Московский государственный университет имени М. В. Ломоносова

Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

Айсина Роза Мунеровна

Тематическое моделирование финансовых потоков корпоративных клиентов банка по транзакционным данным

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:

д.ф-м.н.

Воронцов Константин Вячеславович

Москва, 2017

Содержание

Аннотация

В данной работе ставится задача восстановления латентной информации о видах деятельности компаний по наблюдаемым транзакционным данным. Для выявления латентной кластерной структуры банковских транзакционных данных применяется техника матричных разложений. Но задача матричного разложения является некорректно поставленной, поскольку имеет бесконечно много решений. В работе предлагается регуляризованный алгоритм стохастического разложения, экспериментально исследуются некоторые способы регуляризации и условия устойчивого восстановления исходных матриц по транзакционным данным.

1 Введение

Анализ больших объёмов данных о банковских транзакциях между компаниями (юридическими лицами) открывает перед крупными банками новые возможности по улучшению сервисов и предоставлению новых видов услуг, ранее не свойственных банкам. За отдельными транзакциями могут скрываться общие и отраслевые паттерны финансово-экономической деятельности. Знание этих паттернов может помогать компаниям — клиентам банка — быстрее перенимать успешные для отрасли практики ведения бизнеса, оптимизировать налоги, повышать эффективность управления финансовыми потоками внутри холдингов. Анализ транзакционных данных рассматривается в настоящее время некоторыми крупными банками как важный шаг в сторону таргетирования финансовых услуг и оказания новых услуг в области отраслевого консалтинга. Поэтому становится актуальным созданием инструментов анализа транзакционных данных, способных давать общее понимание структуры финансовых потоков внутри отрасли.

Транзакционные данные могут охватывать сотни тысяч компаний, поэтому важной стадией анализа является агрегирование исходных данных о транзакциях между компаниями и переход к анализу финансовых потоков между видами экономической деятельности. Несмотря на широкое использование общероссийского классификатора видов экономической деятельности (ОКВЭД), его непосредственное использование для анализа транзакционных данных пока сильно затруднено. Обычной практикой является весьма произвольное назначение кодов ОКВЭД как при регистрации компаний, так и при дальнейшем ведении ими экономической деятельности. В банковской практике данные об ОКВЭД конкретных транзакций оказываются ненадёжными или отсутствуют вовсе.

В данной работе ставится задача восстановления латентной информации о видах деятельности компаний по наблюдаемым транзакционным данным. Математическая постановка задачи имеет много общего с задачами кластеризации, вероятностного тематического моделирования и коллаборативной фильтрации.

Вероятностная тематическая модель по наблюдаемой коллекции текстовых документов восстанавливает множество латентных кластеров-тем и описывает каждый документ и каждое слово дискретным распределением над множеством тем. Пары

«слово–документ» являются своего рода транзакциями, а скрытой причиной транзакции является тематическая близость слова и документа.

Латентные модели коллаборативной фильтрации и рекомендательных систем по наблюдаемым данным о пользовательских предпочтениях восстанавливает множество латентных кластеров-интересов. Здесь транзакциями являются пары «пользователь–товар», а скрытой причиной транзакции является способность данного товара удовлетворить интерес (потребность) пользователя.

В этих случаях задачи выявления латентных кластерных структур сводятся к вычислению матричных разложений больших разреженных матриц. В разложении участвуют две матрицы: в случае тематического моделирования это матрицы распределений слов для каждой темы и тем для каждого документа; в случае коллаборативной фильтрации это матрицы распределений товаров для каждого интереса и интересов для каждого пользователя.

Задача матричного разложения является некорректно поставленной, поскольку имеет бесконечно много решений. Для доопределения решения и повышения его устойчивости применяются методы регуляризации [1].

В данной работе техника матричных разложений применяется для выявления латентной кластерной структуры банковских транзакционных данных. В роли кластеров-тем выступают виды экономической деятельности. Главное отличие заключается в том, что разложение состоит не из двух, а из трёх матриц. Соответственно, множество решений становится ещё шире и возрастает роль регуляризации.

В работе предлагается регуляризованный алгоритм стохастического разложения, экспериментально исследуются некоторые способы регуляризации и условия устойчивого восстановления исходных матриц по транзакционным данным.

