Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский университет «Высшая школа экономики»

Факультет математики

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

На тему "Проблемы устойчивости и единственности стохастического матричного разложения."

Студент
Дербаносов Роман Юрьевич

Научный руководитель
проф., доктор физ.-мат. наук
Воронцов Константин Вячеславович

Москва, 2017

Аннотация

В работе исследуется проблема единственности стохастического матричного разложения, т. е. проблема единственности представления фиксированной матрицы в виде произведения двух стохастических матриц. Формулируются и доказываются новые условия, гарантирующие единственность заданного разложения в терминах элементов матриц-факторов.

Содержание

1 Введение **4**

2 Постановка задачи и основная теорема **4**
2.1 Стохастическое матричное разложение 4
2.2 Связь между различными матричными разложениями 5
2.3 Обзор результатов 6
2.4 Условия основной теоремы 8

3 Доказательство достаточных условий **9**
3.1 Когда Φ_i является вершиной многогранника $\text{span}(\Phi) \cap \Delta_n$ 10
3.2 Доказательство теоремы 3 14

4 Связь с тематическим моделированием **14**

5 Заключение **16**

1 Введение

В работе идёт речь 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 её разложение $F = \Phi\Theta$, $\Phi \in \mathbb{R}^{n \times k}$, $\Theta \in \mathbb{R}^{k \times m}$. Такое разложение всегда неединственно даже при фиксированных размерах матриц Φ и Θ, потому что для любой неединичной матрицы перестановки S представление $F = \Phi'\Theta'$, где $\Phi' = \Phi S$, $\Theta' = S^{-1}\Theta$ снова является решением задачи, при этом условия стохастичности матриц Φ' и Θ', дающих разложение, также сохраняются. Поэтому, обычно, разложения рассматривают с точностью до добавления в разложение матрицы перестановки S. Соответственно, единственным называют разложение $F = \Phi\Theta$, для которого все остальные разложения $F = \Phi'\Theta'$ удовлетворяют условию $\Phi = \Phi' S$, $\Theta = S^{-1}\Theta'$ для некоторой матрицы перестановки S. При таком определении единственности уже бывают матрицы F, для которых стохастическое разложение полного ранга единственно.

Проблема неединственности неотрицательного матричного разложения в описанном выше смысле исследовалась в работах [15], [2], [1], 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. Подробный обзор этих результатов и формулировка некоторых из них представлены в разделе "Обзор результатов".

В данной работе представлен результат, гарантирующий единственность данного 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 матричных элементов матриц Φ и Θ, а также простое следствие этого результата, дающее критерий в терминах матрицы F, без заданного разложения. Сравнивается результат нашего следствия и достаточной теоремы статьи [1], также формулируемой в терминах матрицы F, и демонстрируется пример матрицы F, имеющей единственное разложение, для которой работает наше следствие, в то время как теорема статьи [1] не работает.

Последующий текст устроен следующим образом. Во второй секции формулируется задача и определения, требуемые для формулировки основной теоремы, после чего даётся обзор предыдущих результатов и формулируется основная теорема. В третьей секции проводится доказательство основной теоремы. В четвертой секции описывается связь стохастического матричного разложения с тематическим моделированием и смысл условий теоремы с точки зрения тематического моделирования. В пятой секции описываются возможные пути продолжения исследования.

2 Постановка задачи и основная теорема

2.1 Стохастическое матричное разложение

Определение 1. *Будем называть матрицу F неотрицательной, если*

$$\forall i, j \ F_{ij} \geq 0.$$

Определение 2. *Будем называть матрицу F стохастической, если она является неотрицательной и*

$$\forall j \ \sum_i F_{ij} = 1.$$

Определение 3. *Стохастическим матричным разложением полного ранга матрицы $F \in \mathbb{R}^{n\times m}$, $\operatorname{rank} F = k$ называют представление F в виде произведения двух стохастических матриц полного ранга*

$$F = \Phi\Theta,$$

$\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$.

Далее в этой работе будем называть **разложением матрицы** F именно стохастическое матричное разложение полного ранга. Также будем предполагать, что в матрице F нет полностью нулевых столбцов.
Из условий определения разложения 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 $F = \Phi\Theta$, то почти всегда можно найти новое разложение $F = \Phi S S^{-1}\Theta$, где S — матрица перестановки, поэтому дадим следующее определение единственности разложения.

