Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Глушаченков Виталий Владимирович

Устойчивость матричных разложений в задачах тематического моделирования

511656 - Математические и информационные технологии

Магистерская диссертация

Научный руководитель:
с.н.с. ВЦ РАН, д.ф.-м.н.
Воронцов Константин Вячеславович

Москва
2013

Содержание

Аннотация

В данной работе рассматривается проблема неустойчивости решения задачи тематического моделирования текстовых коллекций. Построение тематической модели тесно связано с задачой матричного разложения, решение которой, в силу своей природы, часто является неоднозначным. Неоднозначность разложения, в свою очередь, порождает различные интерпретации исходных данных, по которым строится модель, и это является существенным недостатком.

В работе исследуется влияние разреженности исходных данных на качество восстановления моделей PLSA и LDA. Качественно демонстрируется мера неоднозначности решения. Предложена разреженная модификация алгоритма PLSA-EM и исследовано его качество восстановления оптимальной структуры разреженности исходных данных.

Ключевые слова: *тематическое моделирование, неоднозначность матричных разложений, разреженные тематические модели.*

1 Введение

Актуальность темы. *Тематическое моделирование* — одно из современных приложений методов машинного обучения к анализу текстов, начало которому было положено в конце 90-х годов и с тех пор активно развивающееся. *Тематическая модель* коллекции текстовых документов соотносит каждому документу некоторый набор тем, которым он принадлежит, и определяет какие слова (термины) составляют каждую из тем.

Тематические модели используются для поиска научной информации [10], выявления трендов в научных публикациях и новостных потоках [8], для классификации и категоризации документов [9], в рекомендательных сервисах (коллаборативная фильтрация) [11] и других приложениях.

Вероятностная тематическая модель (ВТМ) коллекции текстовых документов рассматривает каждую тему, как дискретное распределение на множестве терминов, каждый документ, как дискретное распределение на множестве тем. Также предполагается, что коллекция текстовых документов представляет из себя последовательность терминов, выбранных случайно и независимо из смеси этих распределений, и ставится задача восстановления по выборке компонент смеси.

Главной особенностью моделируемых данных, на основе которой строится данная работа, является их разреженность. Естественно предположить, что каждый документ относится к небольшому количеству тем, а для описания сути темы необходима небольшая доля терминов из словаря. Благодаря этой особенности решается одна из актуальных на данный момент проблем, которая связана с неустойчивостью тематических моделей. Суть проблемы заключается в неоднозначности восстановления смесей распределений: для одной и той же выборки, к примеру, в зависимости от начальных приближений могут быть получены различные решения для компонент смесей, каждая из которых дает свою интепретацию исходных данных.

Цель работы. Целью данной работы является: исследование неустойчивости стандартных тематических моделей PLSA-EM и LDA-GS; исследование влияния на неоднозначность разреженности данных; исследование EM-алгоритма с принудительным разреживанием.

2 Вероятностная тематическа модель

2.1 Основные определения и гипотезы

Пусть D — множество (коллеция) текстовых документов, W — множество (словарь) всех употребляемых в документах терминов. Каждый документ $d \in D$ представляет из себя последовательность из n_d терминов $(w_1, w_2, \ldots, w_{n_d})$ из словаря W. Так как один и тот же термин w может встречаться в документе d несколько раз, обозначим число вхождения этого термина в документ через n_{dw}.

Гипотеза о вероятностном пространстве и независимости: Предполагается существование конечного множества тем T и что с каждым употреблением термина w в документе d связана неизвестная тема $t \in T$. Вся коллекция документов рассматривается как множество, выбранных случайно и независимо, троек (d, w, t) из дискретного распределения $p(d, w, t)$, заданного на конечном множестве $D \times W \times T$. Наблюдаемыми переменными являются документы $d \in D$ и термины $w \in W$, латентными (скрытыми) — темы $t \in T$. Выборку из распределения $p(d, w, t)$ можно рассматривать в виде пар $\{(d_i, w_i)\}_{i=1}^n$, где n длина коллекции в терминах. Гипотеза о независимости элементов выборки (гипотеза «мешка слов») означает, что порядок терминов в документах и документов в коллекции не имеет значения:

$$\mathsf{P}\left(\{(d_i, w_i)\}_{i=1}^n\right) = \prod_{i=1}^n \mathsf{P}(d_i, w_i)$$

Гипотеза условной независимости: Вероятность появления термина w в документе d зависит только от темы t, но не от самого документа:

$$p(w \,|\, d, t) = p(w \,|\, t)$$

Гипотеза разреженности: Каждый документ d и каждый термин w связан с небольшим количеством тем t, поэтому большинство условных вероятностей $p(w \,|\, t)$ и $p(t \,|\, d)$ обращается в ноль.

Вероятностная модель порождения данных: Из определения условной вероятности, формуле полной вероятности и гипотезе условной независимости следует:

$$p(w \,|\, d) = \sum_{t \in T} p(w \,|\, t)\, p(t \,|\, d) \tag{2.1}$$

Постановка задачи тематического моделирования: Построить *вероятностную тематическую модель* коллекции документов D — по выборке $\{(d_i, w_i)\}_{i=1}^n$ восстановить совокупность распределений $p(w \mid t)$ для всех тем $t \in T$ и $p(t \mid d)$ для всех документов $d \in D$. Предполагается, что коллекция документов D — это выборка наблюдений $\{(d_i, w_i)\}_{i=1}^n$ полученных согласно (2.1).

Частотные оценки вероятностей: Так как переменные d, w являются наблюдаемыми, то по выборке можно оценить следующие вероятности, как частоты:

$$\hat{p}(d, w) = \frac{n_{dw}}{n}, \quad \hat{p}(d) = \frac{n_d}{n}, \quad \hat{p}(w) = \frac{n_w}{n}, \quad \hat{p}(w \mid d) = \frac{n_{dw}}{n_d} \tag{2.2}$$

n_{dw} — число вхождений термина w в документ d;
$n_d = \sum\limits_{w \in W} n_{dw}$ — длина документа d в терминах;
$n_w = \sum\limits_{d \in D} n_{dw}$ — число вхождений термина w в все документы;
$n = \sum\limits_{d \in D} \sum\limits_{w \in d} n_{dw}$ — длина коллекции в терминах.

Если рассматривать коллекцию, как выборку троект (d, w, t), то также можно оценить вероятности, связанные со скрытой переменной t:

$$\hat{p}(t) = \frac{n_t}{n}, \quad \hat{p}(w \mid t) = \frac{n_{wt}}{n_t}, \quad \hat{p}(t \mid d) = \frac{n_{dt}}{n_d}, \quad \hat{p}(t \mid d, w) = \frac{n_{dwt}}{n_{dw}} \tag{2.3}$$

n_{dwt} — число троек, в которых термин w документа d связан с темой t;
$n_{wt} = \sum\limits_{d \in D} n_{dwt}$ — число троек, в которых термин w связан с темой t;
$n_{dt} = \sum\limits_{w \in W} n_{dwt}$ — число троек, в которых термин документа d связан с темой t;
$n_t = \sum\limits_{d \in D} \sum\limits_{w \in d} n_{dwt}$ — число троек, связанных с темой t.

2.2 Принцип максимума правдоподобия. Обучение модели.

