbb{R}^p$ – вектор-столбец из единиц. Воспользовавшись стохастичностью матрицы $\tilde{F}$, получаем

$$e_n^{\mathrm{T}}\tilde{F} = e_m^{\mathrm{T}} \Leftrightarrow e_n^{\mathrm{T}}\tilde{\Phi}\tilde{\Theta} = e_m^{\mathrm{T}} \Leftrightarrow e_k^{\mathrm{T}}\tilde{\Theta} = e_m^{\mathrm{T}}.$$

Значит, матрица $\tilde{\Theta}$ тоже стохастическая.

Если же матрица F изначально была стохастической и имела неотрицательное разложение, то, используя описанное выше утверждение, можно построить стохастическое разложение этой матрицы. В связи с этим фактом проблемы неотрицательных и стохастических разложений являются очень близкими между собой.

2.3. Обзор результатов

Во всех работах, исследующих единственность неотрицательного матричного разложения, формулируются необходимые либо достаточные условия единственности разложения.

В работе [18] вводится геометрическая интерпретация стохастического матричного разложения. Для матрицы X обозначим $C_X = \left\{x \mid x = \sum_i a_i X_i,\ a_i \geq 0\right\}$ симплициальный конус, порожденный векторами X_i. Геометрическая интерпретация заключается в сопоставлении каждому стохастическому разложению матрицы $F = \Phi\Theta$ симплициального конуса $G_\Phi \supset G_F$. Также в работе получены достаточные условия на матрицы-факторы (Lemma 4), гарантирующие единственность разложения, при этом достаточные условия выражаются в терминах симплициальных конусов G_Φ и G_F. Аналог этой геометрической интерпретации в терминах выпуклых многогранников активно используется в нашей работе (см. лемму 1).

В работе [19] получены необходимые условия единственности (Theorem 5) и достаточные условия единственности для неотрицательных матричных разложений. На примерах демонстрируется выполнение условий теорем и их ограничения. В экспериментах показывается, что в случае единственности разложения $F = \Phi\Theta$ добавление небольшого шума к исходным данным F влечет сходимость к тому же единственному решению (Φ, Θ) с небольшим шумом.

В работе [20] используется геометрическая интерпретация, введенная в работе [18], и доказывается достаточный критерий единственности (Theorem 6), при этом условия даются в терминах матрицы F. Описывается техника предобработки данных, приводящая к устойчивости и разреженности получаемой тематической модели. Демонстрируется эффективность техники на нескольких наборах изображений.

3. ТЕОРЕМА О ЕДИНСТВЕННОСТИ РАЗЛОЖЕНИЯ

3.1. Условия основной теоремы

Теорема 1. *Пусть дано разложение* $F = \Phi\Theta$, $F \in \mathbb{R}^{n\times m}$, $\operatorname{rank} F = k$, $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$. *Пусть выполнены следующие условия.*

Условие 1. $\forall i \in \{1,\dots,k\}\ \exists j : \Theta_{ij} = 1,\ \forall i' \neq j \Theta_{i'j} = 0.$

Условие 2. $\forall j\ \operatorname{rank}(\Phi[\overline{\operatorname{supp}(\Phi_j)}, [1,\dots,k]\backslash[j]]) = k-1.$

Тогда разложение $F = \Phi\Theta$ *единственно.*

Первое условие требует наличия в матрице Θ k столбцов, из которых можно составить единичную матрицу $k \times k$.

Второе условие требует, чтобы для каждого столбца Φ_j матрицы Φ подматрица, соответствующая множеству строк, на которых в Φ_j стоят нули, имела ранг $k-1$. Тривиальным примером матрицы Φ, которая удовлетворяет этому условию, является матрица, из k строк которой можно составить единичную матрицу размера $k \times k$.

Простым следствием из этой теоремы является достаточное условие для единственности разложения стохастической матрицы F.

Следствие 1. *Пусть у стохастической матрицы* $F \in \mathbb{R}^{n\times m}$ *ранга* k *нашлось* k *таких столбцов с номерами* $\{j_1,\dots,j_k\} := J$, *что*

(I) $\forall j \in J\ \operatorname{rank}(F[\overline{\operatorname{supp}(F_j)}, J\backslash[j]]) = k-1,$

(II) *для любого* $j \in J$, $p = 1,\dots,m$, *найдутся коэффициенты* $a_{jp} \geq 0$ *т.ч.* $\sum_{j\in J} a_{jp} = 1$, $F_p = \sum_{j\in J} a_{jp} F_j,\ \forall p.$

Тогда у F *существует единственное разложение* $F = \Phi\Theta$, *где*

$$\Phi = F[:, J],$$
$$\Theta[j, p] = a_{jp}.$$

Доказательство. Действительно, если выполнено условие 1, то k столбцов $F_{j_1},\ldots,F_{j_k}$ можно взять в качестве матрицы Φ, подходящей под условия теоремы 1. Матрица Θ, дающая разложение $F = \Phi\Theta$, найдется благодаря условию 2.

