Московский физико-технический институт
(Государственный университет)

Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛВРА

«Регуляризация вероятностных тематических моделей для повышения устойчивости и интерпретируемости»

Выполнила:
студентка 4 курса 074 группы
Рыскина Мария Никитична

Научный руководитель:
д.ф-м.н.
Воронцов Константин Вячеславович

Москва, 2014

Содержание

Аннотация

В данной работе на модельных и полумодельных данных исследуется устойчивость и интерпретируемость тематических моделей. Оценивается качество сходимости алгоритмов и их комбинаций, качество восстановления исходных матриц и структуры их разреженности. Показано, что на идеальных данных без шума алгоритмы матричного разложения позволяют добиться точной сходимости к исходным данным. В условиях, приближенных к реальным, лучше работают алгоритмы тематического моделирования, и регуляризация позволяет существенно улучшить качество восстановления и устойчивость.

Ключевые слова: *вероятностное тематическое моделирование, устойчивость, интерпретируемость, аддитивная регуляризация*

1 Введение

Актуальность темы. *Тематическое моделирование* — метод решения задачи совместной кластеризации терминов и документов некторой текстовой коллекции. Он применяется для извлечения необходимой информации из текстовых документов или коллекций большого объёма, например, в задачах тематического поиска [1] или анализа трендов [2]. *Вероятностная тематическая модель* восстанавливает два дискретных вероятностных распределения — слов по выделенным кластерам (темам) и тем по документам. При этом считается, что коллекция порождается случайным и независимым выбором терминов из смеси этих распределений. Таким образом, решается задача восстановления смеси распределений, которая имеет неединственное решение — в зависимости от начального приближения результат восстановления может быть разным.

Но так как в задаче тематического моделирования компоненты смесей интерпретируются как темы в текстовых документах, из бесконечного множества решений задачи требуется выбрать наиболее интерпретируемое с человеческой точки зрения решение. Поэтому важна *устойчивость* модели: нужно достичь сходимости к фиксированному решению из любого начального приближения. Кроме того, получаемое решение должно быть наиболее *интерпретируемым*. Таким образом, появляется также задача формализации понятия интерпретируемости с целью уменьшить необходимость вмешательства экспертов в оценку качества модели.

Все эксперименты в данной работе проводились на модельных данных, так как в экспериментах на реальных данных неизвестны истинные распределения, и нет возможности оценить качество их восстановления. Модельные данные были сгенерированы в соответствии со сделанными предположениями о структуре интерпретируемых тем.

Цель работы. Целью данной работы является формализация понятия интерпретируемости построенной модели, исследование сходимости и устойчивости алгоритма PLSA-EM, а также проверка гипотезы о том, что аддитивная регуляризация приближает решение к интерпретируемому.

Обзор литературы. Глобально понятие интерпретируемости означает, что результат машинной кластеризации множеств слов и документов должен быть максимально похож на результат экспертной кластеризации [3].

Наиболее используемые в тематическом моделировании меры качества, такие как перплексия или правдоподобие контрольной выборки [4], характеризуют только качество построения модели, но не учитывают семантические содержательные свойства результатов кластеризации (это показано, например, в [3]). Поэтому интрепретируемость является одной из главных характеристик качества тематической модели, но из-за отсутствия чёткого определения не существует однозначной методики её измерения. Использование для этого экспертной оценки приводит к необходимости просматривать вручную большие объёмы текста. Поэтому необходима автоматизированная методика, где эксперт размечает коллекцию один раз, и эти данные используются для последующей оценки интерпретируемости моделей, построенных на этой коллекции.

Ранее использовались различные методики измерения интерпретируемости, основной из которых была интрузия (впервые предложено в [3]) — внедрение в группу слов, характерных для получившейся темы, слова, не относящегося к ней. Если эксперты определяли добавленное слово как лишнее в большой доле случаев, тема признавалась хорошо интерпретируемой. Аналогично используется тематическая интрузия: эксперту предоставляется документ и группа тем, в которой все, кроме одной, имеют высокую вероятность в данном документе. Но так как проведение такого эксперимента тоже требует больших затрат времени и усилий асессоров, в [5] был предложен автоматизированный метод оценивания интерпретируемости: было показано, что функционал когерентности (PMI) коррелирует с ответами экспертов и может быть использован в качестве меры интерпретируемости. Этот функционал широко используется сейчас для оценки интерпретируемости, например, в задачах topic model labeling [6]. Были предложены различные модификации функционала PMI, например, в [7] и [8].

Таким образом, методика повышения интерпретируемости зависит от того, как именно определять это понятие. В вышеупомянутых работах использовались экспертные измерения качества, а здесь мерой интерпретируемости служит мера бли-

зости к интерпретируемым модельным данным.

2 Постановка задачи

2.1 Исходные гипотезы

Рассмотрим стандартную задачу тематического моделирования: дана коллекция документов D, где каждый документ d представляет собой последовательность слов $W_d = (w_1, \ldots, w_{n_d})$ из словаря W. Число вхождений слова w в документ d обозначим как n_{dw}.

Гипотеза о вероятностном пространстве: существует некоторое множество тем T, и каждое слово в кажом документе связано с некоторой темой из этого множества. Таким образом, вся коллекция рассматривается как набор троек (d, w, t) из $D \times W \times T$, где переменные d и w наблюдаемые, а t – скрытые.

Гипотеза «мешка слов»: порядок слов в каждом из документов не важен:

$$p\left(\{d, w_i\}_{i=1}^{n_d}\right) = \prod_{i=1}^{n_d} p\left(\{d, w_i\}\right)$$

Гипотеза условной независимости: вероятность порождения слова w в документе d темой t зависит только от темы, но не от самого документа:

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

Вероятностное порождение коллекции. Предполагается, что при порождении коллекции для каждого словоместа в документе d сначала из распределения тем в документе $p\left(t|d\right)$ сэмплируется тема, а затем из распределения слов в этой теме $p\left(w|t\right)$ сэмплируется слово. Таким образом, в соответствии с гипотезой условной независимости:

$$p\left(w|d\right) = \sum_{t \in T} p\left(w|t\right) p\left(t|d\right)$$

2.2 Задача тематического моделирования

Постановка задачи. По имеющейся коллекции D нужно восстановить скрытые распределения $p(w|t)$ для всех $t \in T$ и $p(t|d)$ для всех $d \in D$.

Матричная формулировка: Коллекцию можно рассматривать в виде стохастической матрицы $\mathbf{F}$ распределения слов по документам. Требуется представить имеющуюся матрицу в виде

$$\mathbf{F} \approx \mathbf{\Phi} \cdot \mathbf{\Theta},$$

где $\mathbf{\Phi} = [\phi_{ij}] = [p(w_i|t_j)]$, $w_i \in W$, $t_j \in T$ — матрица распределения слов по темам, а $\mathbf{\Theta} = [\theta_{ij}] = [p(t_i|d_j)]$, $t_i \in T$, $d_j \in D$ — матрица распределения тем по документам. Далее для упрощения будем w, t и d использовать в качестве индексов для матриц $\mathbf{\Phi}$ и $\mathbf{\Theta}$.

Тогда в этой задаче, как и в любой задаче матричного разложения, имеет место проблема неединственности разложения:

$$\mathbf{\Phi\Theta} = (\mathbf{\Phi S})\left(\mathbf{S}^{-1}\mathbf{\Theta}\right) = \mathbf{\Phi}'\mathbf{\Theta}'$$

Эта проблема может привести к сходимости к локальным оптимумам и неустойчивости решения.

2.3 Модельные данные

Все эксперименты в этой работе ставятся на модельных или полумодельных данных. Это значит, что были сгенерированы определённые модельные матрицы $\mathbf{\Phi}_0$ и $\mathbf{\Theta}_0$, а затем по ним восстанавливалась коллекция. С помощью алгоритмов тематического моделирования восстановленная матрица коллекции раскладывается в произведение матриц $\mathbf{\Phi}$ и $\mathbf{\Theta}$, а затем исследуется близость полученных матриц к модельным.

Для определения близости матриц вычислялось расстояние Хеллингера между профилями тем — столбцами матриц $\mathbf{\Phi}$ и $\mathbf{\Phi}_0$:

$$H(\mathbf{p}, \mathbf{p}_0) = \frac{1}{m}\sum_{j=1}^{m}\sqrt{\frac{1}{2}\sum_{i=1}^{n}\left(\sqrt{\mathbf{p}(i|j)} - \sqrt{\mathbf{p}_0(i|j)}\right)^2}$$

и с помощью венгерского алгоритма определялось взаимно-однозначное соответствие между темами в новой и исходной моделях (так как матрицы восстанавливаются с

точностью до перестановки тем). Венгерский алгоритм определяет перестановочную матрицу $\mathbf{\Pi}$, минимизирующую функционал

$$f(\mathbf{\Pi}) = D_{\mathbf{\Phi}}(\mathbf{\Phi}_0\mathbf{\Pi}, \mathbf{\Phi}),$$

где

$$D_{\mathbf{\Phi}}(\mathbf{\Phi}_0, \mathbf{\Phi}) = H(\mathbf{\Phi}_0, \mathbf{\Phi}).$$

Мы хотим добиться устойчивости, то есть сходимости решения из любого начального приближения к исходным матрицам (с учётом перестановки тем). Таким образом, задача сводится к одновременной минимизации функционалов:

$$D_{\mathbf{\Phi}}(\mathbf{\Phi}_0, \mathbf{\Phi}) = H(\mathbf{\Phi}_0, \mathbf{\Phi}) \to \min$$

$$D_{\mathbf{\Theta}}(\mathbf{\Theta}_0, \mathbf{\Theta}) = H(\mathbf{\Theta}_0, \mathbf{\Theta}) \to \min$$

$$D_{\mathbf{\Phi\Theta}}(\mathbf{\Phi}_0\mathbf{\Theta}_0, \mathbf{\Phi\Theta}) = H(\mathbf{\Phi}_0\mathbf{\Theta}_0, \mathbf{\Phi\Theta}) \to \min$$

Дополнительные предположения о структуре данных. Так как мы хотим получить наиболее интерпретируемое решение, сделаем некоторые предположения о его структуре. Положим, что в случае интерпретируемого решения темы должны быть сильно декоррелированы, и в каждой из них только малое количество слов имеет существенно отличные от нуля вероятности (они составляют *тематическое ядро* темы). Также полагаем, что в каждом документе может присутствовать только небольшое количество тем, то есть матрица $\mathbf{\Theta}$ является разреженной.

Но все указанные допущения относятся только к *предметным* темам, а помимо них в коллекции также присутствуют *фоновые* темы. В этих темах содержатся слова общей лексики, но и тематические слова встречаются в них с равной вероятностью. Фоновые темы равномерно распределены по всем документам коллекции.

Основная гипотеза, проверяемая в данной работе: в сделанных предположениях о коллекции и структуре интерпретируемых данных восстановить искомое решение можно с помощью *аддитивной регуляризации*.