Введем новые обозначения для неизвестных параметров модели $\varphi_{wt} = p(w \mid t)$, $\theta_{dt} = p(t \mid d)$, а матрицы составленные из различных φ_{wt} и θ_{td} через Φ и Θ соответственно. Матрицы Φ, Θ будем называть «матрица тем» и «матрица документов». Для нахождения параметров Φ и Θ максимизируется правдоподобие выборки:

$$p\,(D; \Phi, \Theta) = C \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}} = C \prod_{d \in D} \prod_{w \in d} p(w \mid d)^{n_{dw}}\, p(d)^{n_{dw}} \to \max_{\Phi, \Theta} \tag{2.4}$$

здесь C — нормировочный множитель мультиноминального распределения, который зависит только от n_{dw} и не влияет на положение максимума. Вероятности $p(d)$ также

можно не учитывать. Подставив выражение для $p(w \mid d)$ из (2.1) и прологарифмировав, получаем следующую задачу максимизации:

$$L\left(D;\Phi,\Theta\right) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta} \tag{2.5}$$

при ограничениях нормировки и неотрицательности распределений:

$$\varphi_{wt} \geqslant 0, \quad \sum_{w\in W} \varphi_{wt} = 1, \quad \theta_{td} \geqslant 0, \quad \sum_{t\in T} \theta_{td} = 1 \tag{2.6}$$

2.3 Связь с задачей матричного разложения

Заметим, что задачу оптимизации (2.5) можно переписать через минимизацию взвешенной суммы расстояний Кульбака-Лейблера $\mathrm{KL}(\hat{p}\,||\,p) = \sum\limits_{w\in d} \hat{p}(w \mid d) \ln \frac{\hat{p}(w \mid d)}{p(w \mid d)}$ между эмпирическими $\hat{p}(w \mid d) = n_{dw}/n_d$ и модельными $p(w \mid d) = \sum\limits_{t\in t} \varphi_{wt}\theta_{td}$ распределениям по всем $d \in D$:

$$\begin{aligned}
&\sum_{d\in D} n_d \sum_{w\in d} \hat{p}(w \mid d)\ \ln \hat{p}(w \mid d) - L\left(D;\Phi,\Theta\right) = \\
&= \sum_{d\in D} n_d \sum_{w\in d} \hat{p}(w \mid d)\ \ln \frac{\hat{p}(w \mid d)}{\sum\limits_{t\in T} \varphi_{wt}\theta_{td}} = \\
&= \sum_{d\in D} n_d\ \mathrm{KL}(\hat{p}(w \mid d)\,||\,p(w \mid d)) \to \min_{\Phi,\Theta}
\end{aligned} \tag{2.7}$$

Таким образом, обозначив взвешенную сумму расстояний Кульбака-Лейблера через D_{KL}, задача обучения эквивалентна задаче поиска приближения известной матрицы частот $F = (\hat{p}(w \mid d))_{W\times D}$ матричным разложением $F' = \Phi\Theta$ ($\Phi = (\varphi_{wt})_{W\times T}$, $\Theta = (\theta_{td})_{T\times D}$), таким что:

$$D_{\mathrm{KL}}(F\,||\,\Phi\Theta) \to \min_{\Phi\Theta} \tag{2.8}$$

2.4 Неоднозначность матричного разложения

Для решения задачи (2.8) существуют различные итерационные алгоритмы (например, алгоритмы LDA-GS и PLSA-EM, которые будут рассмотрены далее). Одной из первых проблем, которые возникают в подобных алгоритмах, является неустойчивость решения при различных начальных приближениях модельных данных, но даже в случае достижения алгоритмом экстремума правдоподобия $L(D;\Phi,\Theta)$, полученное решение необязательно является единственно возможным. Очевидно, что $F' = \Phi\Theta = (\Phi R)(R^{-1}\Theta) = \Phi'\Theta'$, где R — некоторая невырожденная матрица преобразования размера $T\times T$. При этом матрицы Φ', Θ' могут существенно отличаться от

Φ, Θ и давать совершенно иную интерпретацию тематической модели. На матрицу R накладываются ограничения, которые связаны с неотрицательностью и нормированностью распределений, составляющих матрицы «тем» и «документов», но легко показать, что этих ограничений недостаточно для единственности разложения матрицы F'.

Рассмотрим искусственный пример: $|W| = 4$, $|D| = 3$, $|T| = 2$

$$\Phi = \begin{bmatrix} 0.5 & 0.125 \\ 0.25 & 0.125 \\ 0.125 & 0.25 \\ 0.125 & 0.5 \end{bmatrix}, \quad \Theta = \begin{bmatrix} 0.7 & 0.6 & 0.5 \\ 0.3 & 0.4 & 0.5 \end{bmatrix}, \quad F' = \Phi\Theta = \begin{bmatrix} 0.3875 & 0.35 & 0.3125 \\ 0.2125 & 0.2 & 0.1875 \\ 0.1625 & 0.175 & 0.1875 \\ 0.2375 & 0.275 & 0.3125 \end{bmatrix}$$

Легко видно, что столбцы соответствующих распределений в матрицах неотрицательны и нормированы. Возьмем невырожденную матрицу R и преобразуем матрицы «тем» и «документов»:

$$R = \begin{bmatrix} 0.5 & 0.75 \\ 0.5 & 0.25 \end{bmatrix}, \quad R^{-1} = \begin{bmatrix} -1 & 3 \\ 2 & -2 \end{bmatrix}$$

$$\Phi' = \Phi R = \begin{bmatrix} 0.3125 & 0.40625 \\ 0.1875 & 0.21875 \\ 0.1875 & 0.15625 \\ 0.3125 & 0.21875 \end{bmatrix}, \quad \Theta' = R^{-1}\Theta = \begin{bmatrix} 0.2 & 0.6 & 1 \\ 0.8 & 0.4 & 0 \end{bmatrix}$$

$$F' = \Phi'\Theta' = \begin{bmatrix} 0.3875 & 0.35 & 0.3125 \\ 0.2125 & 0.2 & 0.1875 \\ 0.1625 & 0.175 & 0.1875 \\ 0.2375 & 0.275 & 0.3125 \end{bmatrix} = \Phi\Theta, \quad \Phi \neq \Phi', \quad \Theta \neq \Theta'$$

Видно, что матрицы Φ' и Θ' остались неотрицательными, а их столбцы, соответствующие распределениям, остались нормированными. То есть, для одной и той же матрицы частот F' получено совершенно два различных разложения. Данный пример наглядно демонстрирует проблему неодназначности матричных разложений.

Разумно предположить, что разложение будет единственным в случае, когда единственно возможным вариантом преобразования R будет перестановочная матрица. Это очевидно, так как изначально порядок тем неизвестен, но от их произвольной перестановки местами интепретируемость модели не зависит. Вопрос, в каких случаях такое возможно и разложение становится уникальным, представляется

Алгоритм 3.1. EM-алгоритм для тематической модели PLSA.

Вход: коллекция документов D, число тем $|T|$, начальные приближения Θ и Φ;
Выход: распределения Θ и Φ;

повторять
  обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ для всех $d \in D,\ w \in W,\ t \in T$;
  для всех $d \in D,\ w \in d$
    $Z := \sum\limits_{t \in T} \varphi_{wt}\theta_{td}$;
    для всех $t \in T$ таких, что $\varphi_{wt}\theta_{td} > 0$
      увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на $\delta = n_{dw}\varphi_{wt}\theta_{td}/Z$;
  $\varphi_{wt} := \hat{n}_{wt}/\hat{n}_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := \hat{n}_{dt}/n_d$ для всех $d \in D,\ t \in T$;
пока Θ и Φ не стабилизируются.

нетривиальным. Но интуитивно понятно, что устранить или понизить степень неоднозначности может наличие в матрицах «тем» и «документов», на которые производится разложение, большого числа нулевых элементов, другими словами — наличие разреженности.