Определение 4. *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.*

Для вектора v будем обозначать $\operatorname{supp}(v)$ множество позиций, на которых стоят ненулевые элементы; $\overline{\operatorname{supp}(v)}$ — множество позиций, на которых стоят нулевые элементы; X_j — j-ый столбец матрицы X , а подматрицу, состоящую из строк $i_1, \ldots, i_p$ и столбцов $j_1, \ldots, j_q$ будем обозначать

$$X\big[\{i_1, \ldots, i_p\}, \{j_1, \ldots, j_q\}\big].$$

2.2 Связь между различными матричными разложениями

Определение 5. *Неотрицательным рангом матрицы F называют такое наименьшее k, что существует стохастическое разложение $F = \Phi\Theta$, $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$ и обозначают $\operatorname{rank}_+ F$.*

Если произвольная матрица $M \in \mathbb{R}^{n\times m}$ имеет ранг k, то существует её представление в виде произведения $M = AB$, $A \in \mathbb{R}^{n\times k}$, $B \in \mathbb{R}^{k\times m}$. Действительно, достаточно записать в матрицу A какие-нибудь k столбцов, порождающих остальные, а в i-ый столбец матрицы B — веса разложения i-го столбца матрицы M по порождающему набору A.
Для произвольной матрицы F положительный ранг, вообще говоря, не определен, ведь в ней могут быть отрицательные элементы и она может быть не стохастической, в этом случае точно не найдутся множители Φ и Θ, дающие в произведении F. Но даже если F стохастическая, ранг и неотрицательный ранг могут различаться. Пример такой матрицы приведён ниже.

Пример 1.

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

Матрица F имеет ранг 3, однако её неотрицательный ранг равен 4.
Более тесная связь имеется между стохастическим и неотрицательным матричными разложениями. Предположим, что у нас есть неотрицательная матрица $F \in \mathbb{R}^{n\times m}$ и её неотрицательное разложение $F = \Phi\Theta$, $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$, т. е. такое, что матрицы Φ и Θ неотрицательны. Можно рассмотреть матрицу $\tilde{F}$, которая равна отнормированной по столбцам матрице F.

Утверждение 1. $\tilde{F} = \tilde{\Phi}\tilde{\Theta}$ *является стохастическим разложением матрицы* $\tilde{F}$, *где*

$$\tilde{\Phi} = \Phi S,$$
$$\tilde{\Theta} = S^{-1}\Theta,$$
$$S = \mathrm{diag}((\sum_i \Phi_{i1})^{-1}, \dots, (\sum_i \Phi_{ik})^{-1}).$$

Доказательство. Матрица S подобрана таким образом, чтобы матрица $\tilde{\Phi}$ была стохастической. Получается, что стохастическая матрица $\tilde{F}$ представляется в виде произведения стохастической $\tilde{\Phi}$ на неотрицательную $\tilde{\Theta}$. Утверждается, что в таком случае и $\tilde{\Theta}$ является стохастической. Действительно, если бы сумма элементов в j-ом столбце матрицы $\tilde{\Theta}$ была бы равна $\alpha \neq 1$, то и после умножения этого столбца на стохастическую матрицу $\tilde{\Phi}$ сумма элементов в столбце по-прежнему была бы равна $\alpha \neq 1$, потому что

$$\sum_i \tilde{\Phi}\tilde{\Theta}_j = \sum_{i,k} \tilde{\Phi}_{ik}\tilde{\Theta}_{kj} = \sum_k \sum_i \tilde{\Phi}_{ik}\tilde{\Theta}_{kj} = \sum_k \tilde{\Theta}_{kj} = \alpha,$$

но эта сумма должна равняться 1, т. к. это сумма элементов в столбце стохастической матрицы $\tilde{F}$, получаем противоречие. Значит, матрица $\tilde{\Theta}$ тоже стохастическая. □

Если же матрица F изначально была стохастической и имела неотрицательное разложение, то по этому разложению описанным в утверждении выше образом можно сделать стохастическое разложение матрицы F. В связи с этим проблемы единственности неотрицательных и стохастических разложений являются очень близкими между собой.

2.3 Обзор результатов

В работе [15] вводится геометрическая интерпретация стохастического матричного разложения. Благодаря этой интерпретации каждому решению задачи разложения стохастической матрицы F однозначно соответствует некоторый симплициальный конус, удовлетворяющий условиям, определяемым по матрице F. Также в ней получены достаточные условия на матрицы-факторы, гарантирующие единственность разложения, при этом достаточные

условия выражаются в терминах симплициальных конусов. Аналог этой геометрической интерпретации в терминах выпуклых многогранников активно используется в нашей работе (см. теорему 4).
В работе [2] получены необходимое условие (Theorem 5) единственности и достаточное условие единственности для неотрицательных необязательно стохастических матричных разложений. Сформулируем основные определения и теорему с необходимыми условиями работы [2].