Сравним это следствие с теоремой единственности, сформулированной в работе [20]. *Паттерном разреженности* (англ. sparsity pattern) вектора v называется множество $\{i|v_i = 0\}$. Например, паттерн разреженности вектора $(4,0,0,2,0)$ есть $\{1,2,4\}$. *Неотрицательным рангом* матрицы F назовем такое минимальное r, что существует неотрицательное разложение матрицы $F = \Phi\Theta$, $\Phi \in \mathbb{R}^{n\times r}$, $\Theta \in \mathbb{R}^{r\times m}$.

Теорема 2 ([20, теорема 6]). *Пусть дана стохастическая матрица F с рангом, совпадающим с неотрицательным рангом и равным k. Если у матрицы F есть k ненулевых столбцов, каждый из которых имеет $k-1$ нулей таких, что в соответствующих этим нулям строкам паттерны разреженности различны, то матрица F имеет единственное разложение.*

Ниже приведен пример матрицы F, которая имеет единственное разложение, удовлетворяет условиям следствия теоремы 1, описанного выше, но не удовлетворяет условиям теоремы 6 статьи [20].

Пример 1.

$$F = \begin{pmatrix} 0 & \frac{1}{6} & \frac{2}{6} \\ 0 & \frac{2}{6} & \frac{1}{6} \\ \frac{1}{6} & 0 & \frac{2}{6} \\ \frac{2}{6} & 0 & \frac{1}{6} \\ \frac{1}{6} & \frac{2}{6} & 0 \\ \frac{2}{6} & \frac{1}{6} & 0 \end{pmatrix}.$$

У этой матрицы F есть единственное разложение $F = FE$, но условия теоремы [20, Theorem 6] не выполнены, потому что для каждого столбца у множества строк, в которых этот столбец зануляется, одинаковые паттерны разреженности. При этом легко понять, что условия следствия теоремы 1 выполнены.

Определение 6. *Стандартным $(n-1)$-мерным симплексом* называется множество

$$\Delta_{n-1} = \left\{ x \in \mathbb{R}^n \,|\, \sum_{i=1}^{n} x^{(i)} = 1,\ \forall i\ x^{(i)} \geq 0 \right\}.$$

Далее линейной оболочкой матрицы X будем называть линейную оболочку множества ее столбцов $\{X_i\}$ и обозначать $\operatorname{span}(X)$. Аналогично выпуклой оболочкой матрицы X будем называть выпуклую оболочку ее столбцов и обозначать $\operatorname{conv}(X)$.

Определение 7. Точку v выпуклого многогранника M называют *вершиной* M, если не существует такого отрезка $[u_1,u_2] \subset M$, что v является внутренней точкой отрезка $[u_1,u_2]$.

Лемма 1. *Пусть дана стохастическая матрица $F \in \mathbb{R}^{n\times m}$, $\operatorname{rank} F = k$. Каждый выпуклый многогранник $M \subset \Delta_{n-1}$ с k вершинами, т.ч. $\operatorname{conv}(F) \subset M$, биективно соответствует некоторому семейству разложений $F = \Phi\Theta$, которое определяется условием $\operatorname{conv}(\Phi) = M$.*

Доказательство аналогичного утверждения можно найти в [18, лемма 4]. В этой работе доказательство проводится в терминах симплициальных конусов. Каждый симплициальный конус в формулировке работы [18] соответствует некоторому многограннику в приведенной выше формулировке. Соответствие ‘многогранник’ $\to$ ‘конус’ восстанавливается путем взятия всех возможных положительных линейных комбинаций точек из многогранника. Соответствие ‘конус’ $\to$ ‘многогранник’ получается путем пересечения конуса со стандартным симплексом.

Следствие 2. Пусть разложение $F = \Phi\Theta$ таково, что $\mathrm{conv}(F) = \mathrm{conv}(\Phi)$. Тогда, если каждая вершина $\mathrm{conv}(\Phi)$ является вершиной $\mathrm{span}(\Phi) \cap \Delta_{n-1}$, то разложение $F = \Phi\Theta$ единственно.