3 Тематические модели PLSA, LDA

3.1 Вероятностный латентный семантический анализ

Вероятностный латентный семантический анализ (probabilistic latent semantic analysis, PLSA) был предложен Томасом Хофманном в [1]. Для вероятности появления пары «документ-термин» (d, w) используется представление из выражения (2.1). Для решения задачи (2.5) применяется итерационный процесс известный как EM-алгоритм. На каждой итерации используется два шага - E (expectation) и M (maximization) [2]. В начале работы алгоритма (перед первой итерацией) задается начальное приближение параметров φ_{wt} и θ_{td}.

На E-шаге по текущим значениям параметров φ_{wt} и θ_{td} при помощи формулы Байеса вычисляются условные вероятности всех тем $t \in T$ для термина $w \in d$ в каждом документе d:

$$p(t \mid d, w) = \frac{p(w \mid t)p(t \mid d)}{p(w \mid d)} = \frac{\varphi_{wt}\theta_{td}}{\sum\limits_{s \in T} \varphi_{ws}\theta_{sd}} \tag{3.1}$$

На M-шаге для вычисленных на E-шаге вероятностей $p(t \mid d, w)$ уточняются параметры φ_{wt} и θ_{td} через максимизацию правдоподобия (2.5) при ограничениях (2.6). Также этот результат можно получить используя оценку $\hat{n}_{dwt} = n_{dw}\, p(t \mid d, w)$ и оцен-

ки (2.3):

$$\varphi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t} = \frac{\sum\limits_{d \in D} n_{dw}\, p(t \mid d, w)}{\sum\limits_{w' \in W} \sum\limits_{d \in D} n_{dw'}\, p(t \mid d, w')}, \quad \theta_{td} = \frac{\hat{n}_{dt}}{n_d} = \frac{\sum\limits_{w \in d} n_{dw}\, p(t \mid d, w)}{\sum\limits_{t' \in T} \sum\limits_{w \in d} n_{dw}\, p(t' \mid d, w)} \tag{3.2}$$

Стоит сделать важное замечание — если начальные приближения φ_{wt} и θ_{td} были положительны, то после каждой итерации они будут положительны. Наоборот, если значение были нулевыми, то нулевое значение будет сохраняться на протяжении всех итераций.

В данной работе используется эквивалентный вариант реализации EM-алгоритма, который будем называть рациональным. Он представлен в виде Алгоритма (3.1), в котором E-шаг встроен в M-шаг, где условные вероятности $p(t \mid d, w)$ вычисляются в тот момент, когда они необходимы.

3.2 Латентное размещение Дирихле

Тематическая модель латентного размещения Дирихле (latent Dirichlet allocation, LDA) была предложена Дэвидом Блайем в [3]. Она также основана на разложении (2.1), но используется байесовская регуляризация, которая основана на введении априорного распределения вероятности в пространстве параметров. В LDA предполагается, что векторы документов $\theta_d = (\theta_{td}) \in \mathbb{R}^{|T|}$ и векторы тем $\varphi_t = (\varphi_{wt}) \in \mathbb{R}^{|W|}$ порождаются распределениями Дирихле с параметрами $\alpha \in \mathbb{R}^{|T|}$ и $\beta \in \mathbb{R}^{|W|}$ соответственно:

$$\mathrm{Dir}(\theta_d; \alpha) = \frac{\Gamma(\alpha_0)}{\prod\limits_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \qquad \alpha_t > 0, \quad \alpha_0 = \sum_t \alpha_t, \qquad \theta_{td} > 0, \qquad \sum_t \theta_{td} = 1;$$

$$\mathrm{Dir}(\varphi_t; \beta) = \frac{\Gamma(\beta_0)}{\prod\limits_w \Gamma(\beta_w)} \prod_w \varphi_{wt}^{\beta_w - 1}, \quad \beta_w > 0, \quad \beta_0 = \sum_w \beta_w, \quad \varphi_{wt} > 0, \quad \sum_w \varphi_{wt} = 1.$$

где $\Gamma(z)$ — гамма-функция. Векторы α и β называются гиперпараметрами.

Для того, чтобы модифицировать PLSA-EM в LDA, достаточно заменить правила обновления параметров (3.2) на сглаженные оценки:

$$\varphi_{wt} = \frac{\hat{n}_{wt} + \beta_w}{\hat{n}_t + \beta_0}, \quad \theta_{td} = \frac{\hat{n}_{dt} + \alpha_t}{n_d + \alpha_0} \tag{3.3}$$

В данной работе применяется вариант алгоритма обучения модели LDA-GS (Алгоритм 3.2), в котором используется сэмплирование Гиббса. Данный подход предложен в [4], а строгий вывод формул приводится в отчёте [5].

Алгоритм 3.2. Сэмплирование Гиббса LDA-GS.

Вход: коллекция D, число тем $|T|$, начальные приближения Θ и Φ, гиперпараметры α, β;
Выход: распределения Θ и Φ;

обнулить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$, $\forall d \in D,\ \forall w \in W,\ \forall t \in T$;
повторять
  для всех $d \in D,\ w \in d,\ i = 1, \ldots, n_{dw}$
    если не первый проход коллекции **то**
      $t := t_{dwi}$; уменьшить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на 1;
    вычислить φ_{wt}, θ_{td} согласно (**??**);
    сэмплировать t_{dwi} из $p(t \mid d, w) \propto \varphi_{wt}\theta_{td}$;
    $t := t_{dwi}$; увеличить $\hat{n}_{wt}$, $\hat{n}_{dt}$, $\hat{n}_t$ на 1;
пока Θ и Φ не стабилизируются.
обновить φ_{wt}, θ_{td}, $\forall d \in D,\ \forall w \in W,\ \forall t \in T$;

3.3 Выбор начального приближения

В данной работе во всех алгоритмах обучения начальные приближения φ_{wt} и θ_{td} задаются посредством обхода всей коллекции, где каждой паре (d, w) назначается случайная тема t из равномерного распределения на темах и вычисляются частотные оценки (2.3) вероятностей φ_{wt} и θ_{td} для всех $d \in D$, $w \in W$, $t \in T$.

4 Разреживание модели PLSA в ходе алгоритма обучения

Согласно гипотезе разреженности, каждый документ $d \in D$ и каждый термин $w \in W$ связан с малым количеством тем $t \in T$. По этой причине значительная часть вероятностей φ_{wt} и θ_{td} должна быть нулевой. Описанные выше алгоритмы PLSA и LDA не позволяют определить, какие конкретно из значений необходимо обнулить.

Для определения позиций для обнуления предлагается применить метод, аналогичный описанному в алгоритме по разреживанию нейронных сетей Optimal Brain Damage (OBD) [6]. За основу для нового алгоритма берется PLSA-EM.