Определение 6. *Набор векторов S в $\mathbb{R}^d$ называется близким к границе (англ. boundary close), если для всех $j \neq i$ и $k > 0$ существует элемент $s \in S$, т. ч.*

$$s_j < k s_i.$$

В случае замкнутых множеств (а определение используется далее для конечного множества — множества столбцов матрицы), условие означает, что для всех $j \neq i$ существует $s \in S$, т. ч. $s_j = 0, s_i \neq 0$.

Определение 7. *Множество векторов S в $\mathbb{R}^d$ называется сильно близким к границе (англ. strongly boundary close), если оно близко к границе и существует $z > 0$ и перенумерация векторов из S, т. ч. для любого $k > 0$ и $n \in \{1, \ldots, d-1\}$ существует $n - d$ векторов $s_1, \ldots, s_{d-n}$ из S, удовлетворяющих следующим условиям:*

- $s_j^{(n)} < k \sum\limits_{i>n} s_j^{(i)}$ *для всех j,*
- $k_2([b_1, \ldots, b_n]) \leq z$, *где k_2 — число обусловленности матрицы, отношение между наибольшим и наименьшим сингулярным значением. $b_j = P_n s_j$ и $P_n \in \mathbb{R}^{(d-n)\times d}$ есть матрица проекции, переводящая $d - n$ последних элементов вектора в $\mathbb{R}^d$.*

Определение 8. *Набор векторов S в $\mathbb{R}^d$ называется достаточно распространенным (англ. sufficiently spread), если для всех j и $k > 0$ существует элемент $s \in S$, т. ч.*

$$s_j > k \sum_{i \neq j} s_i.$$

Теорема 1. *([2], Theorem 5) Если набор векторов Φ^T является сильно близким к границе (англ. strongly boundary close), а Θ —* достаточно распространенным *(англ. sufficiently spread), то разложение $F = \Phi\Theta$ единственно.*

В работе [1] используется геометрическая интерпретация, введённая в работе [15] и доказывается достаточный критерий единственности (Theorem 6), при этом условия даются в терминах матрицы F. Теперь сформулируем необходимые определения и результат работы [1].

Определение 9. *Будем называть* ***паттерном разреженности*** *(англ. sparsity pattern) строки s номера мест, на которых стоят нули.*

Например, у вектора $(1,0,2,0,5)$ паттерном разреженности является множество $\{2,4\}$, а вектора $(0,4,2)$ — множество $\{1\}$.

Теорема 2. *([1], Theorem 6) Пусть $F \in \mathbb{R}^{m\times n}$, $\operatorname{rank} F = \operatorname{rank}_+ F = k$. Если у F есть r ненулевых столбцов, у каждого из которых есть $r-1$ ноль, такие, что у строк, соответствующих этим нулям, различны паттерны разреженности, тогда у F есть единственное разложение $F = \Phi\Theta$.*

2.4 Условия основной теоремы

Теорема 3. *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 условия:*

- $\forall i \in \{1,\ldots,k\}\ \exists j : \Theta_{ij} = 1, \forall i' \neq j\ \Theta_{i'j} = 0;$
- $\forall j\ \operatorname{rank}\Big(\Phi\big[\overline{\operatorname{supp}(\Phi_j)}, \{1,\ldots,k\} \setminus \{j\}\big]\Big) = k-1.$

Тогда разложение 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_k матрицы Φ подматрица, соответствующая множеству строк, на которых в Φ_k 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 $\{j_1,\ldots,j_k\} := J$, что*

- *(1)* $\forall j \in J\ \operatorname{rank}\Big(\Phi\big[\overline{\operatorname{supp}(\Phi_j)}, J \setminus \{j\}\big]\Big) = k-1,$
- *(2) для любого $p \in \{1,\ldots,m\} \setminus J$ найдутся коэффициенты $a_1,\ldots,a_k$, т. ч. $a_i \geq 0$, $\sum\limits_i a_i = 1$, $F_p = \sum\limits_i a_i F_{j_p}$.*

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 3. Матрица Θ, дающая разложение $F = \Phi\Theta$, найдётся благодаря условию 2. $\square$

Ниже приведен пример матрицы F, которая имеет единственное разложение, удовлетворяет условиям следствия теоремы 3, описанного выше, но не удовлетворяет условиям Theorem 6 статьи [1].

Пример 2.

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

У этой матрицы F есть единственное разложение $F = FE$, но условия теоремы [1] 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 3 выполнены.