3 Используемые алгоритмы

3.1 Вероятностный латентный семантический анализ

PLSA (Probabilistic Latent Semantic Analysis, [9]) — классический алгоритм тематического моделирования. EM-алгоритм решает задачу максимизации правдоподобия:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max$$

Каждая итерация алгоритма состоит из двух шагов. Перед первой итерацией выбираются начальные значения параметров $\phi_{wt},\ \theta_{td}$.

На E-шаге по текущим значениям $\phi_{wt},\ \theta_{td}$ по формуле Байеса вычисляются условные вероятности $p(t|d,w)$ для всех тем $t \in T$ и всех пар «слово-документ» $(w,d),\ w \in d$:

$$H_{dwt} = p(t|w,d) = \frac{p(w|t)\,p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s \in T} \phi_{ws}\theta_{sd}}.$$

На M-шаге по условным вероятностям тем H_{dwt} вычисляется новое приближение $\phi_{wt},\ \theta_{td}$. Для этого вычисляется оценка $\hat{n}_{dwt} = n_{dw} H_{dwt}$ и по ней рассчитываются приближения параметров:

$$\phi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \quad \hat{n}_t = \sum_{w \in W} \hat{n}_{wt}, \quad \hat{n}_{wt} = \sum_{d \in D} \hat{n}_{dw} H_{dwt}.$$

$$\theta_{td} = \frac{\hat{n}_{dt}}{\hat{n}_d}, \quad \hat{n}_d = \sum_{t \in T} \hat{n}_{dt}, \quad \hat{n}_{dt} = \sum_{w \in d} \hat{n}_{dw} H_{dwt}.$$

В данном случае использовался рациональный EM-алгоритм, в котором E-шаг встроен внутрь M-шага.

Матричная реализация алгоритма позволяет ускорить его выполнение в 100 раз (среда MATLAB). Введём обозначения:

$$\mathbf{N} \in \mathbb{R}^{|W| \times |D|} : \mathbf{N}[w,d] = n_{dw};$$

$$\mathbf{Q} \in \mathbb{R}^{|W| \times |D|} : \mathbf{Q} = \mathbf{N} \oslash (\mathbf{\Phi} \cdot \mathbf{\Theta})$$

$$\mathbf{R}_{wt} \in \mathbb{R}^{|W| \times |T|} : \mathbf{R}_{wt} = \mathbf{\Phi} \otimes \left(\mathbf{Q} \cdot \mathbf{\Theta}^T\right)$$

$$\mathbf{R}_{td} \in \mathbb{R}^{|T| \times |D|} : \mathbf{R}_{td} = \left(\mathbf{\Phi}^T \cdot \mathbf{Q}\right) \otimes \mathbf{\Theta}$$

Здесь $\oslash$ и $\otimes$ обозначают операции поэлементного деления и умножения.

Заметим, что

$$\mathbf{Q}\left[w,d\right]=\frac{n_{dw}}{\sum_{s\in T}\phi_{ws}\theta_{sd}};$$

$$\mathbf{R}_{wt}\left[w,t\right]=\phi_{wt}\sum_{d\in D}\left[\theta_{td}\left(\frac{n_{dw}}{\sum_{s\in T}\phi_{ws}\theta_{sd}}\right)\right]=\sum_{d\in D}\left[n_{dw}\left(\frac{\phi_{wt}\theta_{td}}{\sum_{s\in T}\phi_{ws}\theta_{sd}}\right)\right]=\sum_{d\in D}\left[n_{dw}H_{dwt}\right]=\hat{n}_{wt}$$

$\mathbf{R}_{td}\left[t,d\right]=\hat{n}_{td}$ аналогично.

Тогда итерация EM-алгоритма примет вид:

$$\mathbf{Q}\leftarrow\mathbf{N}\oslash\left(\mathbf{\Phi}\cdot\mathbf{\Theta}\right);$$

$$\mathbf{R}_{wt}\leftarrow\mathbf{\Phi}\otimes\left(\mathbf{Q}\cdot\mathbf{\Theta}^{T}\right);\qquad\mathbf{R}_{td}\leftarrow\left(\mathbf{\Phi}^{T}\cdot\mathbf{Q}\right)\bullet\mathbf{\Theta};$$

$$\mathbf{\Phi}\leftarrow\Xi\mathbf{R}_{wt};\qquad\mathbf{\Theta}\leftarrow\Xi\mathbf{R}_{td};$$

где Ξ — оператор нормировки матрицы по столбцам.

3.2 Аддитивная регуляризация

Регуляризованный EM-алгоритм [10] решает задачу максимизации регуляризованного правдоподобия

$$L'\left(\mathbf{\Phi},\mathbf{\Theta}\right)=\sum_{d\in D}\sum_{w\in d}n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td}+R\left(\mathbf{\Phi},\mathbf{\Theta}\right)\rightarrow\max$$

$$R\left(\mathbf{\Phi},\mathbf{\Theta}\right)=\sum_{i=1}^{n}\tau_{i}R_{i}\left(\mathbf{\Phi},\mathbf{\Theta}\right)$$

при тех же ограничениях; τ_i называются коэффициентами регуляризации.

Тогда изменятся формулы M-шага:

$$\phi_{wt}\propto\left(\hat{n}_{wt}+\phi_{wt}\frac{\partial R}{\partial\phi_{wt}}\right)_{+};\qquad\hat{n}_{wt}=\sum_{d\in D}n_{dw}H_{dwt}$$

$$\theta_{td}\propto\left(\hat{n}_{td}+\theta_{td}\frac{\partial R}{\partial\theta_{td}}\right)_{+};\qquad\hat{n}_{td}=\sum_{w\in d}n_{dw}H_{dwt};$$

Используемые в данной работе регуляризаторы:

- декорреляция Φ:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T, s \neq t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max;$$

- разреживание Φ:

$$R(\Phi) = -\beta_0 \sum_{t \in T} \sum_{w \in W} \ln \phi_{wt} \to \max;$$

- разреживание Θ:

$$R(\Theta) = -\alpha_0 \sum_{d \in D} \sum_{t \in T} \ln \theta_{td} \to \max;$$

- регуляризатор частичного обучения:

$$R\left(\Phi\right) = \sum_{t \in T} \left(\tau_+ \sum_{w \in W_{t+}} \phi_{wt} - \tau_- \sum_{w \in W_{t-}} \phi_{wt} \right) \to \max,$$

  где $W_{t+} \in W$ и $W_{t-} \in W$ — некоторые экспертно заданные множества тематических («белых») и фоновых («чёрных») слов для темы t.

Первые три регуляризатора используют сделанные нами предположения о структуре интерпретируемых тем, а последний — некоторую дополнительную информацию о темах, полученную от экспертов. Будет оценено влияние этих факторов на качество восстановления исходных матриц.

4 Вычислительные эксперименты

4.1 Нерегуляризованный алгоритм, модельные данные

Сначала покажем, что нерегуляризованный EM-алгоритм действительно не решает проблему неустойчивости.

Модельные данные: матрица $\mathbf{\Phi}_0$ распределения слов по темам является блочно-диагональной с непересекающимися темами (фоновые темы исключены), столбцы внутри блоков сгенерированы из равномерного распределения. Матрица $\mathbf{\Theta}_0$ распределения тем по документам — разреженная. Параметры эксперимента: $|W| = 1000$,

$|D| = 500$, $|T| = 20$, $n_d = 500$ для всех $d \in D$. Заметим, что коллекция восстанавливалась простым перемножением $\mathbf{\Phi} \cdot \mathbf{\Theta}$, то есть шум в коллекции отсутствует.

Рассмотрим влияние начального приближения на сходимость. Было сгенерировано 20 пар начальных приближений $\mathbf{\Phi}$ и $\mathbf{\Theta}$ из равномерного распределения. На графиках 1 показана сходимость $D_{\mathbf{\Phi}}\left(\mathbf{\Phi}_0, \mathbf{\Phi}\right)$, $D_{\mathbf{\Theta}}\left(\mathbf{\Theta}_0, \mathbf{\Theta}\right)$, $D_{\mathbf{\Phi\Theta}}\left(\mathbf{\Phi}_0\mathbf{\Theta}_0, \mathbf{\Phi\Theta}\right)$.

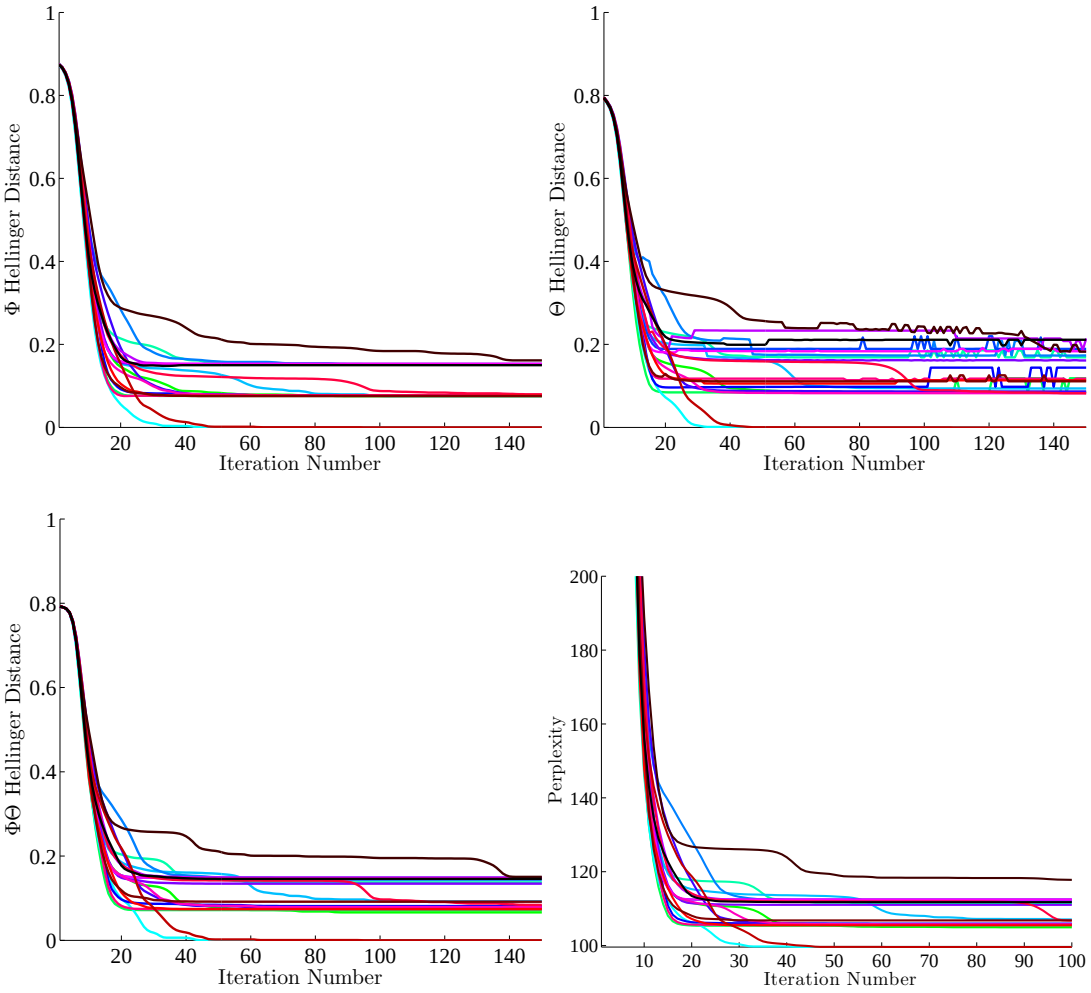

Рис. 1: Сходимость расстояния Хеллингера для матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$, $\mathbf{\Phi\Theta}$ и перплексии модели соответственно (модельные данные)

Итак, мы наблюдаем «застревание» в локальных оптимумах, и лишь малое число приближений даёт возможность восстановить истинные распределения. Наиболее вероятная причина несходимости — обращение в 0 тех элементов матриц, которые не равнялись 0 в модельных матрицах. После этого эти элементы перестают изменяться, т. е. решение «застревает» в локальном оптимуме. Оценим долю ошибки, приходящуюся на обнуление таких элементов. На графиках 2 представлены расстояния между модельными и выходными матрицами, измеренные по метрике городских кварталов: $D_{\mathbf{\Phi}}\left(\mathbf{\Phi}_0, \mathbf{\Phi}\right)$, $D_{\mathbf{\Theta}}\left(\mathbf{\Theta}_0, \mathbf{\Theta}\right)$ по всем элементам и $D^{small}_{\mathbf{\Phi}}\left(\mathbf{\Phi}_0, \mathbf{\Phi}\right)$, $D^{small}_{\mathbf{\Theta}}\left(\mathbf{\Theta}_0, \mathbf{\Theta}\right)$ по стремящимся к нулю. Используется одно из плохо сходящихся начальных приближений.