Доказательство этого следствия можно найти в [20, лемма 4].

3.2. Доказательство теоремы 3

Целью этого подраздела является доказательство теоремы 3. Основой доказательства является лемма 3 о необходимых и достаточных условиях на то, чтобы точка Φ_i являлась вершиной многогранника $\mathrm{span}(\Phi) \cap \Delta_{n-1}$.

Для начала докажем несколько вспомогательных лемм.

Лемма 2. *Пусть многогранник M задан системой*

$$\begin{gathered}\pi_i(x) \ge 0, \quad i = 1,\ldots,m,\\ \sum_s x^{(s)} = 1,\end{gathered}$$

где π_i являются линейными функциями. Точка v является вершиной M тогда и только тогда, когда система

$$\begin{gathered}\pi_i(x) = 0, \quad i \in I,\\ \pi_i(x) > 0, \quad i \notin I,\\ \sum_s x^{(s)} = 1\end{gathered} \tag{3.1}$$

имеет единственное решение, *где*

$$I = \{i \,|\, \pi_i(v) = 0\}.$$

Доказательство. Единственность решения системы (3.1) $\Rightarrow$ v — вершина.

Предположим, что v не вершина, тогда, т.к. она удовлетворяет набору условий $\pi_i(x) \ge 0$, $i = 1,\ldots,m$, значит, точка v лежит в многограннике M, и следовательно, она является серединой некоторого отрезка L с концами u_1 и u_2. Заметим, что в силу симметрии u_1 и u_2 относительно v выполнено $\forall i \in I$ $\pi_i(u_1) = -\pi_i(u_2)$. При этом $\forall x \in M$ $\pi_i(x) \ge 0$ и $\sum_s x^s = 1$. Значит, $\forall i \in I$ $\pi_i(u_1) = \pi_i(u_2) = 0$. Но тогда $\forall x \in L$ x является решением системы (3.1), получаем противоречие с тем фактом, что v является единственным решением системы (3.1).

Таким образом, v — вершина $\Rightarrow$ единственность решения системы (3.1).

Предположим, что нашлось еще одно решение системы (3.1). Значит, система $\pi_i(x) = 0$, $i \in I$ имеет пространство решений размерности больше 0, а пересечение этого пространства решений с системой

$$\begin{gathered}\pi_i(x) > 0, \quad i \notin I,\\ \sum_s x^{(s)} = 1\end{gathered}$$

является многогранником M' размерности больше 0. Отметим, что точка v обязана по определению I быть внутренней для M', в противном случае на ней бы затуплялась некоторая линейная функция π_i для некоторого индекса $i \notin I$. Поэтому точка v представляется как средняя точка некоторого отрезка $[u_1, u_2] \subset M' \subset M$, а значит, не является вершиной M.

Лемма 3. *Пусть дано разложение $F = \Phi\Theta$, $F \in \mathbb{R}^{n\times m}$. Тогда вершина Φ_j многогранника* $\mathrm{conv}(\Phi)$ *является вершиной* $\mathrm{span}(\Phi) \cap \Delta_{n-1}$ *тогда и только тогда, когда*

$$\mathrm{rank}(\Phi[\overline{\mathrm{supp}(\Phi_k)}, [1,\ldots,k]\backslash[j]]) = k-1.$$

Доказательство. Без ограничения общности будем считать, что $j = k$.

Далее будет доказано, что следующие утверждения являются эквивалентными:

(I) Φ_k является вершиной многогранника $\mathrm{span}(\Phi) \cap \Delta_{n-1}$;

(II) $\exists!$ решение системы

$$\begin{aligned}
&\sum_{s=1}^{k}\Phi_{is}x^{(s)} = 0, \quad i \in \overline{\operatorname{supp}(\Phi_k)},\\
&\sum_{s=1}^{k}\Phi_{is}x^{(s)} = 0, \quad i \in \operatorname{supp}(\Phi_k),\\
&\sum_{s=1}^{k}x^{(s)} = 1;
\end{aligned} \tag{3.2}$$

(III) $\exists!$ решение системы уравнений

$$\sum_{s=1}^{k-1}\Phi_{is}x^{(s)} = 0, \quad i \in \overline{\operatorname{supp}(\Phi_k)}; \tag{3.3}$$

(IV) $\operatorname{rank}(\Phi[\overline{\operatorname{supp}(\Phi_k)}, [1,\dots,k]\backslash[k]]) = k-1$;

(III) $\Leftrightarrow$ (IV)

Система (3.3) всегда имеет нулевое решение. То, что это решение единственно, равносильно тому, что ядро отображения, задаваемого матрицей $\Phi[\overline{\operatorname{supp}(\Phi_k)}, [1,\dots,k]\backslash[k]]$, нулевое, что равносильно (IV).