3 Доказательство достаточных условий

Определение 10. *Стандартным n-мерным симплексом называется множество*

$$\{x \in \mathbb{R}^n \mid \sum_{i=1}^{n} x^{(i)} = 0, \forall i \ x^{(i)} \geq 0\}.$$

Далее будем обозначать стандартный n-мерный симплекс Δ_n.

Определение 11. *Выпуклой оболочкой векторов $v_1, \ldots, v_n$ называют множество*

$$\{x \in \mathbb{R}^n \mid \exists a_1, \ldots, a_n : \sum_{i=1}^{n} a_i v_i = x, \forall i \ a_i \geq 0\}$$

и обозначают $\mathrm{conv}(v_1, \ldots, v_n)$.

Далее выпуклой оболочкой матрицы X будем называть выпуклую оболочку множества её столбцов $\{X_i\}$ и обозначать $\mathrm{conv}(X)$.

Определение 12. *Линейной оболочкой векторов $v_1, \ldots, v_n$ называют множество*

$$\{x \in \mathbb{R}^n \mid \exists a_1, \ldots, a_n : \sum_{i=1}^{n} a_i v_i = x\}$$

и обозначают $\mathrm{span}(v_1, \ldots, v_n)$.

Далее линейной оболочкой матрицы X будем называть линейную оболочку множества её столбцов $\{X_i\}$ и обозначать $\mathrm{span}(X)$.

Определение 13. *Точку v многогранника M называют вершиной M, если не существует такого отрезка $[u_1, u_2] \subset M$, что v является внутренней точкой отрезка $[u_1, u_2]$.*

Теорема 4. *Пусть дана неотрицательная стохастическая матрица $F \in \mathbb{R}^{n\times m}$. Каждый многогранник $M \subset \Delta_n$ с k вершинами, т.ч. $\mathrm{conv}(F) \subset M$, биективно соответствует некоторому разложению $F = \Phi\Theta$, которое определяется условием $\mathrm{conv}(\Phi) = M$.*

Доказательство аналогичного утверждения можно найти в [15], Lemma 4. В этой работе доказательство проводится в терминах симплициальных конусов. Каждый симплициальный конус в формулировке работы [15] соответствует некоторому многограннику в приведённой выше формулировке. Соответствие 'многогранник' $\to$ 'конус' восстанавливается путём взятия всех возможных положительных линейных комбинаций точек из многогранника. Соответствие 'конус' $\to$ 'многогранник' получается путём 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 $\mathrm{span}(\Phi) \cap \Delta_n$, то разложение $F = \Phi\Theta$ единственно.*

Доказательство. По теореме 4, любая матрица Φ', участвующая в разложении, должна удовлетворять условию $\mathrm{conv}(\Phi') \supset \mathrm{conv}(F)$. Но $\mathrm{conv}(F) = \mathrm{conv}(\Phi)$, поэтому $\forall i\ \Phi_i \in \mathrm{conv}(\Phi')$.
Также заметим, что, т.к. $\forall i\ \Phi_i$ являются вершинами $\mathrm{span}(\Phi) \cap \Delta_n$, поэтому Φ_i не могут быть не вершинами $\mathrm{conv}(\Phi')$. В противном случае, если для некоторого i Φ_i не вершина, то существует открытый интервал $I \subset \mathrm{conv}(\Phi')$, т.ч. $\Phi_i \in I$. Но $\mathrm{conv}(\Phi') \subset \mathrm{span}(\Phi) \cap \Delta_n$, поэтому этот же интервал I лежит и в $\mathrm{span}(\Phi) \cap \Delta_n$, значит, Φ_i не может быть вершиной $\mathrm{span}(\Phi) \cap \Delta_n$, что противоречит условию.
Получаем, что $\forall i\ \Phi_i$ являются вершинами $\mathrm{conv}(\Phi')$. Но $\mathrm{conv}(\Phi')$ имеет ровно k вершин. Значит, $\mathrm{conv}(\Phi') = \mathrm{conv}(\Phi)$ и разложение $F = \Phi\Theta$ единственно. □

3.1 Когда Φ_i является вершиной многогранника $\mathrm{span}(\Phi) \cap \Delta_n$

Целью этого подраздела является доказательство леммы 3 о необходимых и достаточных условиях на то, чтобы точка Φ_i являлась вершиной многогранника $\mathrm{span}(\Phi) \cap \Delta_n$.
Для начала докажем несколько вспомогательных лемм.