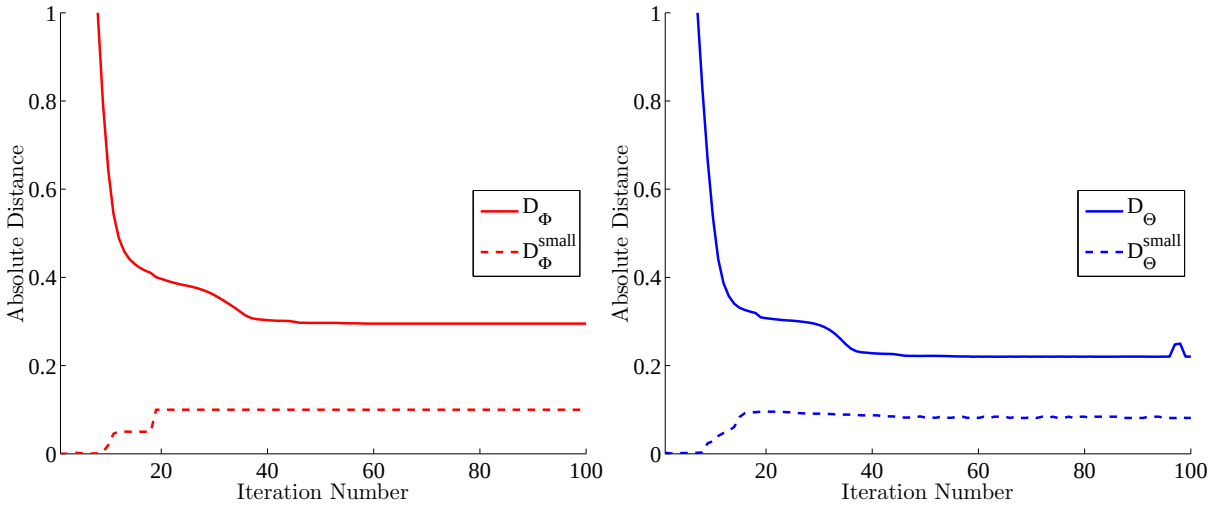

Рис. 2: Сходимость расстояния городских кварталов для матриц $\mathbf{\Phi}$ и $\mathbf{\Theta}$ (модельные данные)

Для устранения этого эффекта можно применить следующий метод [11]: отделять от нуля те элементы матриц, в которых градиент правдоподобия положителен, то есть нарушается условие стационарности. Для этого, когда они достигают малых значений, их значения заменяются некой малой положительной константой. Результаты сходимости метода представлены на графиках 3. Как можно видеть, этот метод не помогает улучшить сходимость.

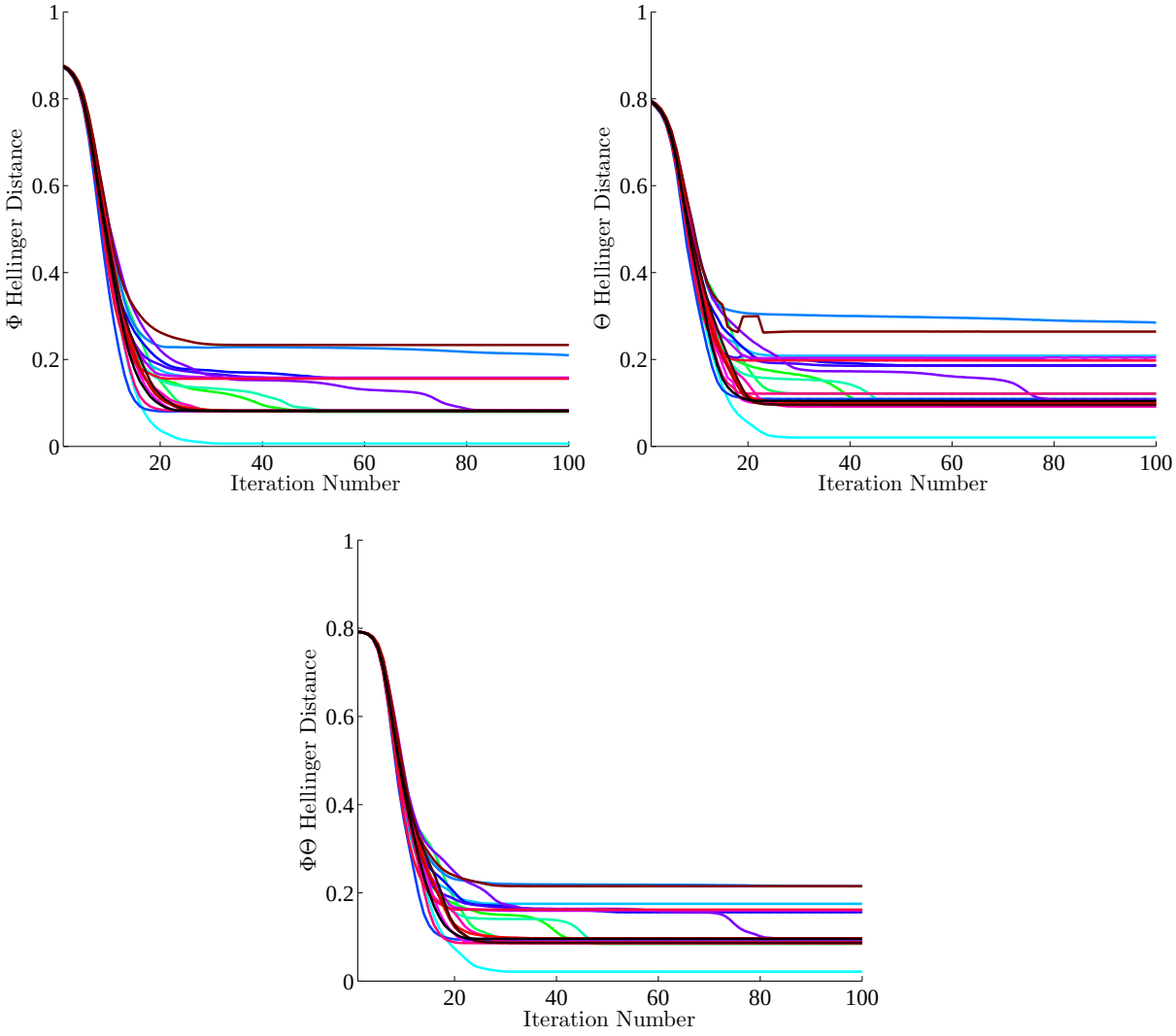

Рис. 3: Сходимость расстояния Хеллингера для матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$ и $\mathbf{\Phi\Theta}$ соответственно (модельные данные, метод [11])

Воспользуемся алгоритмом ALS (Alternating Least Squares, [12]). Это алгоритм неотрицательного матричного разложения, который на каждой итерации решает задачу наименьших квадратов $\min \|\mathbf{F} - \mathbf{\Phi\Theta}\|_F$, где $\|A\|_F = \sqrt{\sum a_{ij}^2}$ — норма Фробениуса, следующим образом:

$$\mathbf{\Phi} \leftarrow \max\left(0,\ \underset{Z \in \mathbb{R}^{|W| \times |T|}}{\arg\min}\ \|\mathbf{F} - \mathbf{Z\Theta}\|_F\right)$$

$$\mathbf{\Theta}^T \leftarrow \max\left(0,\ \underset{Y \in \mathbb{R}^{|D| \times |T|}}{\arg\min}\ \left\|\mathbf{F}^T - \mathbf{Y\Phi}^T\right\|_F\right)$$

Так как ALS не учитывает стохастичности матриц, добавим принудительную нормировку $\mathbf{\Phi}$ и $\mathbf{\Theta}$ после каждой итерации алгоритма (нормируем столбцы матриц и полученные стохастические матрицы подаём на вход следующей итерации).

Добавим 15 итераций ALS с нормировкой перед EM-алгоритмом и посмотрим, как это влияет на сходимость. Из рисунка 4 видно, что все приближения сходятся к очень близким к 0 значениям. Те же эксперименты для 5 и 10 итераций ALS не дали такого эффекта (некоторые приближения продолжили застревать в локальных оптимумах). Вероятнее всего, EM-алгоритм будет всегда сходиться хорошо, если с помощью ALS подойти достаточно близко к модельным матрицам.

Теперь попробуем совсем убрать PLSA и пользоваться только методами матричного разложения (рис. 5). Для сходимости расстояний Хеллингера ничего не меняется.