(III) $\Rightarrow$ (II)

Заметим, что при $i \in \overline{\operatorname{supp}(\Phi_k)}\Phi_{ik} = 0$, поэтому

$$\sum_{s=1}^{k}\Phi_{is}x^{(s)} = 0, \quad i \in \overline{\operatorname{supp}(\Phi_k)} \Leftrightarrow \sum_{s=1}^{k-1}\Phi_{is}x^{(s)} = 0, \quad i \in \overline{\operatorname{supp}(\Phi_k)}.$$

Пусть система (3.3) имеет единственное решение, тогда система (3.2) имеет не более одного решения, потому что в ней больше условий. При этом система (3.2) всегда имеет решение $e_k = (0,\dots,0,1)$, которое соответствует точке Φ_k, $\Phi_k = \Phi e_k$. Значит, и система (3.2) имеет единственное решение.

(II) $\Rightarrow$ (III)

Пусть у системы (3.2) существует единственное решение. Тогда оно равно e_k (вектор размера k), потому что такое решение всегда существует. Также 0 (только в данном случае это уже вектор размера $k-1$) является решением системы (3.3). Предположим, что у системы (3.3) нашлось еще одно решение $u \neq 0$. Тогда мы можем определить новое решение $\bar{u}$ системы (3.2), подобрав константу c в векторе $\tilde{u} = (u^{(1)},\dots,u^{(k-1)},c)$ и отнормировав его:

$$\bar{u} = \begin{cases} norm(\tilde{u}), & \sum_s \tilde{u}^{(s)} \neq 0,\\ norm((\tilde{u}^{(1)},\dots,\tilde{u}^{(k-1)},\tilde{u}^{(k)}+1)), & \sum_s \tilde{u}^s = 0,\end{cases}$$

где

$$norm(x)^{(i)} = \frac{x^{(i)}}{\sum_s x^{(s)}}, \quad c = \max(0,\tilde{c}+1),$$

$$\tilde{c} = -\min_{w\in\operatorname{supp}\Phi_k}\frac{\sum_{s=1}^{k-1}\Phi_{ws}u^{(s)}}{\Phi_{wk}}.$$

(I) $\Leftrightarrow$ (II)

Следует из леммы 2. Заметим, что система

$$\begin{aligned}
&\sum_{s=1}^{k}\Phi_{is}x^{(s)} \geq 0,\\
&\sum_{s=1}^{k}x^{(s)} = 1
\end{aligned}$$

задает в точности многогранник $\mathrm{span}(\Phi) \cap \Delta_n$ в пространстве $\mathrm{span}(\Phi)$ в координатах Φ_i, который соответствует многограннику M из леммы 2. Множеству I из леммы 2 соответствует $\overline{\mathrm{supp}(\Phi_k)}$.

Теперь есть все необходимое, чтобы доказать теорему 1.

Доказательство. Из условия

$$\forall i \in [1,\dots,k] \quad \exists j : \Theta_{ij} = 1 \quad \forall i' \neq j, \quad \Theta_{i'j} = 0,$$

следует, что k точек F совпадают с вершинами $\mathrm{conv}(\Phi)$, а значит, $\mathrm{conv}(F) = \mathrm{conv}(\Phi)$.

Условие же на ранги

$$\forall j \quad \mathrm{rank}(\Phi[\overline{\mathrm{supp}(\Phi_j)}, [1,\dots,k]\backslash[j]]) = k-1$$

по лемме 3 равносильно тому, что каждая вершина $\mathrm{conv}(\Phi)$ является вершиной многогранника $\mathrm{span}(\Phi) \cap \Delta_{n-1}$. Значит, теорема 1 следует из следствия 2.

4. ЭКСПЕРИМЕНТЫ

4.1. Интерпретация условий теоремы

Проинтерпретируем в терминах тематического моделирования условия теоремы 1.