Лемма 1. *Пусть многогранник M задан системой*

$$\begin{cases} \pi_i(x) \geq 0, \ i = 1, \dots, m \\ \sum_s x^{(s)} = 1 \end{cases},$$

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

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

имеет единственное решение, где

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

Доказательство.
Единственное решение системы (1) $\Rightarrow$ v — вершина.
Предположим, что v не вершина, тогда, т. к. она удовлетворяет набору условий $\pi_i(x) \geq 0,\ i = 1, \ldots, m$, она является внутренней точкой для некоторой грани, поэтому мы можем выбрать некоторую окрестность $U \ni v$ в этой грани. Рассмотрим, отрезок $L \subset U$ с концами u_1 и u_2, для которого v является серединой. Заметим, что, в силу симметрии u_1 и u_2 относительно v, $\forall i \in I\ \pi_i(u_1) = -\pi_i(u_2)$. При этом $\forall x \in M\ \pi_i(x) \geq 0$ и $\sum\limits_s x^s = 1$. Значит, $\forall i \in I\ \pi_i(u_1) = \pi_i(u_2) = 0$. Но тогда $\forall x \in L\ x$ является решением системы (1), получаем противоречие с тем фактом, что v является единственным решением системы (1).

v — вершина $\Rightarrow$ единственное решение системы (1).
Предположим, что нашлось ещё одно решение системы (1). Значит, система $\pi_i(x) = 0,\ i \in I$ имеет пространство решений размерности больше 0, а пересечение этого пространства решений с системой

$$\begin{cases} \pi_i(x) > 0,\ i \notin I \\ \sum\limits_s x^{(s)} = 1 \end{cases} \tag{2}$$

является многогранником M' размерности больше 0. Отметим, что точка v обязана по определению I быть внутренней для M', в противном случае на ней бы занулялась 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. $\square$

Лемма 2. *Пусть $A \in Mat(n, m)$. Тогда следующие два условия равносильны*

- *(1) $\exists!$ решение системы $Ax = 0$;*
- *(2)* $\operatorname{rank} A = m$.

Доказательство. (1) $\Rightarrow$ (2)
Пусть $\operatorname{rank} A < m$ ($> m$ он быть не может). Тогда у A есть ненулевое ядро, значит, существует целое семейство решений системы $Ax = 0$.
(2) $\Rightarrow$ (1)
Пусть $\operatorname{rank} A = m$. Тогда в матрице A найдётся m линейно независимых строк с индексами $\{i_1, \ldots, i_m\} =: I$. Тогда система $A[I, \{1, \ldots, m\}]x = 0$ имеет единственное нулевое решение. Если к набору уравнений из системы $A[I, \{1, \ldots, m\}]x = 0$ добавить остальные уравнения системы $Ax = 0$, то решений может стать только меньше. Но нулевое решение у такой системы $Ax = 0$ всегда есть, поэтому $Ax = 0$ имеет единственное решение. $\square$

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

$$\operatorname{rank}\Big(\Phi\big[\overline{\operatorname{supp}(\Phi_k)}, \{1, \ldots, k\} \setminus \{j\}\big]\Big) = k - 1. \tag{3}$$

Доказательство. Без ограничения общности будем считать, что $j = k$.
Далее в ходе доказательства будет доказано, что следующие утверждения являются эквивалентными:

- (1) Φ_k является вершиной многогранника $\mathrm{span}(\Phi) \cap \Delta_n$
- (2) $\exists!$ решение системы

$$\begin{cases} \sum\limits_{s=1}^{k} \Phi_{is} x^{(s)} = 0, \ i \in \overline{\mathrm{supp}(\Phi_k)} \\ \sum\limits_{s=1}^{k} \Phi_{is} x^{(s)} > 0, \ i \in \mathrm{supp}(\Phi_k) \\ \sum\limits_{s=1}^{k} x^{(s)} = 1 \end{cases}$$

- (3) $\exists!$ решение системы уравнений $\sum\limits_{s=1}^{k-1} \Phi_{is} x^{(s)} = 0, \ i \in \overline{\mathrm{supp}(\Phi_k)}$
- (4) $\mathrm{rank}\Big(\Phi\big[\overline{\mathrm{supp}(\Phi_k)}, \{1, \ldots, k\} \setminus \{j\}\big]\Big) = k - 1$