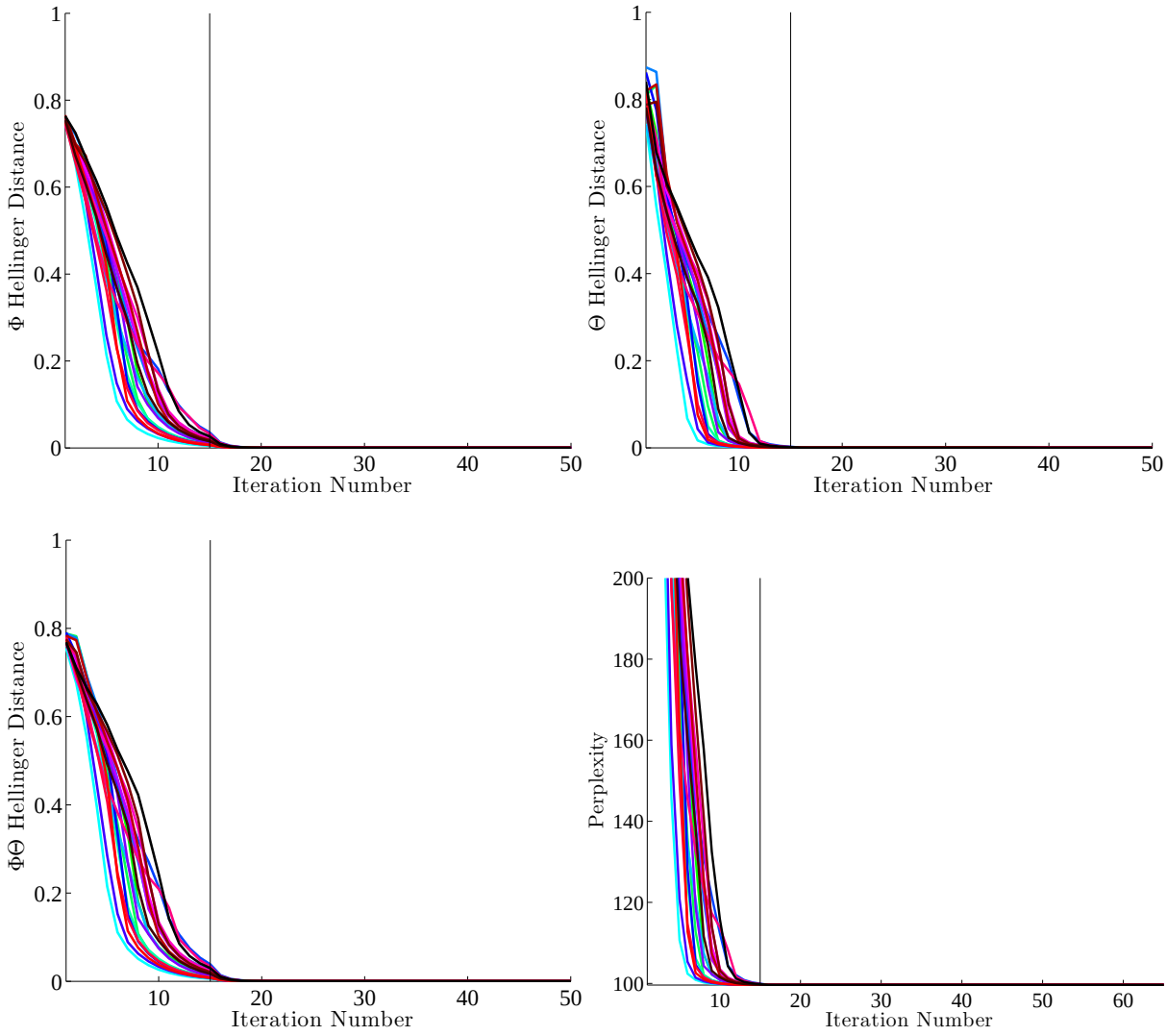

Рис. 4: Сходимость расстояний Хеллингера для матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$ и $\mathbf{\Phi\Theta}$ и перплексии соответственно (модельные данные, 15ALS+PLSA); вертикальная черта обозначает остановку ALS и начало работы EM.

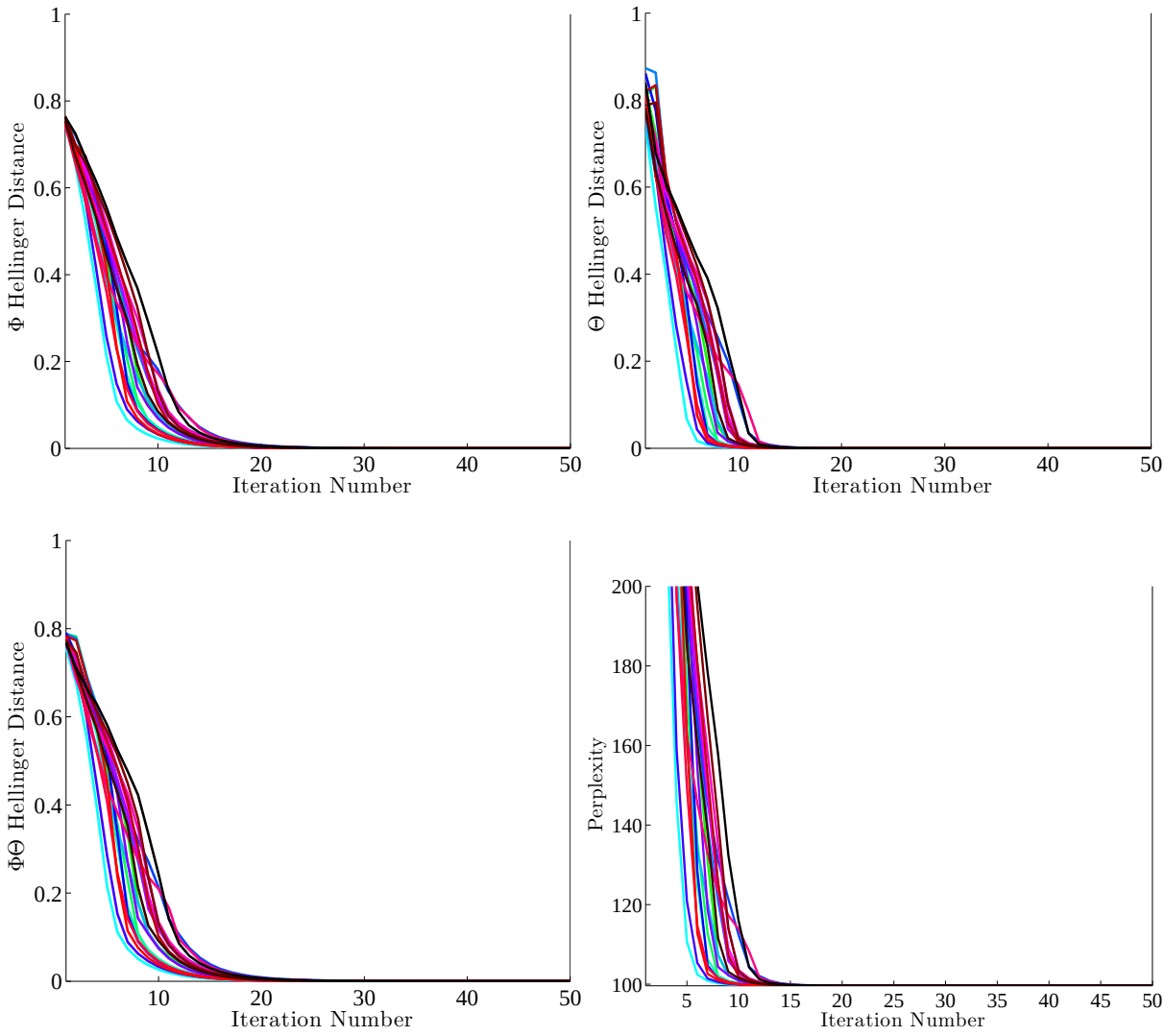

Рис. 5: Сходимость расстояния Хеллингера для матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$ и $\mathbf{\Phi\Theta}$ и перплексии соответственно (модельные данные, 50ALS)

Изучим также восстановление структуры разреженности. Рассмотрим два типа ошибок восстановления структуры. Ошибкой первого рода, или ошибкой $0 \to 1$, назовём ситуацию, когда нулевой элемент исходной матрицы становится ненулевым в восстановленной. Ошибка второго рода, или $1 \to 0$ — противоположная ситуация.

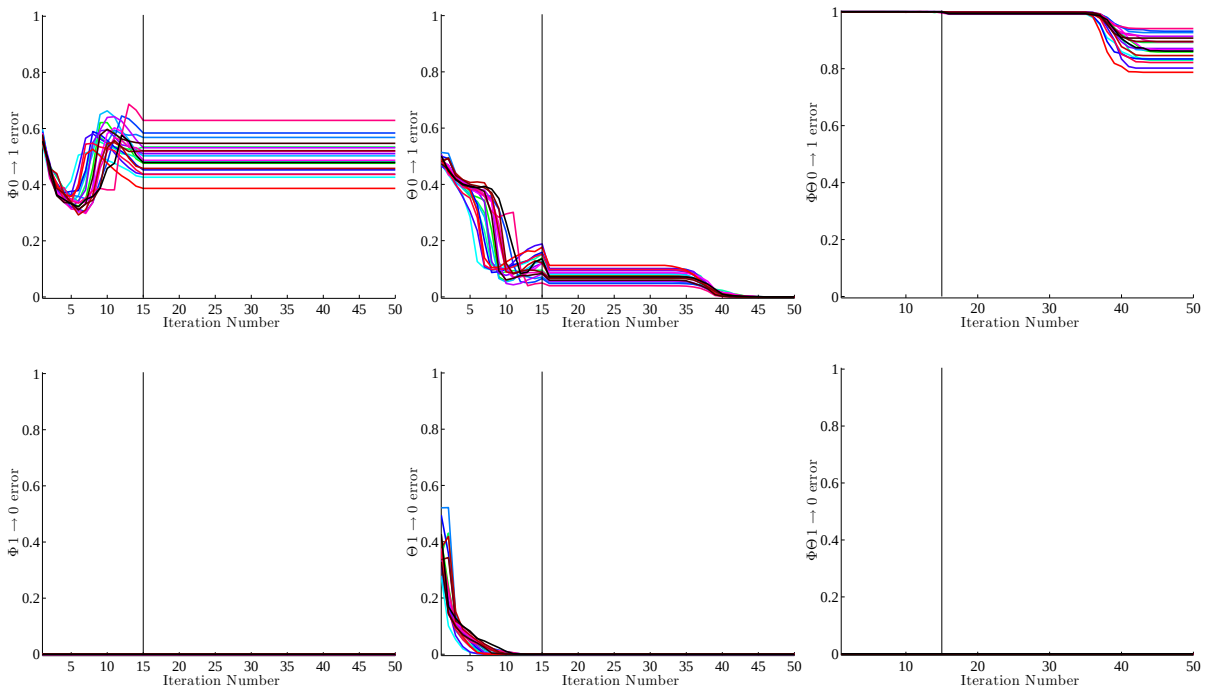

Рис. 6: Доля ошибок первого и второго рода при восстановлении структуры разреженности матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$ и $\mathbf{\Phi\Theta}$ соответственно (модельные данные, 15ALS+PLSA)

При переходе к ALS качество восстановления структуры разреженности ухудшилось (график 7). При сравнении с графиком 6 видно, что EM-алгоритм помогает снизить ошибку 1 рода для матриц $\mathbf{\Theta}$ и $\mathbf{\Phi\Theta}$.

Вывод: Итак, мы получаем, что для модельных данных без шума сходимости к исходному разложению из почти любого начального приближения можно достичь переходом к методу ALS, или выполнением нескольких итераций ALS перед запуском EM-алгоритма. Это решает проблему неустойчивости и застревания в локальных оптимумах. Но нужно проверить, связано ли это с существованием точного разложения или со структурой конкретных модельных данных.