Условие 1, заключающееся в наличии в матрице темы-документы Θ столбцов, из которых можно составить единичную матрицу $k \times k$, говорит о наличии k унитематических документов, то есть таких, в которых вероятности встретить любую тему, кроме одной, нулевые. Выполнение этого условия можно гарантировать, добавив в коллекцию k искусственно созданных унитематических документов, слова для которых подбираются, например, экспертами.

Условие 2 говорит о том, что для любого j матрицы

$$\Phi[\overline{\mathrm{supp}(\Phi_j)}, [1,\dots,k]\backslash[j]],$$
$$\Theta[[1,\dots,k]\backslash[j], :]$$

задают неотрицательное разложение полного ранга матрицы $F[\overline{\mathrm{supp}(\Phi_j)}, :]$. Это означает, что матрица Φ, из которой убрали j-ю тему-столбец и все слова, вероятность которых ненулевая в этой теме, и матрица Θ, из которой убрали j-ю тему-строку, задают тематическую модель для матрицы слова-документы F, из которой убрали все слова, встречающиеся в j-й теме.

Для проверки выполнения второго условия на реальной текстовой коллекции был проведен эксперимент.

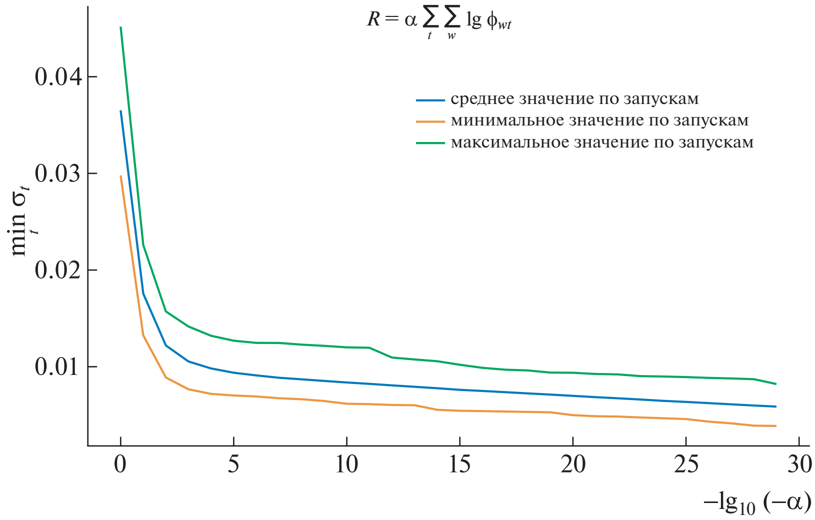
4.2. Описание эксперимента

Эксперимент проводился на лемматизированной коллекции 20newsgroups [21]. Тематическая модель строилась алгоритмом оптимизации ARTM (см. [3], [4]) с регуляризатором разреживания $R(\Phi) = \alpha \sum_t \sum_w \ln \phi_{wt}$ с коэффициентом регуляризации α. В использованной реализации алгоритма ARTM по умолчанию матрица Φ не может содержать нулей. Регуляризатор разреживания позволяет добиться зануления малых значений в этой матрице и контролировать количество нулей.

Для проверки второго условия теоремы требовалось для каждой темы t проверять полноту ранга матрицы $\Phi[\overline{\mathrm{supp}(\Phi_t)}, [1,\dots,T]\backslash[t]]$. Эффективный вычислительный способ сделать это – нахождение минимального сингулярного значения матрицы $\Phi[\overline{\mathrm{supp}(\Phi_t)}, [1,\dots,T]\backslash[t]]$ и сравнение его с нулем. Далее будем обозначать это минимальное сингулярное значение σ_t для темы t.

4.3. Результаты

Как показывает фиг. 1, даже незначительного разреживания ($\alpha = -10^{-25}$) достаточно, чтобы обеспечить единственность разложения. Однако, если запустить исходный алгоритм PLSA без какого либо разреживания, то условия теоремы не будут выполняться из-за отсутствия нулей в матрице Φ.



Фиг 1. Изменение $\min_t \sigma_t$ при стремлении коэффициента α к нулю в регуляризаторе $R(\Phi) = \alpha\sum_t \sum_w \ln \phi_{wt}$, т.е. при уменьшении силы разреживания.

5. ЗАКЛЮЧЕНИЕ

Была описана проблема неединственности решения в задаче тематического моделирования, которая декомпозируется на неоднозначность выбора локального экстремума $\tilde{F}$ оптимизируемого функционала и неединственность разложения $\tilde{F}$ на Φ и Θ.