(3) $\Leftrightarrow$ (4)
Является прямым следствием леммы 2.
(3) $\Rightarrow$ (2)
Заметим, что при $i \in \overline{\mathrm{supp}(\Phi_k)}$ $\Phi_{ik} = 0$, поэтому $\sum\limits_{s=1}^{k} \Phi_{is} x^{(s)} = 0, \ i \in \overline{\mathrm{supp}(\Phi_k)} \Leftrightarrow \sum\limits_{s=1}^{k-1} \Phi_{is} x^{(s)} = 0, \ i \in \overline{\mathrm{supp}(\Phi_k)}$. Пусть система (3) имеет единственное решение, тогда система (2) имеет не более одного решения, потому что в ней больше условий. При этом система (2) всегда имеет решение $e_k = (0, \ldots, 0, 1)$, которое соответствует точке Φ_k, $\Phi_k = \Phi e_k$. Значит, и система (2) имеет единственное решение.
(2) $\Rightarrow$ (3)
Пусть у системы (2) существует и единственно решение. Тогда оно равно 0 (вектор размер k), потому что это решение всегда точно есть. Также 0 (только в данном случае это уже вектор размера $k - 1$) является и решением системы (3). Предположим, что у системы (3) нашлось ещё одно решение $u \neq 0$. Определим следующий вектор, из которого мы позже сделаем новое решение системы (2):

$$\tilde{u} = (u^{(1)}, \ldots, u^{(k-1)}, c).$$

Опишем, каким образом подбирается константа c.
Во-первых, заметим, что какую бы c мы не выбрали, $\tilde{u}$ будет удовлетворять условию $\sum\limits_{s=1}^{k} \Phi_{is} x^{(s)} = 0, \ i \in \overline{\mathrm{supp}(\Phi_k)}$. Давайте подберём такое c, чтобы выполнялись неравенства системы (2) и

$\sum_{s=1}^{k} x^{(s)} = 1.$

Сначала добьёмся выполнения неравенств системы (2). Для этого возьмём c следующим образом:

$$c = \max(0, \tilde{c})$$

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

Объясним именно такой выбор. Разберём два случая.

1. $\tilde{c} < 0$. В этом случае все неравенства системы (2) выполняются при $c = 0$. Тогда неравенства системы (2) выполняются при $c > 0$. Поэтому возьмём $c = 0$.

2. $\tilde{c} > 0$. В этом случае некоторые из неравенств системы (2) не выполняются. Тогда $\tilde{c}$ подобрана таким образом, что при таком $c > \tilde{c}$ все неравенства системы (2) выполнялись.

Получается, что c выбирается таким образом, чтобы при всех $c' \geq c$ все неравенства системы (2) выполнялись.

Осталось ещё равенство $\sum_{s=1}^{k} x^{(s)} = 1$. Добьёмся его выполнения, тем самым получив решение $\overline{u}$ всей системы (2).

$$\overline{u} = \begin{cases} norm(\tilde{u}),\ if\ \sum_{s} \tilde{u}^{(s)} \neq 0 \\ norm((\tilde{u}^{(1)}, \tilde{u}^{(s)}, \dots, \tilde{u}^{(k-1)}, \tilde{u}^{(k)} + 1)),\ if\ \sum_{s} \tilde{u}^{s} = 0 \end{cases},$$

где

$$norm(x)^{(i)} = \frac{x^{(i)}}{\sum_{s} x^{(s)}}.$$

Суть заключается в том, что мы хотим отнормировать $\tilde{u}$ с помощью функции $norm$. Но может возникнуть проблема, заключающаяся в том, что $\sum_{s} \tilde{u} \neq 0$. Тогда мы прибавляем к последнее координате вектора $\tilde{u}$ единицу, теперь уже можно нормировать, чтобы удовлетворить равенство $\sum_{s=1}^{k} x^{(s)} = 1$ системы (2), а все остальные равенства и неравенства системы (2) продолжат выполняться, потому что такое прибавление соответствует увеличению константы c в векторе $\tilde{u} = (u^{(1)}, \dots, u^{(k-1)}, c)$ на 1, а это только усиливает нужные равенства и неравенства (как уже говорилось выше).
Значит, $\overline{u}$ является ещё одним решением системы (2).

(1) $\Leftrightarrow$ (2)
Заметим, что система

$$\begin{cases} \sum_{s=1}^{k} \Phi_{is} x^{(s)} \geq 0, \ i \in \overline{\mathrm{supp}(\Phi_k)} \\ \sum_{s=1}^{k} \Phi_{is} x^{(s)} \geq 0, \ i \in \mathrm{supp}(\Phi_k) \\ \sum_{s=1}^{k} x^{(s)} = 1 \end{cases}$$

задаёт в точности многогранник $\mathrm{span}(\Phi) \cap \Delta_n$. Поэтому требуемая эквивалентность следует из леммы 1. $\square$

Теперь есть всё необходимое, чтобы доказать теорему 3.