4.1 Алгоритм разреживания на основе метода OBD

Допустим, что EM-алгоритм сошёлся в точку локального максимума правдоподобия (2.5) при ограничениях (2.6), то есть будем рассматривать функцию Лагранжа

данной задачи условной максимизации:

$$\mathscr{L}(\Phi,\Theta)=\sum_{d\in D}\sum_{w\in d}n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td}-\sum_{t\in T}\lambda_t\left(\sum_{w\in W}\varphi_{wt}-1\right)-\sum_{d\in D}\mu_d\left(\sum_{t\in T}\theta_{td}-1\right)$$

Производные первого порядка:

$$\frac{\partial\mathscr{L}(\Phi,\Theta)}{\partial\varphi_{wt}}=\sum_d n_{dw}\frac{\theta_{td}}{p(w\,|\,d)}-\lambda_t=0,\quad \frac{\partial\mathscr{L}(\Phi,\Theta)}{\partial\theta_{td}}=\sum_w n_{dw}\frac{\varphi_{wt}}{p(w\,|\,d)}-\mu_d=0$$

Ненулевые производные второго порядка:

$$\begin{aligned}
\frac{\partial^2\mathscr{L}(\Phi,\Theta)}{\partial\varphi_{wt}\partial\varphi_{wt'}}&=-\sum_d n_{dw}\frac{\theta_{td}\theta_{t'd}}{p(w\,|\,d)^2},\\
\frac{\partial^2\mathscr{L}(\Phi,\Theta)}{\partial\theta_{td}\partial\theta_{t'd}}&=-\sum_w n_{dw}\frac{\varphi_{wt}\varphi_{wt'}}{p(w\,|\,d)^2},\\
\frac{\partial^2\mathscr{L}(\Phi,\Theta)}{\partial\varphi_{wt}\partial\theta_{t'd}}&=n_{dw}\left(\frac{[t=t']}{p(w\,|\,d)}-\frac{\varphi_{wt'}\theta_{td}}{p(w\,|\,d)^2}\right)
\end{aligned}$$

Запишем ряд Тейлора без учета нулевых производных:

$$\mathscr{L}(\Phi+\Delta\Phi,\Theta+\Delta\Theta)=\mathscr{L}(\Phi,\Theta)+\frac{1}{2}\sum_w\sum_t\sum_{t'}\Delta\varphi_{wt}\Delta\varphi_{wt'}\frac{\partial^2\mathscr{L}(\Phi,\Theta)}{\partial\varphi_{wt}\partial\varphi_{wt'}}+$$
$$+\frac{1}{2}\sum_d\sum_t\sum_{t'}\Delta\theta_{td}\Delta\theta_{t'd}\frac{\partial^2\mathscr{L}(\Phi,\Theta)}{\partial\theta_{td}\partial\theta_{t'd}}+\sum_w\sum_d\sum_t\sum_{t'}\Delta\varphi_{wt}\Delta\theta_{t'd}\frac{\partial^2\mathscr{L}(\Phi,\Theta)}{\partial\varphi_{wt}\partial\theta_{t'd}}+o\left(\Delta\Phi,\Delta\Theta\right)$$

Обнулить параметр φ_{wt} означает положить $\varphi_{wt}+\Delta\varphi_{wt}=0$, откуда следует $\Delta\varphi_{wt}=-\varphi_{wt}$. Аналогично, $\Delta\theta_{td}=-\theta_{td}$. Смешанные производные взаимно сокращаются:

$$\begin{aligned}
\sum_t\sum_{t'}\varphi_{wt}\theta_{t'd}\,\frac{\partial^2\mathscr{L}(\Phi,\Theta)}{\partial\varphi_{wt}\partial\theta_{t'd}}&=\sum_t\sum_{t'\neq t}\varphi_{wt}\theta_{t'd}\frac{\partial^2\mathscr{L}(\Phi,\Theta)}{\partial\varphi_{wt}\partial\theta_{t'd}}+\sum_t\varphi_{wt}\theta_{td}\,\frac{\partial^2\mathscr{L}(\Phi,\Theta)}{\partial\varphi_{wt}\partial\theta_{td}}=\\
&=-\sum_t\sum_{t'\neq t}\varphi_{wt}\theta_{t'd}\,n_{dw}\frac{\varphi_{wt'}\theta_{td}}{p(w\,|\,d)^2}+\sum_t\varphi_{wt}\theta_{td}\,n_{dw}\,\frac{\sum\limits_{t'\neq t}\varphi_{wt'}\theta_{t'd}}{p(w\,|\,d)^2}=\\
&=-\sum_t\sum_{t'\neq t}\varphi_{wt}\theta_{t'd}\,n_{dw}\frac{\varphi_{wt'}\theta_{td}}{p(w\,|\,d)^2}+\sum_t\sum_{t'\neq t}\varphi_{wt}\theta_{td}\,n_{dw}\,\frac{\varphi_{wt'}\theta_{t'd}}{p(w\,|\,d)^2}=0
\end{aligned}$$

Подставивим ненулевые производные в разложение и перегруппируем слагаемые:

$$\begin{aligned}
&\mathscr{L}(\Phi+\Delta\Phi,\Theta+\Delta\Theta)-\mathscr{L}(\Phi,\Theta)=\\
&=-\frac{1}{2}\sum_{t\in T}\hat{n}_t\sum_{w\in W}\varphi_{wt}-\frac{1}{2}\sum_{d\in D}\hat{n}_d\sum_{t\in T}\theta_{dt}+o(\Delta\Phi,\Delta\Theta)=\\
&=-\frac{1}{2}\sum_{wt}\hat{n}_{wt}-\frac{1}{2}\sum_{td}\hat{n}_{td}+o(\Delta\Phi,\Delta\Theta)
\end{aligned}\tag{4.1}$$

Из полученых формул (4.1) можно сделать вывод, что счетчики $\hat{n}_{wt}$ и $\hat{n}_{td}$ вносят аддитивные вклады в изменение правдоподобия. Это обосновывает интуитивно понятную стратегию разреживания: после каждого прохода коллекции в каждом распределении $\varphi_{wt} = \hat{n}_{wt}/\hat{n}_t$ и $\theta_{td} = \hat{n}_{td}/n_d$ обнуляются наименьшие значения вероятностей.

В данной работе используется следующая стратегия разреживания, которая встраивается в EM-алгоритм (3.1) после 8-ого шага:

1. После каждой итерации, начиная с некоторой заданной i_0, обнуляются наименьшие значения векторов вероятностей (хвосты распределений) φ_t, θ_d по очереди для всех $t \in T$, $d \in D$.

2. Доля ненулевых элементов в каждом из распределений, которая подлежит обнулению, не превышает некоторый порог M, но так, чтобы сумма обнуляемых элементов не превышала порог S_φ для φ_t и S_θ для θ_d.

4.2 Коррекция вероятностной модели

При разреживании матриц Φ и Θ значение

$$Z_{dw} = \sum_{t \in T} \varphi_{wt}\theta_{td}$$

может обратиться в ноль для некоторых (d, w). Подобные термины можно интерпретировать как нетематические, шумовые. Однако несмотря на интерпретируемость, такая ситуация вызывает появление в $L(D; \Phi, \Theta)$ нуля под логарифмом. Чтобы избежать неопределенности, вероятностная модель корректируется:

$$p(w \,|\, d) = \nu_d \sum_{t \in T} \varphi_{wt}\theta_{td} + [Z_{dw} = 0]\pi_{dw}$$

где π_{dw} — новые параметры модели, $\pi_{dw} > 0$ тогда и только тогда, когда $Z_{dw} = 0$; нормировочный множитель ν_d выбирается из условия $\sum_{w \in d} p(w \,|\, d) = 1$.

Отметим, что корректировка не влияет на вид матриц Φ, Θ.

5 Постановка экспериментов

5.1 Генерация модельных данных

Во всех экспериментах количество документов полагалось равным $|D| = 500$, количество терминов — $|W| = 1000$, количество тем — $|T| = 30$, а длина документа n_d — равномерная случайная величина из интервала $[100, 600]$.

Для генерации исходных столбцов матриц $\Phi_{wt} = p(w \mid t)$, $\Theta_{td} = p(t \mid d)$ применялось два подхода:

1. Распределения Дирихле с симметричными гиперпараметрами β размерности $|W|$, α размерности $|T|$ — $\varphi_t \sim \mathrm{Dir}(\beta)$, $\theta_d \sim \mathrm{Dir}(\alpha)$.