1.1 Формальная постановка задачи

Пусть F — конечное множество фирм одной отрасли. Транзакционные данные — это множество записей вида (b_i, s_i, m_i), $i = 1, \ldots, n$, где $b_i \in F$ — фирма-покупатель, $s_i \in F$ — фирма-продавец, m_i — объём i-й транзакции. Если последовательность (время) транзакций не важны для целей проводимого анализа, то агрегированные данные

можно представить в виде матрицы объёмов $M = (m_{bs})_{F \times F}$:

$$m_{bs} = \sum_{i=1}^{n} m_i [b_i = b] \, [s_i = s]. \tag{1}$$

Если важны лишь факты транзакций, то данные можно представлять в виде матрицы числа транзакций $N = (n_{bs})_{F \times F}$:

$$n_{bs} = \sum_{i=1}^{n} [b_i = b] \, [s_i = s],$$

или в виде бинарной матрицы индикаторов транзакций $I = \big([n_{bs} > 0]\big)_{F \times F}$.

Пусть V — конечное множество видов экономической деятельности. Каждая транзакция вызвана тем, что для осуществления деятельности $u \in V$ покупателю b требуются некоторые товары или услуги, которые имеются у продавца s в результате его деятельности $v \in V$. Предполагается, что число видов деятельности много меньше числа фирм, $|V| \ll |F|$.

Целью данной работы является разработка модели и метода оптимизации ее параметров для восстановления множества видов деятельности и связей между ними внутри отрасли, которые описываются графом. Вершинами графа являются виды экономической деятельности, а рёбра соответствуют характерным для данной отрасли типам сделок между фирмами. Для построения этого графа ставится задача выявления латентного множества видов деятельности (вершин графа) и связей между ними (рёбер графа) по наблюдаемым транзакционным данным.

1.2 Обзор различных методов построения взаимодействия между фирмами

Авторы [2] сравнили структуру связей между фирмами в двух секторах промышленности. Для этого была построена так называемая сеть транзакций (*transaction network*) — граф, вершинами которого являются фирмы, а ребрами — транзакции между данным фирмами (рис. 1). Авторы построили сети для автомобильной и электронной промышленности в Японии. Одной из целей сравнения было проверить гипотезу о том, что иерархия является общим свойством таких производственных сетей. Эмпирические результаты опровергли эту гипотезу и показали, что сектор электро-

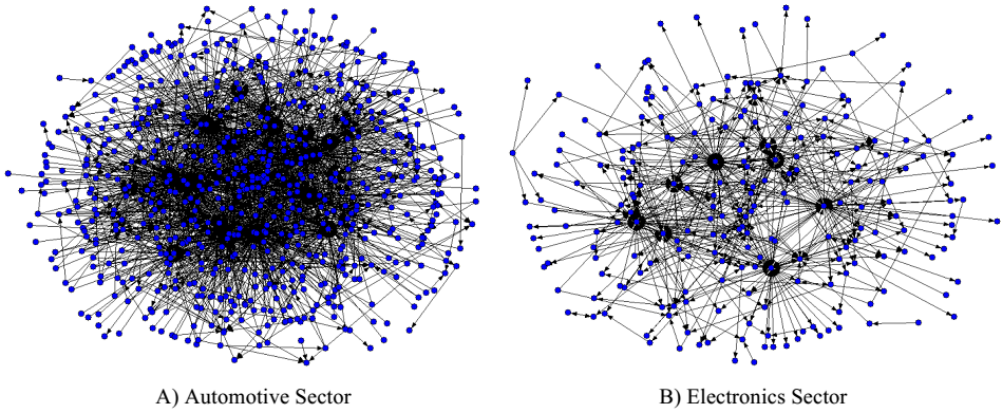

Рис. 1: Сети транзакций между фирмами для автомобильной и электронной промышленности Японии в 1993 году [2].

ники демонстрирует значительно более низкую степень иерархии, чем автомобильный сектор, из-за наличия многочисленных межфирменных циклов в сети.

Наличие циклов, т.е. замкнутых ориентированных путей в графе, позволяет выявлять фирмы, которые занимаются разными направлениями. Было обнаружено, что одна из 10 крупнейших компаний входит в большинство циклов в сети электроники. Также авторы выяснили, что крупные ИТ-компании, которые входят в большое количество циклов, поддерживают широкий спектр возможностей по созданию комплексных решений для своих клиентов.