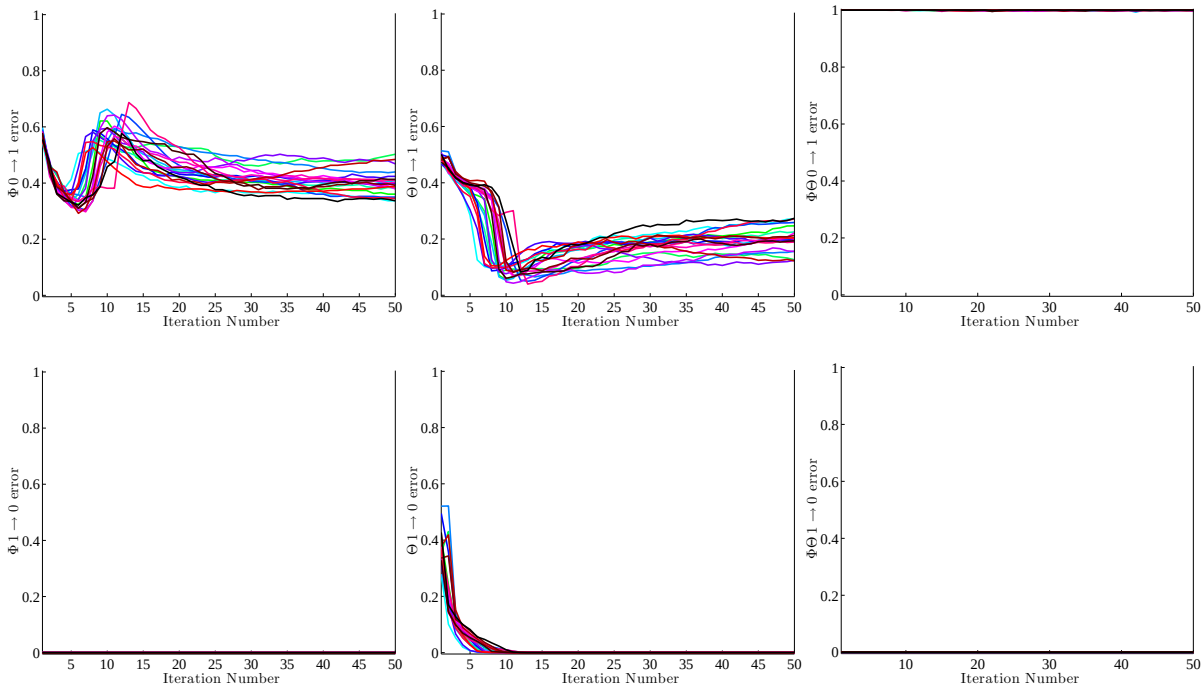

Рис. 7: Доля ошибок первого и второго рода при восстановлении структуры разреженности матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$ и $\mathbf{\Phi\Theta}$ соответственно (модельные данные, 50ALS)

4.2 Нерегуляризованный алгоритм, полумодельные данные

Полумодельные данные: В качестве $\mathbf{\Phi}_0$ и $\mathbf{\Theta}_0$ возьмём фрагменты матриц $\mathbf{\Phi}$ и $\mathbf{\Theta}$, полученных с помощью PLSA-EM в реальном эксперименте на коллекции NIPS [10]. Параметры эксперимента остались теми же: $|W| = 1000$, $|D| = 500$, $|T| = 20$, $n_d = 500$ для всех $d \in D$. Теперь матрицы стали менее разреженными и декоррелированными.

Проделаем те же эксперименты: из 10 начальных приближений запустим нерегуляризованный EM-алгоритм. Сходимости снова нет (рис. 8).

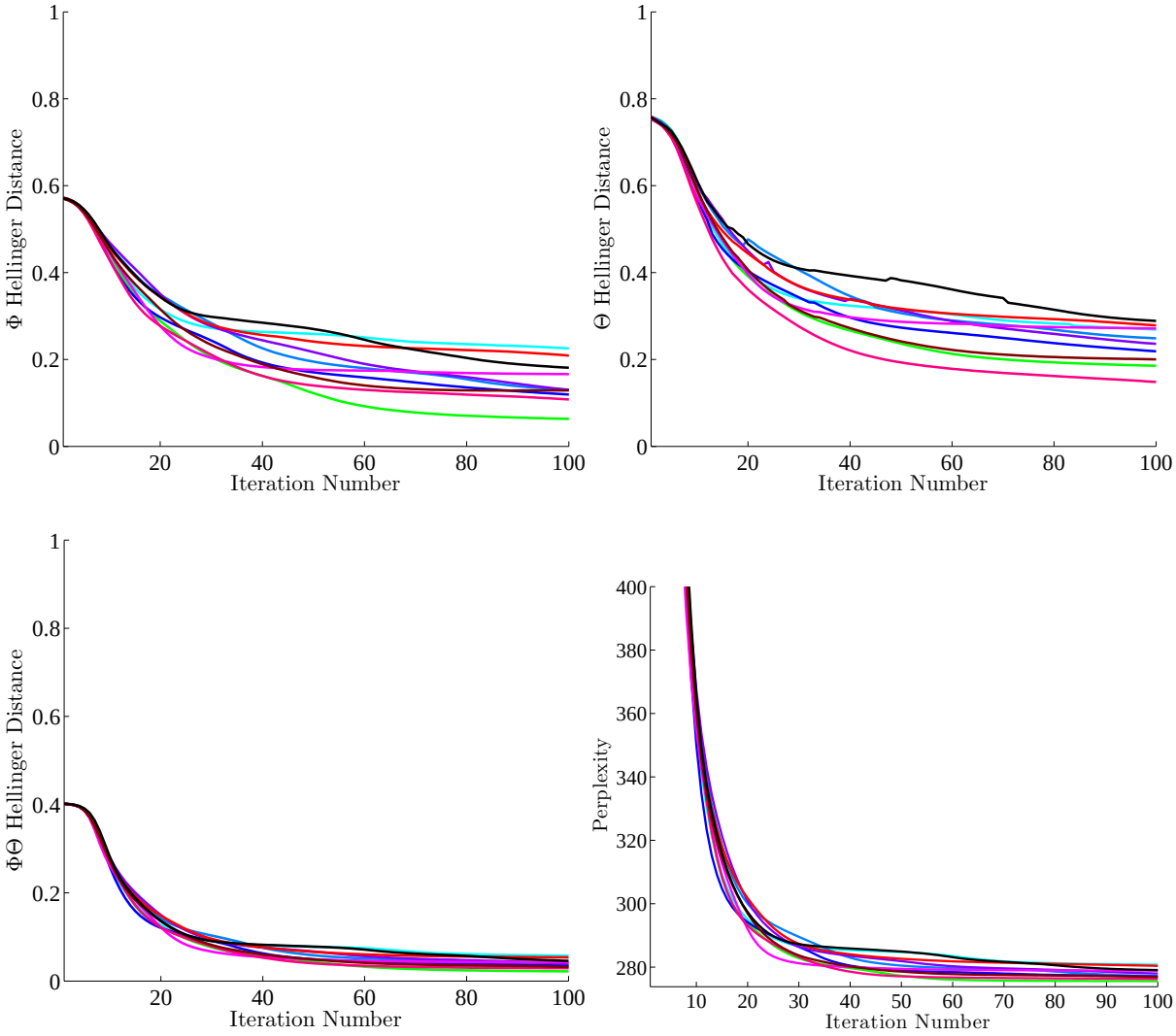

Рис. 8: Сходимость расстояния Хеллингера для матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$ и $\mathbf{\Phi\Theta}$ и перплексии соответственно (полумодельные данные)

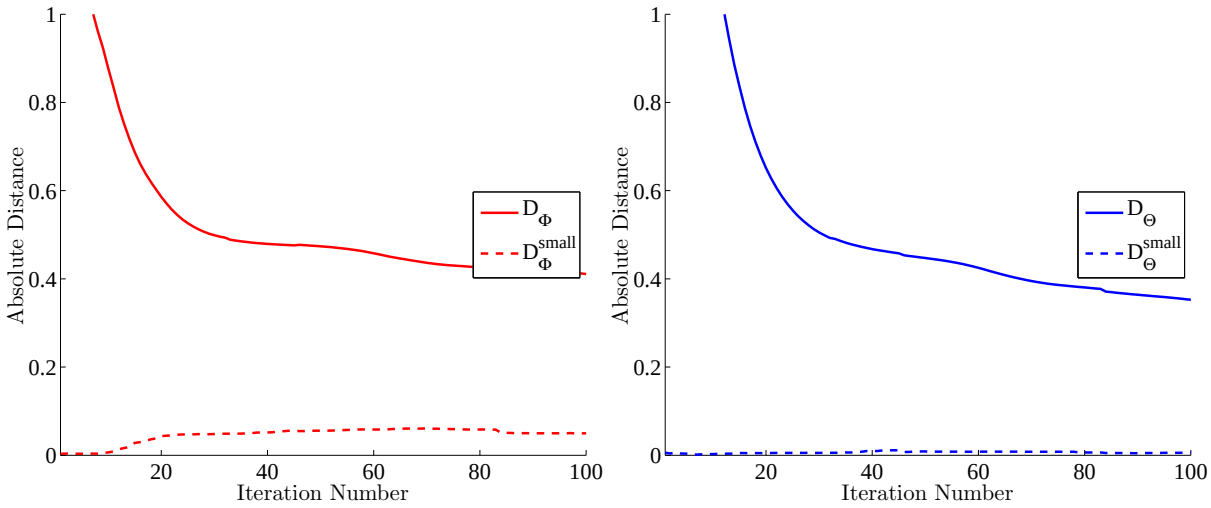

Рис. 9: Сходимость расстояния городских кварталов для матриц $\mathbf{\Phi}$ и $\mathbf{\Theta}$ (полумодельные данные)

Из графиков 9 видно, что в случае неразреженных данных обращение в 0 составляет лишь малую долю ошибки.

Добавление 50 итераций ALS (графики 10) существенно улучшает сходимость.

Теперь посмотрим на результат работы только ALS (графики 11). Видно, что при достаточном количестве итераций ALS достигается сходимость к модельным данным ($D_{\mathbf{\Phi}}\left(\mathbf{\Phi}_0, \mathbf{\Phi}\right), D_{\mathbf{\Phi\Theta}}\left(\mathbf{\Phi}_0\mathbf{\Theta}_0, \mathbf{\Phi\Theta}\right) \approx 10^{-3}$, $D_{\mathbf{\Theta}}\left(\mathbf{\Theta}_0, \mathbf{\Theta}\right) \approx 10^{-2}$). Перплексия также сходится, хотя на начальных итерациях она может принимать бесконечные значения (это связано с обращением в 0 выражения под логарифмом).

Исследовать ошибку первого рода восстановления структуры разреженности не имеет смысла, так как в марицах нет нулевых элементов (только близкие к нулю). Посмотрим на ошибку второго рода (рис. 12).

Вывод: Таким образом, мы получаем, что и для более реалистичных данных исходные матрицы можно восстановить с помощью алгоритма ALS достаточно точно, поэтому регуляризацию применять нет необходимости. Однако стоит отметить, что в данном случае сходимость к локальным оптимумам является искусственной проблемой, связанной с тем фактом, что число тем изначально известно. В экспериментах на реальных данных такая проблема не возникает, но и устранение этого эффекта в экспериментах на модельных данных является полезным результатом.