2. Равномерное распределение с обнулением заданной доли элементов в каждом столбце R_φ в матрице Φ и R_θ в матрице Θ.

Коллекция документов генерилась на основе модели порождения данных из смеси распределений (2.1). Также следует учесть, что в полученной коллекции не всегда предсталены все термины из словаря, поэтому для них может быть $|W| < 1000$. При сравнениях матриц, термины которые не попали в коллекцию, просто игнорируются.

5.2 Функционалы качества восстановления модели

В качестве меры отклонения восстановленного семейства распределений от истинного предлагается усредненное по столбцам расстояние Хеллингера между распределениями:

$$H(P, Q) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{P_{ij}} - \sqrt{Q_{ij}} \right)^2} \tag{5.1}$$

Φ и Θ — истинные значения матриц, по которым генерилась коллекция. $\hat{\Phi}$ и $\hat{\Theta}$ — восстановленные.

$$D_{\Phi\Theta}(\hat{\Phi}\hat{\Theta}, \Phi\Theta) = H(\hat{\Phi}\hat{\Theta}, \Phi\Theta) \tag{5.2}$$

$$D_{\Phi}(\hat{\Phi}, \Phi) = H(\hat{\Phi}, \Phi) \tag{5.3}$$

$$D_{\Theta}(\hat{\Theta}, \Theta) = H(\hat{\Theta}, \Theta) \tag{5.4}$$

Очевидно, что применение одной и той же перестановки к столбцам в матрице Φ и к строкам в матрице Θ, не меняет их произведение $\Phi\Theta$, а значит распределение $p(w \mid d)$ также не изменяется. Это значит, что матрицы Φ, Θ восставливаются с

точность до перестановки тем. Поэтому перед сравнением восстановленных матриц с истинными значениями ищется перестановка тем π, которая минимизирует функционал:

$$\arg\min_{\pi} f(\pi) = \sum_{t\in T}\left(\sqrt{\frac{1}{2}\sum_{w\in W}\left(\sqrt{\hat{\varphi}_{w\pi_t}} - \sqrt{\varphi_{wt}}\right)^2} + \sqrt{\frac{1}{2}\sum_{d\in D}\left(\sqrt{\hat{\theta}_{\pi_t d}} - \sqrt{\theta_{td}}\right)^2}\right) \quad (5.5)$$

Минимизация данного функционала эквивалентна задаче о назначениях, которая решается Венгерским алгоритмом [7], где матрица стоимости равна сумме расстояний Хеллингера:

$$M(t,t') = \sqrt{\frac{1}{2}\sum_{w\in W}\left(\sqrt{\hat{\varphi}_{wt'}} - \sqrt{\varphi_{wt}}\right)^2} + \sqrt{\frac{1}{2}\sum_{d\in D}\left(\sqrt{\hat{\theta}_{t'd}} - \sqrt{\theta_{td}}\right)^2} \quad (5.6)$$

5.3 Функционалы качества восстановления структуры разреженности

Ошибки первого рода:

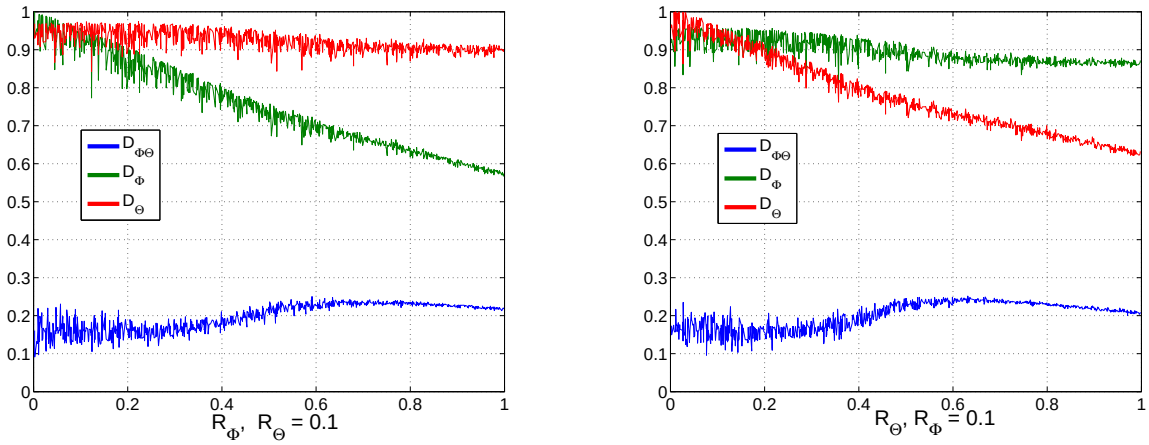
- $S^1_\Phi = \sum\limits_{w,t}[\hat{\Phi}_{wt} > 0][\Phi_{wt} = 0]$
- $S^1_\Theta = \sum\limits_{d,t}[\hat{\Theta}_{dt} > 0][\Theta_{dt} = 0]$

Ошибки второго рода:

- $S^2_\Phi = \sum\limits_{w,t}[\hat{\Phi}_{wt} = 0][\Phi_{wt} > 0]$
- $S^2_\Theta = \sum\limits_{d,t}[\hat{\Theta}_{dt} = 0][\Theta_{dt} > 0]$

6 Эксперименты

6.1 Влияние перестановки тем



a) LDA-GS b) PLSA-EM

Рис. 1: Зависимость расстояний без отыскания перестановки тем

R_Φ — доля ненулей в истинной матрице Φ, R_Θ — доля ненулей в истинной матрице Θ. Левый график — $R_\Theta = 0.1$ фиксирована, R_Φ изменяется. Правый — $R_\Phi = 0.1$ фиксирована, R_Θ изменяется. Данные графики наглядно демонстрируют необходимость отыскания перестановки тем. Отметим, что уже из этих графиков видно, что изменяемая переменная восстанавливается лучше. $\Phi\Theta$ одинаково хороша для всех точек.

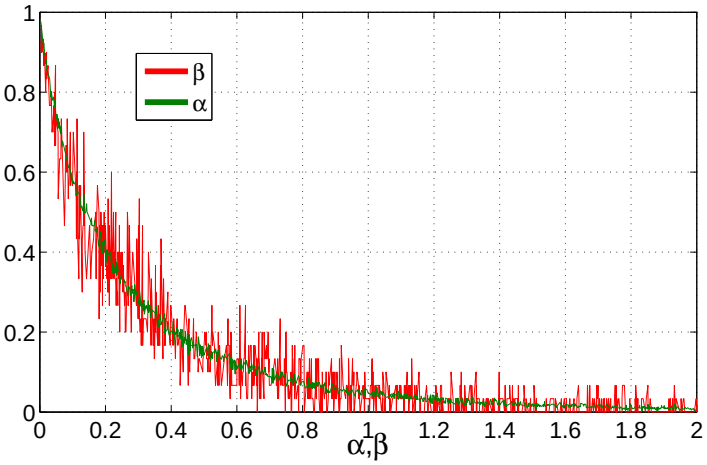
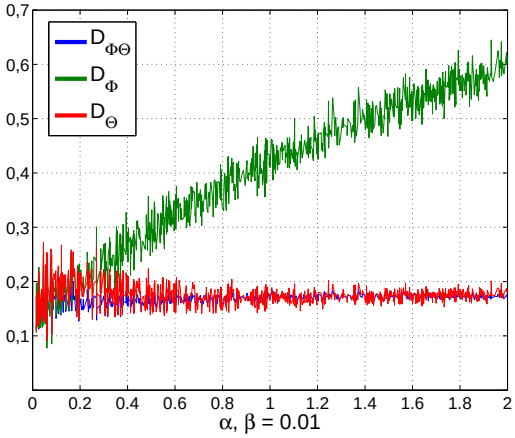