Был сформулирован ранее неизвестный результат (теорема 1), дающий достаточные условия для единственности решения задачи стохастического матричного разложения. Также был реализован эксперимент, в котором подтвердилось выполнение условий теоремы 1 на реальной текстовой коллекции. Таким образом, было показано, что неединственность решения в задаче тематического моделирования возникает в основном из-за неоднозначности выбора $\tilde{F}$.

СПИСОК ЛИТЕРАТУРЫ

1. *Wenwu Wang*. Instantaneous Versus Convolutive Non-Negative Matrix Factorization. P. 353–370.
2. *David M. Blei, Andrew Y. Ng, Michael I. Jordan, John Lafferty*. Latent dirichlet allocation // J. of Machine Learning Research. 2003. V. 3. P. 993–1022.
3. *Vorontsov K.V.* Additive regularization for topic models of text collections // Doklady Mathematics. 2014. V. 89. № 3. P. 301–304.
4. *Vorontsov K., Potapenko A.* Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. 2014. P. 29–46.
5. *Vorontsov K., Potapenko A.* Additive regularization of topic models // Machine Learning. 2014. V. 101. № 1–3. P. 303–323.
6. *Feng, Yansong, Lapata, Mirella*. Topic Models for Image Annotation and Text Illustration. Stroudsburg, PA, USA, 2010. P. 831–839.
7. *Timothy Hospedales, Shaogang Gong, Tao Xiang*. Video Behaviour Mining Using a Dynamic Topic Model // International Journal of Computer Vision. 2011. V. 98. № 3. P. 303–323.
8. *Xiao-xu Li, Chao-bo Sun, Peng Lu, Xiao-jie Wang, Yi-xin Zhong*. Simultaneous image classification and annotation based on probabilistic model // The Journal of China Universities of Posts and Telecommunications. 2012. V. 19. № 2. P. 107–115.
9. *Pritchard J.K., Stephens M., Donnelly P.* Inference of population structure using multilocus genotype data // Genetics. 2000. V. 155. P. 945–959.
10. *Shivashankar S., Srivathsan S., Ravindran B., Tendulkar A.V.* Multi-view methods for protein structure comparison using latent dirichlet allocation // Bioinformatics. 2011. V. 27. № 13. P. i61–i68.
11. *Vulić I., Wim De Smet, Marie-Francine Moens*. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora // Information Retrieval. 2012. V. 16. № 3. P. 331–368.

12. *Vulić I., Wim De Smet, Jie Tang, Marie-Francine Moens.* Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications // Information Processing & Management. 2015. V. 51. № 1. P. 111–147.
13. *Ianina A., Golitsyn L., Vorontsov K.* Multi-objective Topic Modeling for Exploratory Search in Tech News // Communications in Computer and Information Science. Springer International Publishing. 2017. P. 181–193.
14. *Mikolov T., Sutskever I., Chen K., Corrado. G.S., Dean J.* Distributed Representations of Words and Phrases and their Compositionality // Advances in Neural Information Processing Systems 26 / Red. by C.J.C. Burges, L. Bottou, M. Welling et al. Curran Associates, Inc., 2013. P. 3111–3119.
15. *Potapenko A., Popov A., Vorontsov K.* Interpretable Probabilistic Embeddings: Bridging the Gap Between Topic Models and Neural Networks // Communications in Computer and Information Science. Springer International Publishing, 2017. P. 167–180.
16. *Hofmann T.* Probabilistic latent semantic indexing // Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval – SIGIR 99. ACM Press, 1999.
17. *Kochedykov D., Apishev M., Golitsyn L., Vorontsov K.* Fast and modular regularized topic modelling // 2017 21st Conference of Open Innovations Association (FRUCT). IEEE, 2017.
18. *Donoho D., Stodden V.* When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts? 2004. P. 1141–1148.
19. *Hans Laurberg, Mads Græsbøll Christensen, Mark D. Plumbley, Lars Kai Hansen, Søren Holdt Jensen*. Theorems on Positive Data: On the Uniqueness of NMF // Comput. Intelligence and Neuroscience. 2008. V. 2008. P. 1–9.
20. *Gillis N*. Sparse and unique nonnegative matrix factorization through data preprocessing // The Journal of Machine Learning Research. 2012. V. 13. № 1. P. 3349–3386.
21. *Ken Lang*. 20 Newsgroups. 2008. Data retrieved from the dataset's official website, http://qwone.com/~jason/20Newsgroups/.