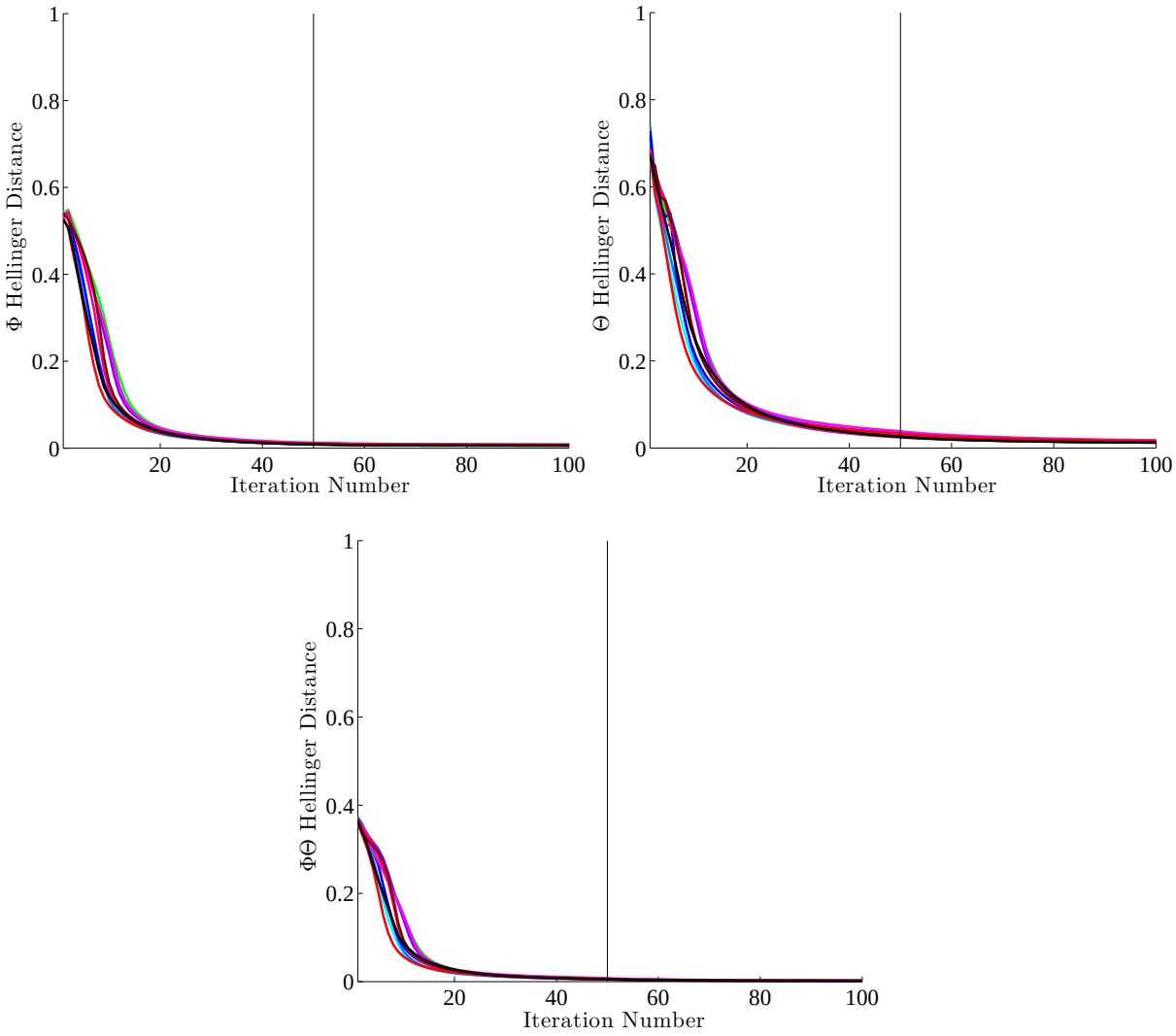

Рис. 10: Сходимость расстояния Хеллингера для матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$ и $\mathbf{\Phi\Theta}$ соответственно (полумодельные данные, 50ALS+EM)

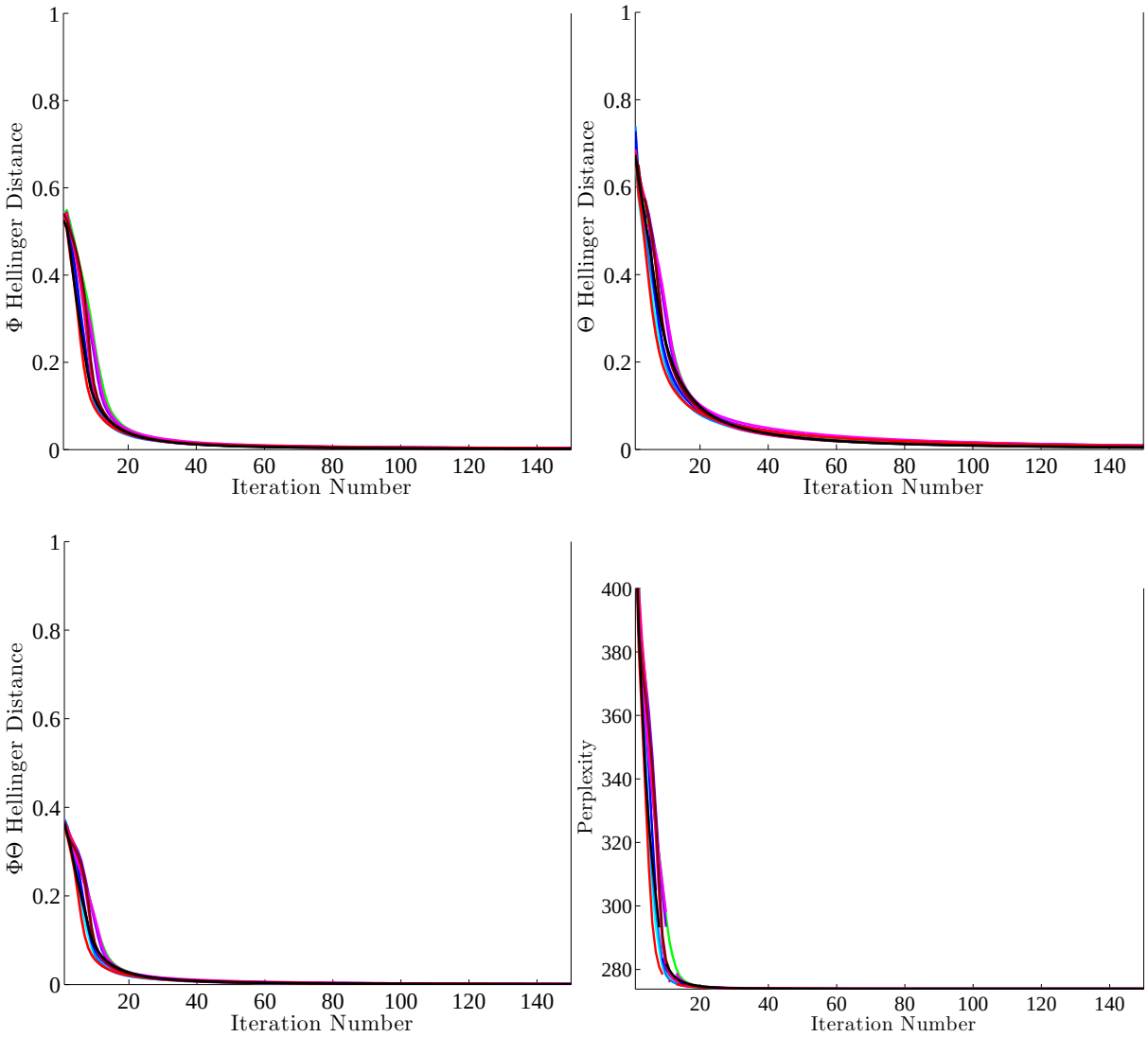

Рис. 11: Сходимость расстояния Хеллингера для матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$ и $\mathbf{\Phi\Theta}$ соответственно (полумодельные данные, 150ALS)

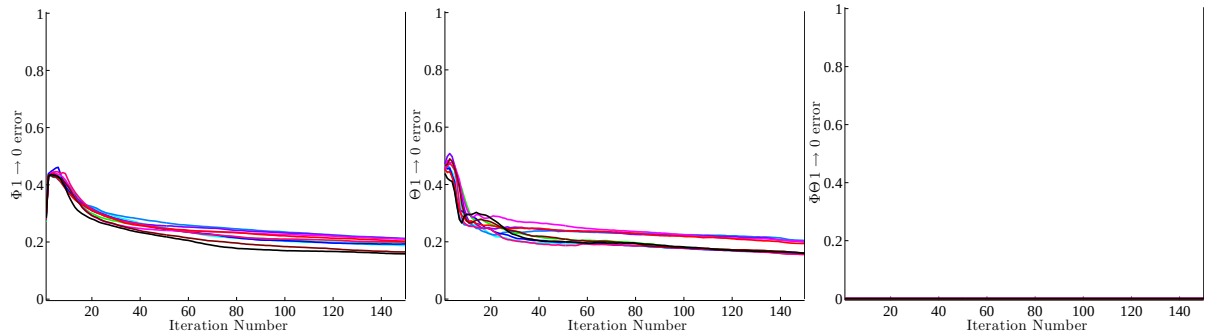

Рис. 12: Доля ошибок первого и второго рода при восстановлении структуры разреженности матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$ и $\mathbf{\Phi\Theta}$ соответственно (полумодельные данные, 150ALS)

4.3 Реалистичные данные

В предыдущих разделах было показано, что в случае существования точного разложения алгоритмы неотрицательного матричного разложения добиваются достаточно точной сходимости. Но для работы с реальными данными требуется получение устойчивого решения и при невыполнении этого условия. В следующих экспериментах попробуем приблизить исследуемые модельные данные к реальным.

Первое отличие от реальных заключается в том, что матрица коллекции, полученная простым перемножением матриц распределений $\mathbf{\Phi}_0 \cdot \mathbf{\Theta}_0 \cdot n_d$, не является целочисленной. Избавимся от этого с помощью вероятностного округления. Тогда повторение экспериментов, показанных на графиках 1 для модельных данных, даёт результат, показанный на рис. 13. Как видно из графиков, картина локальных экстремумов не изменилась. Однако теперь точная сходимость не достигается даже из приближений, которые давали близкий результат в изначальном эксперименте. Таким образом, округление существенно ухудшает восстановление исходных матриц, несмотря на малость самого эффекта: существенную роль вероятностное округление играет только для элементов со значениями между 0 и 1.

Теперь получим целочисленную коллекцию другим способом: будем из описанных выше неразреженных полумодельных матриц $\mathbf{\Theta}_0$ и $\mathbf{\Phi}_0$ сэмплировать для каждого словоместа номер темы и слова соответственно (следуя гипотезе вероятностного порождения коллекции). В этом случае матрица коллекции будет целочисленная и точного разложения не будет. На графиках 14 представлена сходимость для EM (сверху), ALS (посередине) и ALS+EM (снизу) соответственно. Видно, что при рандомизации коллекции ни один из ранее использованных алгоритмов не сходится.