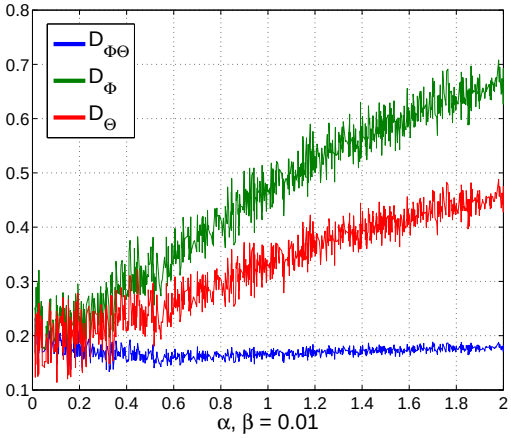
Рис. 2: Приблизительная зависимость степени разреженности от гиперпараметров Дирихле

6.2 Сравнение LDA-GS и PLSA-EM

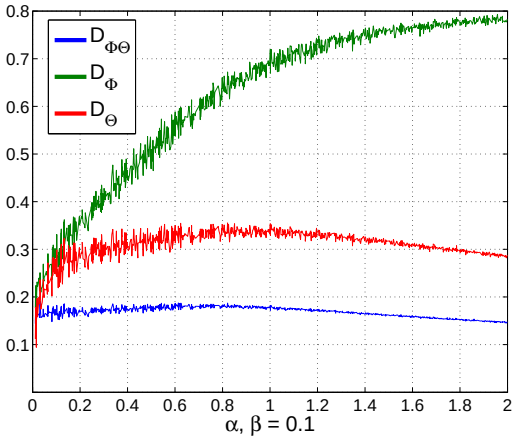
Цель: выявить неустойчивость в зависимости от разреженности данных.



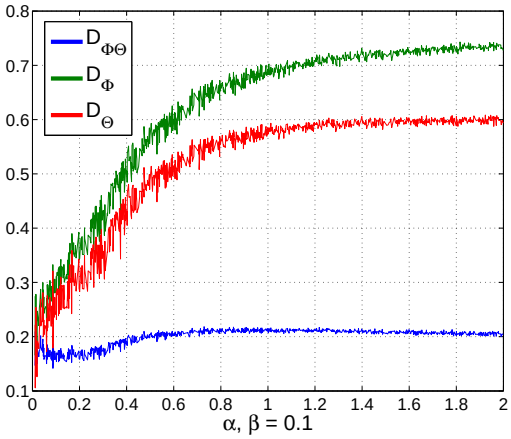
a) LDA-GS



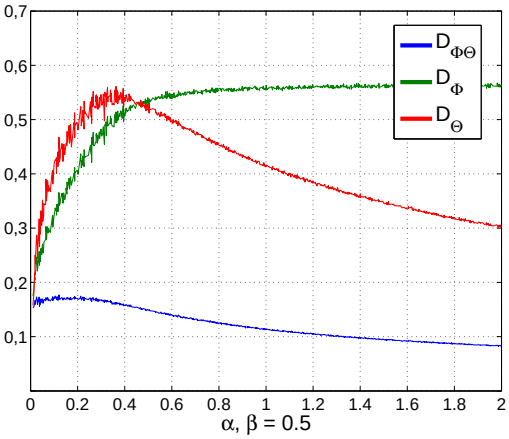
b) PLSA-EM



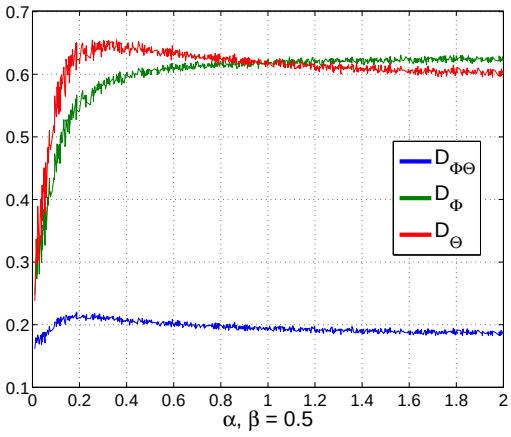
c) LDA-GS



d) PLSA-EM

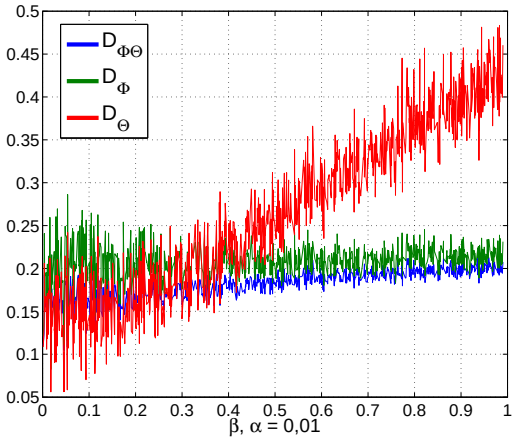


e) LDA-GS

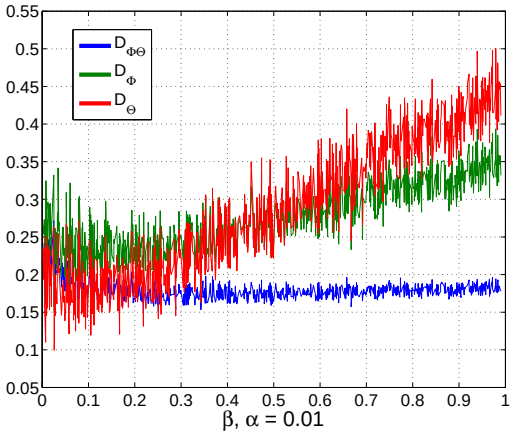


f) PLSA-EM

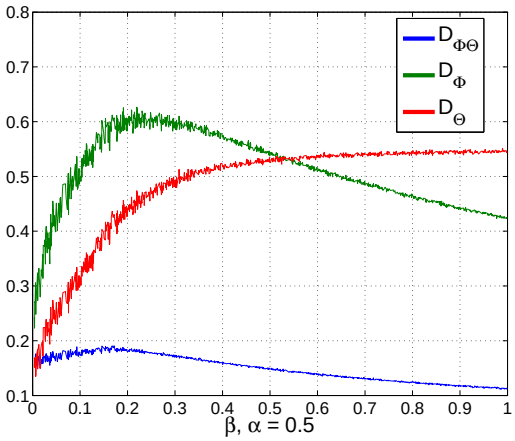
Рис. 3: Сравнение LDA-GS и PLSA-EM при фиксированных β



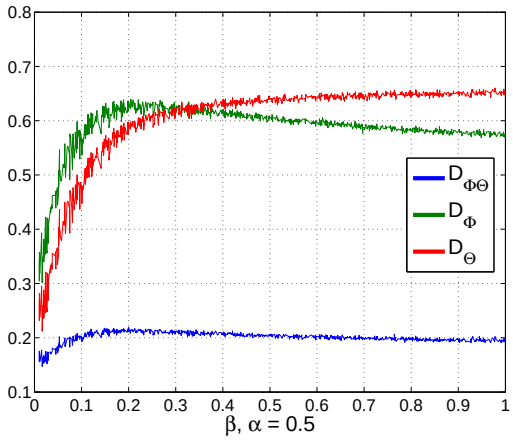
a) LDA-GS



b) PLSA-EM



c) LDA-GS



d) PLSA-EM

Рис. 4: Сравнение LDA-GS и PLSA-EM при фиксированных α

На графиках из 3-4 истинные матрицы Φ, Θ и коллекция порождались при изменяемом одном гиперпараметре (второй фиксирован), и наоборот. Модель восстанавливалась по коллекции при тех же известных истинных параметрах.