3.2 Доказательство теоремы 3

Доказательство. Докажем, что теорема 3 следует из следствия 2.
Условие

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

равносильно тому, что из k векторов матрицы Θ можно составить единичную матрицу, значит, в матрице F есть k линейно независимых столбцов, а именно, это столбцы матрицы Φ. Но, раз k точек F совпадают с вершинами $\mathrm{conv}(\Phi)$ (которых всего k) и $\mathrm{conv}(F) \subset \mathrm{conv}(\Phi)$, то $\mathrm{conv}(F) = \mathrm{conv}(\Phi)$.
Условие же на ранги

$$\forall j \ \ \mathrm{rank}\Big(\Phi\big[\mathrm{supp}(\Phi_j), \{1, \ldots, k\} \setminus \{j\}\big]\Big) = k - 1$$

по лемме 3 равносильно тому, что каждая вершина $\mathrm{conv}(\Phi)$ является вершиной $\mathrm{span}(\Phi) \cap \Delta_n$. $\square$

4 Связь с тематическим моделированием

Одно из приложений стохастического матричного разложения — тематическое моделирование. Тематическое моделирование применяют в задачах анализа аудио ([3]), текстов ([4], [5], [6], [7]), изображений и видео ([13], [12], [14]), биоинформатике ([11], [10]), в задачах информационного поиска ([8], [9]).
Сформулируем основную задачу тематического моделирования. Пусть у нас есть некоторая коллекция документов. Вообще говоря, каждый документ обычно состоит из последовательности слов (сигналов, действий, транзакций и т.д.). Из этой последовательности слов нужно извлечь набор характеристик (в машинном обучении их называют *признаками*), которые

затем будут использоваться для построения модели. На практике часто используется модель *мешка слов*. В этой модели каждому документу сопоставляется набор значений, каждое из которых соответствует уникальном слову из коллекции и равно числу раз, которое это слово встречается в документе. Теперь всей коллекции можно сопоставить матрицу слова-документы F, где строкам отвечают различные слова из коллекции, столбцам — документы, а на пересечении i-ой строки и j-го столбца стоит число раз, которое i-ое слово встречается в j-ом документе. После этого мы можем перейти от чисел встречаемости к эмпирическим вероятностям, отнормировав значения по столбцам: $F_{ij} = \frac{F_{ij}}{\sum\limits_i F_{ij}}$. Таким образом, мы получим стохастическую матрицу слова-документы F, где в каждом столбце будут записаны дискретные распределения слов в документах.

Темой будем называть дискретное распределение на словах. Пусть матрица F имеет размер $n \times m$. Предположим, что каждый документ является выпуклой комбинацией (т.е. веса комбинации неотрицательны и в сумме дают 1) k фиксированных тем, где $k < \min(n, m)$. В этом случае у матрицы F есть естественное представление в виде произведение $F = \Phi\Theta$. В матрице Φ по столбцам записаны распределения слов в темах, j-ому столбцу отвечает j-ая тема. В матрице Θ в j-ом столбце записаны веса выпуклой комбинации для j-го документа, элемент Θ_{ij} равен весу i-ой темы для j-го документа.

Пусть матрица слова-документы F сгенерирована описанным выше образом. Задача тематического моделирования заключается в том, чтобы по известной матрице F восстановить матрицы Φ и Θ.

Одной из проблем восстановления исходных матриц является то, что в общем случае представление матрицы F в виде произведения $F = \Phi\Theta$ неединственно. Неединственность же разложения ведёт к тому, что на практике алгоритмы ведут себя неустойчиво по отношению к шуму в данных и начальным инициализациям. Также стоит вопрос о том, когда мы можем однозначно восстановить матрицы Φ и Θ, породившие F.

Теперь интерпретируем в терминах тематического моделирования достаточные условия теоремы 3.

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. *Унитематическими документами* называют такие документы, в которых вероятности встретить любую тему, кроме одной, нулевые согласно построенной тематической модели. Эксперименты на небольших коллекциях kos, nips, enron, nytimes (коллекция взяты с ресурса UCI) показывают, что при малом числе тем k ($k < 5$) условие наличия k унитематических документов выполняется автоматически. Однако при большем числе тем это уже не так. Тем не менее, обычной практикой является добавление в коллекцию k искусственно созданных унитематических документов, слова для которых подбираются экспертами, после чего в матрице Θ фиксируется единичная подматрица, не меняющаяся на протяжении всей работы алгоритма, подбирающего разложение $F = \Phi\Theta$. С точки зрения машинного обучения, алгоритмы поиска разложения $F = \Phi\Theta$ по матрице F являются алгоритмами обучения без учителя, в то время как алгоритмы, использующие искусственно созданные унитематические документы, являются алгоритмами частичного обучения.