Вывод: Мы получили, что при приближении данных к реалистичным (отсутствии точного разложения) описанные ранее алгоритмы не находят истинное решение. Поэтому перейдём к регуляризованному алгоритму.

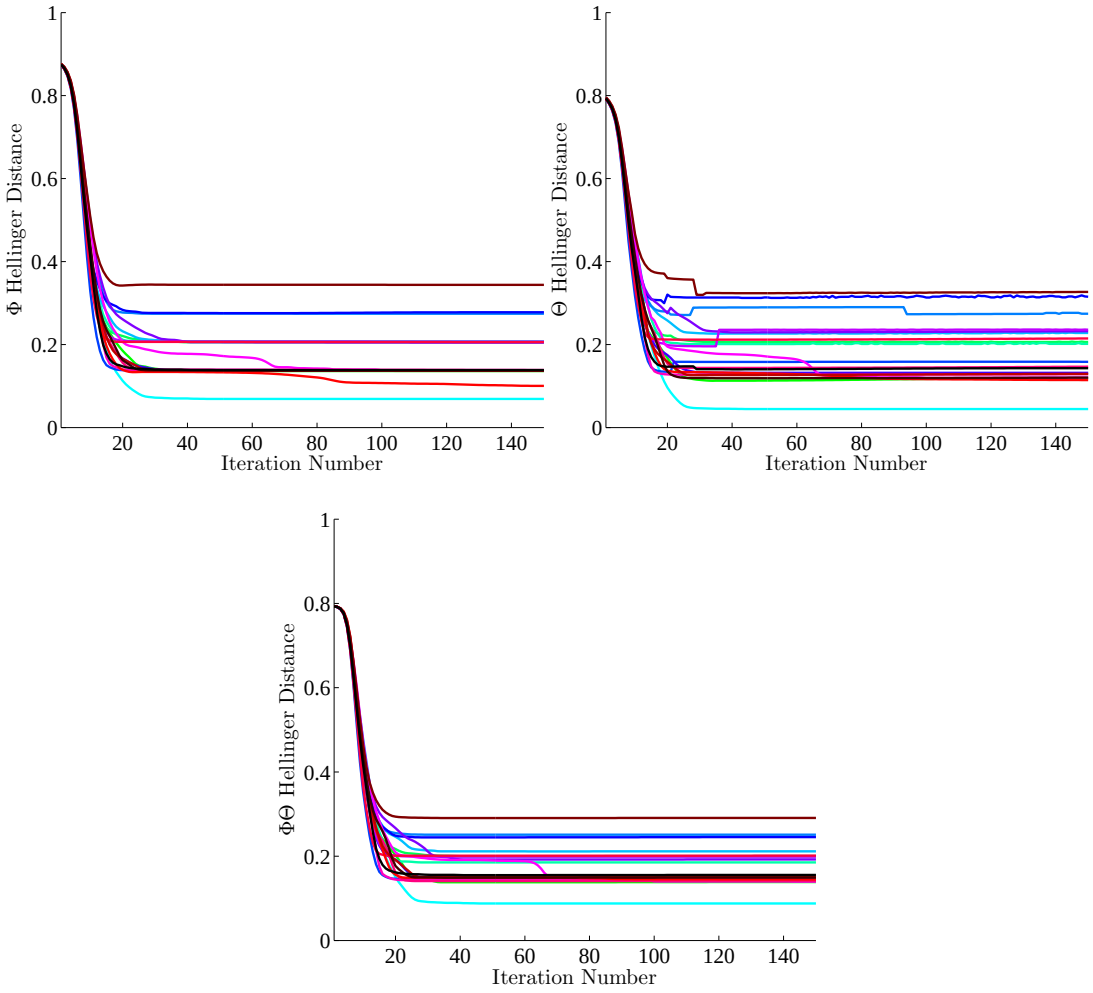

Рис. 13: Сходимость расстояния Хеллингера для матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$, $\mathbf{\Phi\Theta}$ и перплексии модели соответственно (модельные данные с округлением)

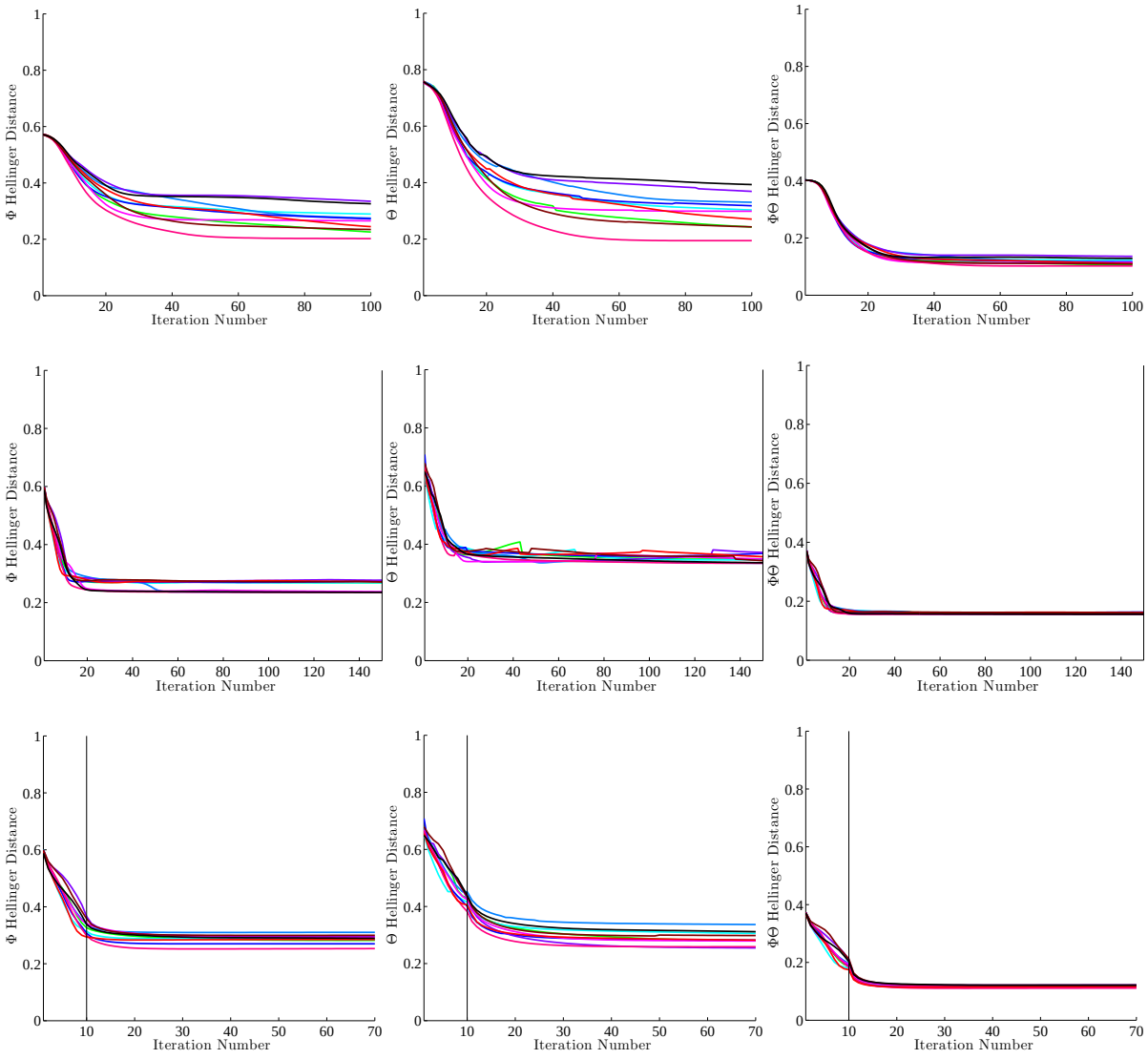

Рис. 14: Сходимость расстояния Хеллингера для матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$, $\mathbf{\Phi\Theta}$ и перплексии модели соответственно (полумодельные данные с рандомизацией, EM, ALS и 10ALS+EM соответственно)

4.4 Регуляризованный алгоритм

Теперь проверим, улучшает ли регуляризация сходимость алгоритма. Данные для этого эксперимента были сгенерированы согласно изначальным предположениям о структуре данных: к описанным разреженным и декоррелированным модельным матрицам были добавлены 2 равномерные фоновые темы. Фоновые строки в матрице $\mathbf{\Theta}$ имеют тот же порядок, что и остальные ненулевые элементы, чтобы зашумление коллекции было ощутимым.

Сгенерируем 20 начальных приближений и на них будем сравнивать сходимость регуляризованного и нерегуляризованного алгоритма. Сначала используем только регуляризатор частичного обучения. D качестве «белых» слов выбираются 50 наиболее вероятных слов в теме (то есть всё тематическое ядро), а в качестве «чёрных» — 50 наименее вероятных. В силу разреженности матрицы $\mathbf{\Phi}_0$ «чёрные» слова в модельных темах имеют нулевые вероятности.

Подбором весов «чёрного» и «белого» регуляризаторов можно получить картину сходимости, показанную на графике 15. Коэффициенты регуляризации $\tau_+ = 10^4$, $\tau_- = 10^6$. Синим показана сходимость нерегуляризованного алгоритма, красным — регуляризованного. Как видно из графика, мы можем сильно выиграть в сходимости матрицы $\mathbf{\Phi}$ (что логично, так как регуляризатор основан на дополнительной информации об этой матрице), но меньше улучшаем сходимость $\mathbf{\Theta}$, и почти не влияем на сходимость их произведения.

Теперь добавим регуляризаторы, учитывающие структуру матриц. После нового подбора весов ($\alpha_0 = 5$, $\beta_0 = 2$, $\tau = 10^4$; регуляризаторы разреживания включаются на 10 итерации, после некоторой стабилизации матриц) из тех же начальных приближений получаем более существенный выигрыш в сходимости (рис. 16). Видно, что регуляризатор разреживания матрицы $\mathbf{\Theta}$ помогает приблизиться к модельной матрице $\mathbf{\Theta}_0$.

Попробуем уменьшить объём экспертной информации. При тех же коэффициентах регуляризации для 25 «чёрных» и «белых» слов получаем меньший выигрыш в сходимости, но всё же заметный (графики 17). Значит, можно при относительно малом количестве информации о темах восстанавливать их существенно лучше.

Следует также отметить, что коэффициенты регуляризации подбирались вруч-

ную по порядку величины. Возможно, при автоматизированной процедуре настройки можно добиться более существенных результатов.