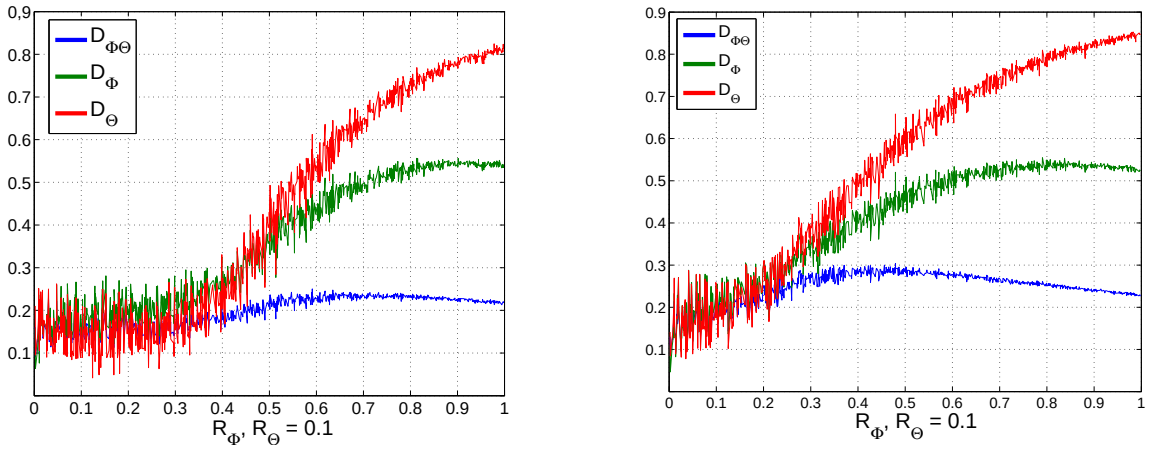
Из рисунка 2 видно, что разреженность начинает сильно расти при гиперпараметрах менее 0.1. При значениях больше 0.4 разреженность резко падает и постепенно переходит во всё более равномерное распределение на матрицах Φ, Θ.

Видно, что LDA-GS и PLSA-EM одинаково хорошо работают в области с сильно разреженными данными. При падении разреженности начинает резко проявляться неустойчивость, а PLSA-EM становится хуже по сравнению с LDA-GS.

6.3 Сравнение EM без и с разреживанием

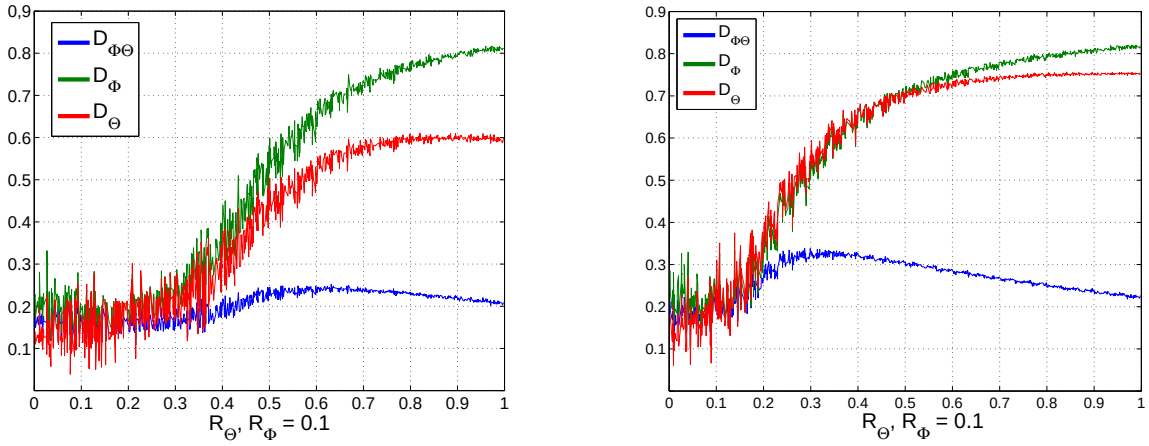
Цель: выявить неустойчивость в зависимости от разреженности данных.

В данном эксперименте R_Φ — доля ненулей в истинной матрице Φ, R_Θ — доля ненулей в истинной матрице Θ. Верхние графики — $R_\Theta = 0.1$ фиксирована, R_Φ изменяется. Нижние — $R_\Phi = 0.1$ фиксирована, R_Θ изменяется.



a) без разреживания b) с разреживанием

Рис. 5: Сравнение без и с разреживанием при фиксированном R_Θ



a) без разреживания b) с разреживанием

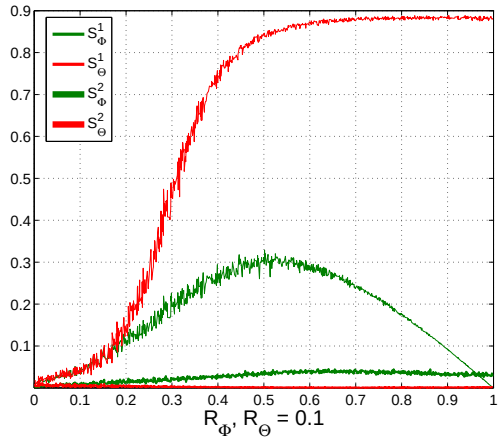
Рис. 6: Сравнение без и с разреживанием при фиксированном R_Φ

Параметры разреживания $i_0 = 10$, $M = 0.1$, $R_\Phi = 0.001$, $R_\Theta = 0.1$.

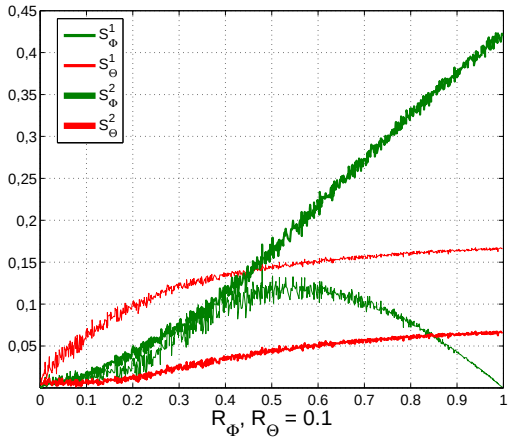
На графиках видно, что стандартный EM и EM с принудительным разреживанием работают хорошо в областях разреженности 70% и более. Далее будет показано, что при этом разреживающий EM лучше восстанавливает структуру разреженности. В остальной области начинает проявляться неустойчивость.

6.4 Качество EM с разреживанием

Цель: показать, что стратегия принудительного разреживания лучше выявляет структуру разреженности.