Условие 2 говорит о том, что для любого j матрицы $\Phi\big[\overline{\mathrm{supp}(\Phi_j)}, \{1, \ldots, k\} \backslash \{j\}\big]$ и $\Theta\big[\{1, \ldots, k\} \backslash$

$j, :\big]$ задают стохастическое разложение полного ранга матрицы $F\big[\overline{\mathrm{supp}(\Phi_j)}, :\big]$. Тот факт, что это разложения полного ранга, говорит о том, что матрица Φ, из которой выкинули j-ую тему-столбец и все слова, вероятность которых ненулевая в этой теме, и матрица Θ, из которой выкинули j-ую тему-строку, задают тематическую модель для матрицы слова-документы F, из которой выкинули все слова, встречающиеся в j-ой теме.

5 Заключение

На защиту выносится новая теорема, дающая достаточные условия единственности стохастического матричного разложения $F = \Phi\Theta$, $F \in \mathbb{R}^{n\times m}$, $\operatorname{rank} F = k$ при условиях $\Phi \in \mathbb{R}^{n\times k}$, $\Theta \in \mathbb{R}^{k\times m}$ в терминах элементов матриц Φ и Θ.

В дальнейших исследованиях планируется ослабление условий доказанной теоремы и, тем самым, её усиление. В идеале — формулировка необходимых и достаточных условий для единственности разложения, выражающихся в терминах матрицы F или матриц-факторов Φ и Θ. Следующими шагами может более подробное изучение топологии пространства всех возможных разложений фиксированной матрицы F, например, определения числа компонент связности пространства разложений. С другой стороны, интерес также представляет введение естественной метрики на пространстве матричных разложений и определение объёма каждой из компонент связности пространства разложений по матрице F или одному из её разложений $F = \Phi\Theta$.

Список литературы

[1] Gillis, N. 2012. *Sparse and unique nonnegative matrix factorization through data preprocessing.* The Journal of Machine Learning Research, 13(1), 3349-3386.

[2] Hans Laurberg, Mads Graesboll Christensen, Mark D. Plumbley, Lars Kai Hansen, and Soren Holdt Jensen. 2008. *Theorems on Positive Data: On the Uniqueness of NMF*. Computational Intelligence and Neuroscience, Volume 2008.

[3] Wang, Wenwu. 2010. *Instantaneous Versus Convolutive Non-Negative Matrix Factorization: Models, Algorithms and Applications to Audio Pattern Separation*. Machine Audition: Principles, Algorithms and Systems. IGI Global. pp. 353–370.

[4] Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. *Latent dirichlet allocation.* the Journal of machine Learning research 3:993–1022.

[5] Vorontsov, K. 2014. *Additive regularization for topic models of text collections*. Doklady Mathematics. 89(3):301–304.

[6] Vorontsov, K., and A. Potapenko. 2014. *Tutorial on probabilistic topic modeling: additive regularization for stochastic matrix factorization.* Analysis of Images, Social networks and Texts. Springer. 29–46.

[7] Vorontsov, K., and A. Potapenko. 2015. *Additive regularization of topic models*. Machine Learning 101(1-3):303–323. doi: http://dx.doi.org/10.1007/s10994-014-5476-6.

[8] Vulic I., Smet W., Moens M.-F. 2012. *Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora*. Information Retrieval. Pp. 1–38.

[9] Vulic I., De Smet W., Tang J., Moens M.-F. 2015. *Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications* Information Processing and Management. Vol. 51, no. 1.Pp. 111–147.

[10] Shivashankar S., Srivathsan S., Ravindran B., Tendulkar A. V. 2011. *Multi-view methods for protein structure comparison using latent dirichlet allocation.* Bioinformatics [ISMB/ECCB]. Vol. 27, no. 13. Pp. 61–68.

[11] Pritchard J. K., Stephens M., Donnelly P. 2000. *Inference of population structure using multilocus genotype data.* Genetics. Vol. 155. Pp. 945–959.

[12] Hospedales T., Gong S., Xiang T. 2012. *Video behaviour mining using a dynamic topic model.* International Journal of Computer Vision. Vol. 98, no. 3. Pp. 303–323.

[13] Feng Y., Lapata M. 2010. *Topic models for image annotation and text illustration.* Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Pp. 831–839.

[14] Li X.-X., Sun C.-B., Lu P., Wang X.-J., Zhong Y.-X. 2012. *Simultaneous image classification and annotation based on probabilistic model*. The Journal of China Universities of Posts and Telecommunications. Vol. 19, no. 2. Pp. 107–115.

[15] David Donoho, Victoria Stodden. 2003. *When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts?*