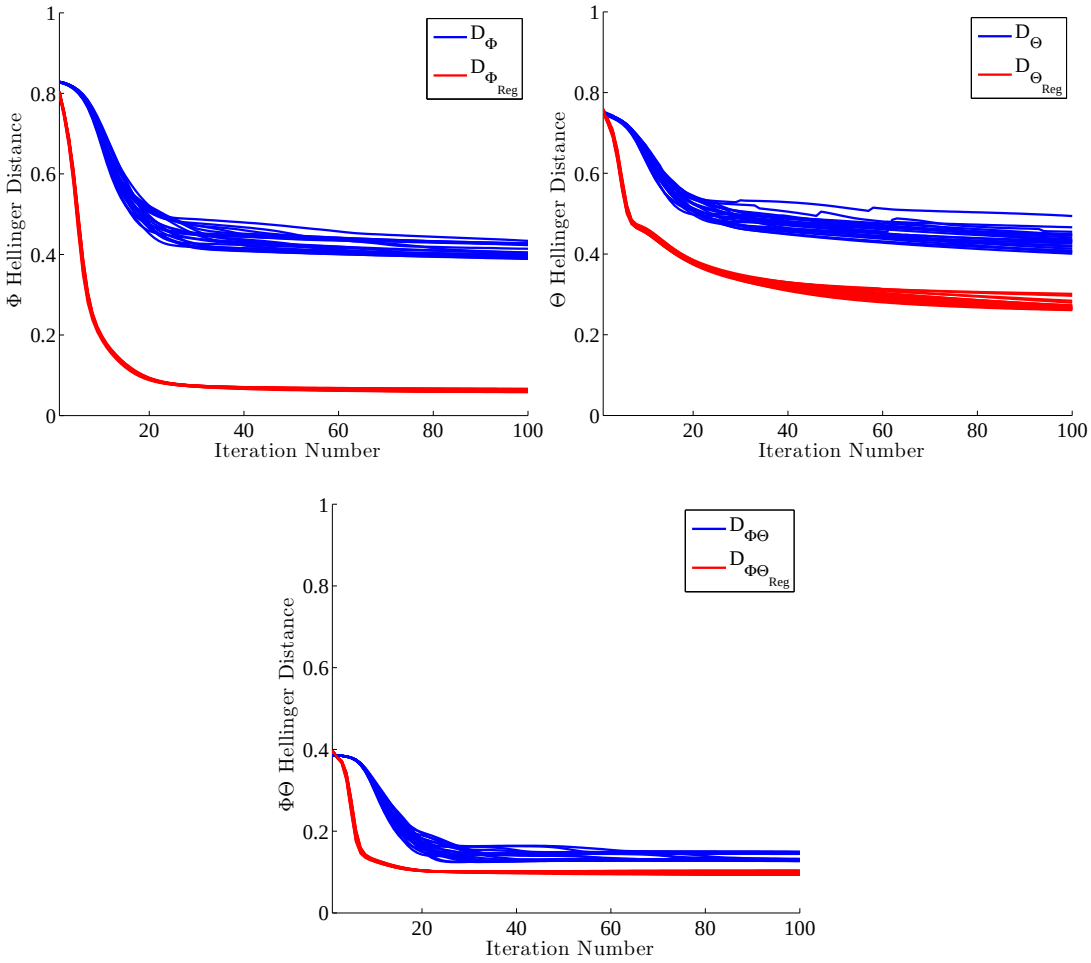

Рис. 15: Сходимость расстояния Хеллингера для матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$, $\mathbf{\Phi\Theta}$ (нерегуляризованный EM-алгоритм и EM с частичным обучением)

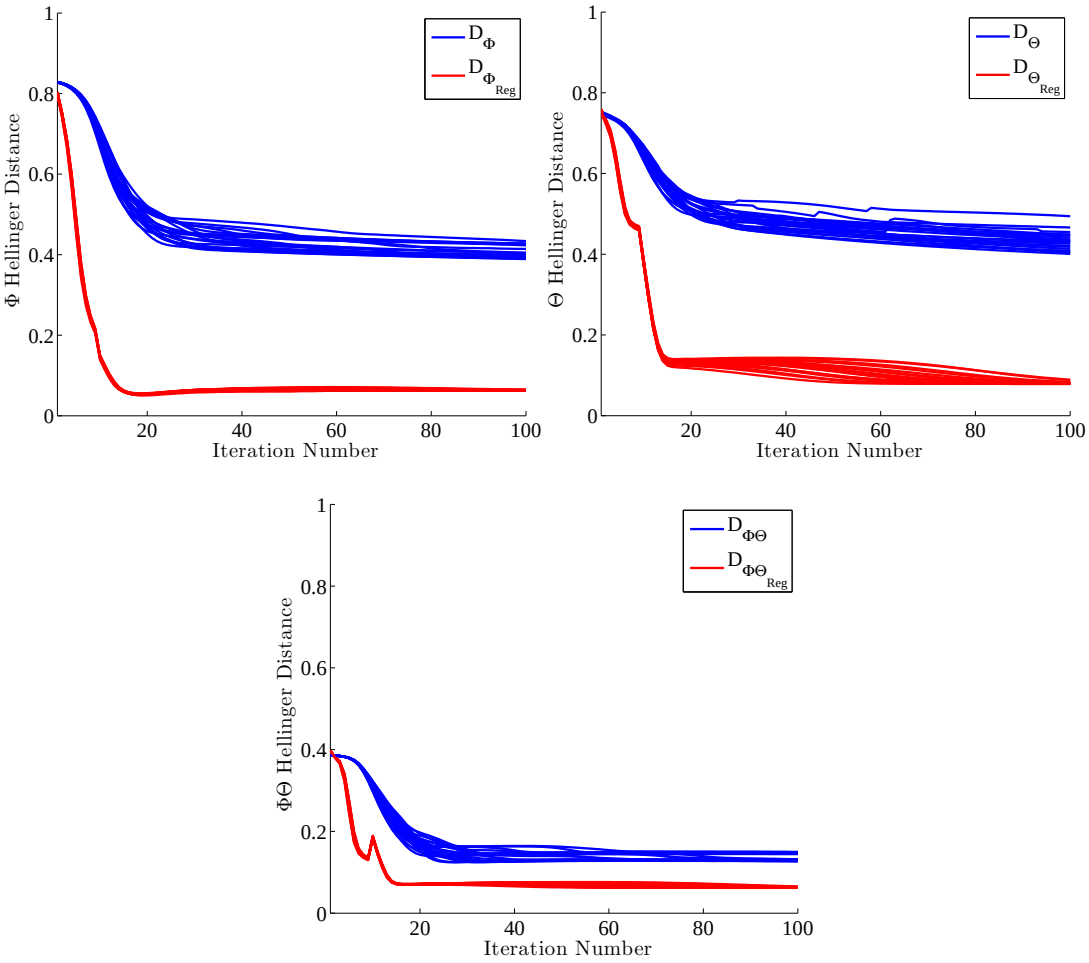

Рис. 16: Сходимость расстояния Хеллингера для матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$, $\mathbf{\Phi\Theta}$ (нерегуляризованный и регуляризованный EM-алгоритм)

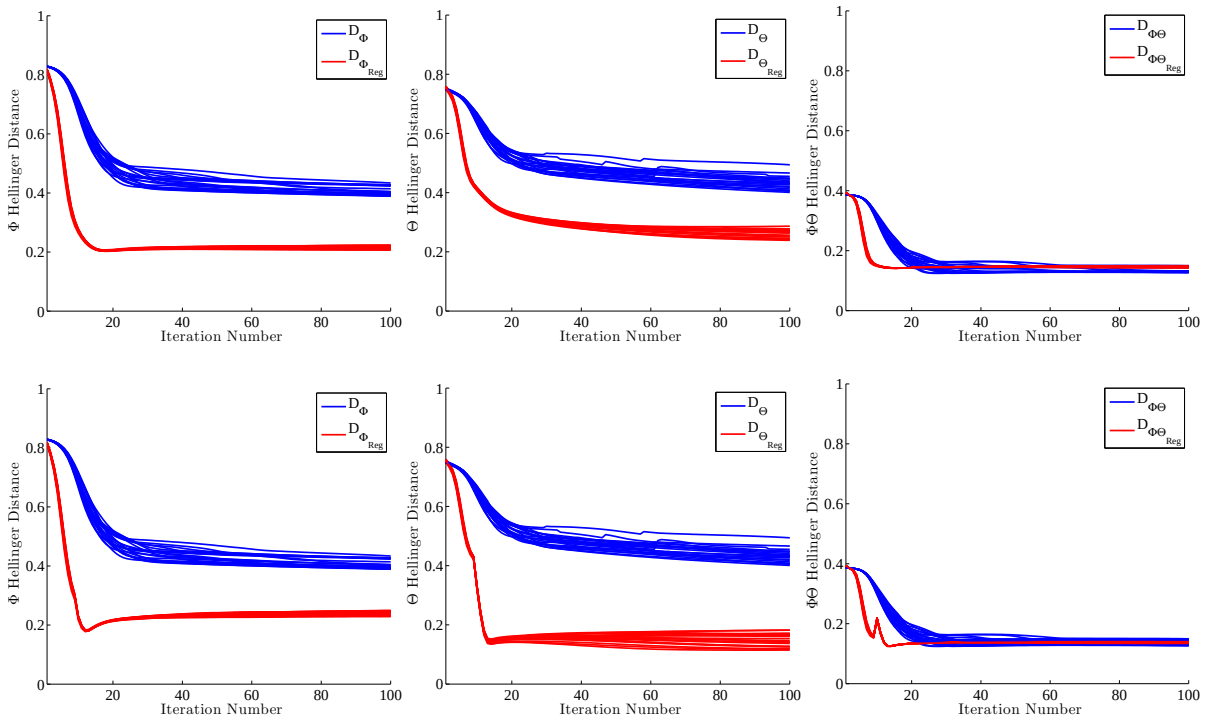

Рис. 17: Сходимость расстояния Хеллингера для матриц $\mathbf{\Phi}$, $\mathbf{\Theta}$, $\mathbf{\Phi\Theta}$ (нерегуляризованный и регуляризованный EM-алгоритм)

5 Заключение

Итак, в работе было формализовано понятие интерпретируемости тем в виде предположений об их структуре. Был предложен набор регуляризаторов тематической модели, позволяющих восстановить интерпретируемую структуру. В серии экспериментов на модельных данных показано, что при чистых модельных данных с существованием точного разложения проблемы несходимости и неуйстойчивости можно решить с помощью алгоритмов матричного разложения. В условиях, максимально приближенных к реальным, точная сходимость не достигается, но преложенные регуляризаторы при верном подборе весов помогают существенно улучшить её.

Список литературы

[1] Xuerui Wang, Andrew McCallum, and Xing Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 697–702. IEEE, 2007.

[2] Levent Bolelli, Şeyda Ertekin, and C Lee Giles. Topic and trend detection in text collections using latent dirichlet allocation. In *Advances in Information Retrieval*, pages 776–780. Springer, 2009.

[3] Jonathan Chang, Jordan L Boyd-Graber, Sean Gerrish, Chong Wang, and David M Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, volume 22, pages 288–296, 2009.

[4] К.В. Воронцов. Вероятностное тематическое моделирование. 2013.

[5] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

[6] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499. ACM, 2007.

[7] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

[8] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality.

[9] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[10] К.В. Воронцов and А.А. Потапенко. Регуляризация, робастность и разреженность вероятностных тематических моделей. *Компьютерные исследования и моделирование*, 4(4):693–706, 2012.

[11] Eric C Chi and Tamara G Kolda. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1272–1299, 2012.

[12] Andrzej Cichocki and PHAN Anh-Huy. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 92(3):708–721, 2009.