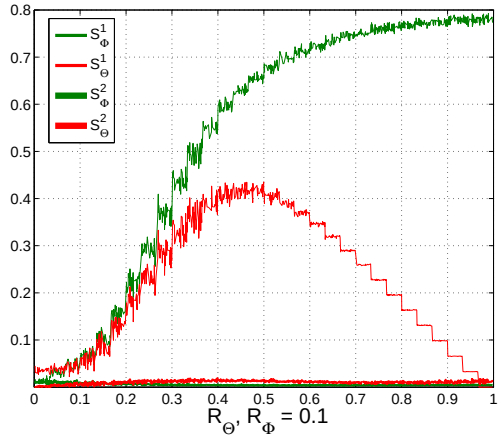


a) без разреживания

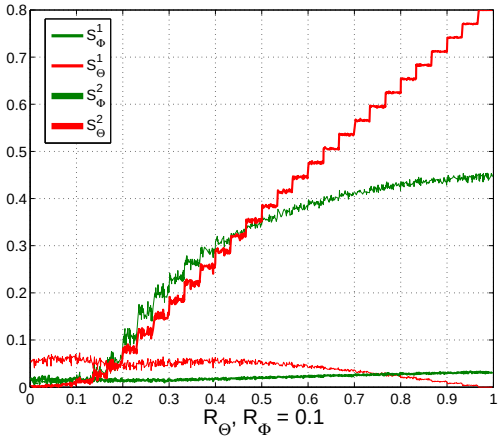


b) с разреживанием

Рис. 7: Сравнение без и с разреживанием при фиксированном $R_\Theta = 0.1$



a) без разреживания

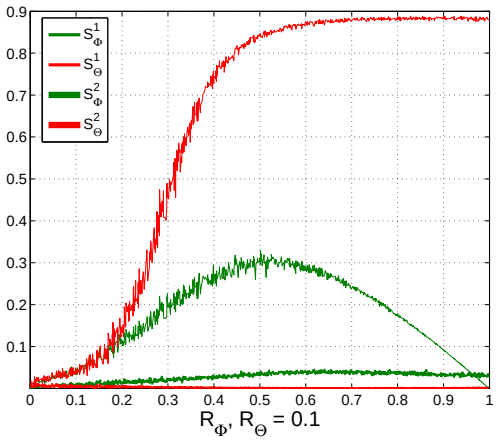


b) с разреживанием

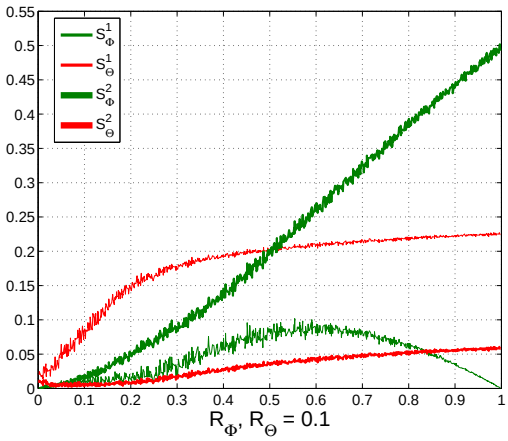
Рис. 8: Сравнение без и с разреживанием при фиксированном $R_\Phi = 0.1$

Параметры разреживания $i_0 = 10$, $M = 0.1$, $S_\Phi = 0.001$, $S_\Theta = 0.1$.

Параметры разреживания $i_0 = 10$, $M = 0.1$, $S_\Phi = 0.0015$, $S_\Theta = 0.075$.

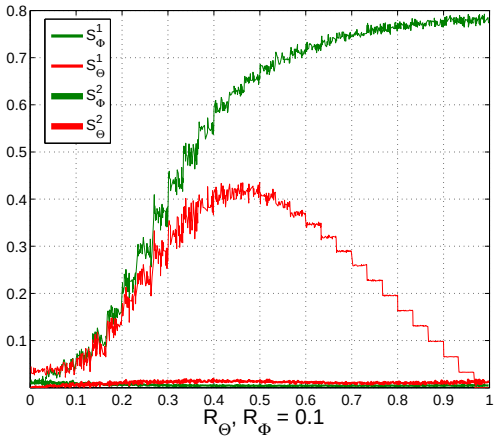


a) без разреживания

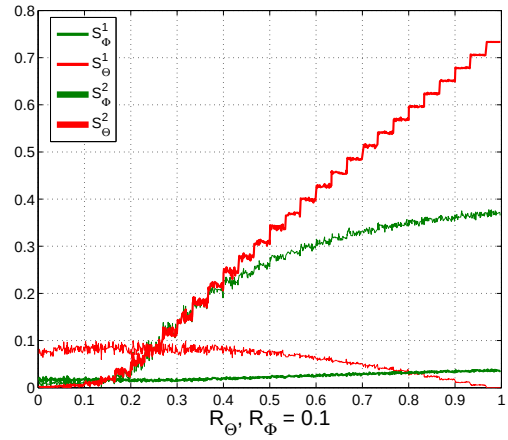


b) с разреживанием

Рис. 9: Сравнение без и с разреживанием при фиксированном $R_\Theta = 0.1$



a) без разреживания



b) с разреживанием

Рис. 10: Сравнение без и с разреживанием при фиксированном $R_\Phi = 0.1$

Из графиков 7-10 по ошибкам 1-ого рода видно, что разреживающий ЕМ действительно лучше восстанавливает структуру разреженности почти на всем интервале. Ошибки 1-ого рода меньше, а значит алгоритм правильнее зануляет элементы. В области разреженности 80% и более ошибки 2-ого рода одинаково малы.

7 Заключение

Проведено исследование по выявлению неустойчивости тематических моделей в зависимости от разреженности данных Предложен простой подход, который заключается в постепенном обнулении наименьших значений вероятностных распределений. Применение данного подхода сохраняет качество модели и оптимизирует структуру разреженности.

Основные выводы:

- В экспериментах на модельных данных показано, что восстановление матриц «тем» Φ и «документов» Θ устойчиво только когда исходные матрицы разрежены на 80% и более.
- Произведение $\Phi\Theta$ восстанавливаются примерно с одинаковой точностью на любых данных.
- Предложен алгоритм разреживания, позволяющий обнулять элементы матриц Φ и Θ в процессе сходимости EM-алгоритма
- Показано, что когда исходные матрицы сильно разрежены, данный алгоритм неплохо восстанавливает структуру разреженности

Список литературы

[1] Hofmann T. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 1999. Pp. 50–57.

[2] Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm // J. of the Royal Statistical Society, Series B. 1977. no. 34. Pp. 1–38.

[3] Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. Vol. 3. Pp. 993–1022.

[4] Steyvers M., Griffiths T. Finding scientific topics // Proceedings of the National Academy of Sciences. 2004. Vol. 101, no. Suppl. 1. Pp. 5228–5235.

[5] Wang Y. Distributed Gibbs sampling of latent dirichlet allocation: The gritty details. 2008.

[6] Y. LeCun and J. Denker and S. Solla and R. E. Howard and L. D. Jackel, Optimal Brain Damage, 1990

[7] J. Munkres, Algorithms for the Assignment and Transportation Problems // Journal of the Society for Industrial and Applied Mathematics, 5(1):32–38, 1957 March.

[8] TextFlow: Towards better understanding of evolving topics in text. // W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, X. Tong // IEEE transactions on visualization and computer graphics. 2011. Vol. 17, no. 12. Pp. 2412–2421.

[9] Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multilabel document classification // Machine Learning. 2012. Vol. 88, no. 1-2. Pp. 157–208.

[10] Yi X., Allan J. A comparative study of utilizing topic models for information retrieval // Advances in Information Retrieval. Springer Berlin Heidelberg, 2009. Vol. 5478 of Lecture Notes in Computer Science. Pp. 29–41.

[11] Yeh J.-h., Wu M.-l. Recommendation based on latent topics and social network analysis // Proceedings of the 2010 Second International Conference on Computer Engineering and Applications. Vol. 1. IEEE Computer Society, 2010. Pp. 209–213.