Авторы [3] объясняют, как возникают транзакции. Они используют несколько другие, более широкие, понятия: трансфер (*transfer*) — некоторое взаимодействие между фирмами или внутри фирмы (например, передача товара или информации) и транзакция (*transaction*) — сделка между фирмами, то есть трансфер, обладающий следующими свойствами: стандартизованность (*standardized*, обе стороны обладают соответствующей информацией о сделке), с определенным количеством товара (*counted*, обе стороны знают, какое количество товара передается) скомпенсированность (*compensated*, покупатель каким-либо образом компенсирует передачу товара продавцом, например, платит за него). Важно, что не всякий трансфер есть транзакция, например, взаимодействия внутри фирмы могут не быть транзакциями. Для визуализации авторы используют матрицу смежности (рис. 2), где отображены как внутренние, так и внешние трансферы фирм.

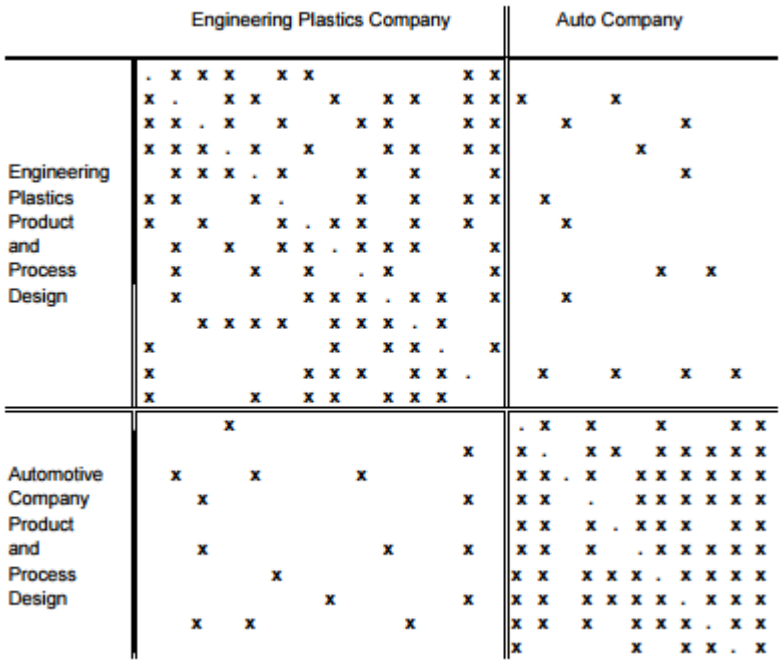

Рис. 2: Матрица смежности внутренних и внешних трасферов фирм [3].

Авторы описывают задачу образования транзакции из трансфера с минимальными издержками. Вводится понятие модулярности (*modularity*), или блочности, когда фирма сама образует блок (например, холдинг из нескольких фирм) или несколько фирм образуют блок путем тесного взаимодействия. Для минимизации издержек необходимо строить области, в которых может появляться определенное множество трансферов (*transaction-free zones*, такое название означает, что по определению такие трансферы не смогут стать транзакциями), и локально замкнутые области (*encapsulated local system*), для которых движение товара вовне и внутрь области осуществляется только транзакциями. Тогда новые транзакции образуются на границах блоков. Таким образом, одной из подзадач задачи минимизации издержек является построение блоков внутреннего и внешнего взаимодействия фирм.

В работе [4] показано, как динамический граф отрасли облегчает ее анализ на трех различных уровнях. Во-первых, он показывает, как развивается рынок отрасли. Во-вторых, он позволяет выделять банки, паттерн изменения которых представляет экономический интерес. И, наконец, это облегчает сравнение одного и того же рынка

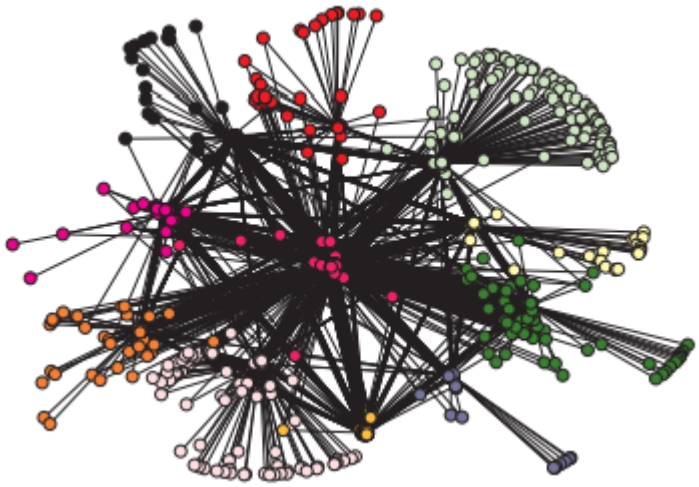

Рис. 3: Граф транзакций банков Австрии, кластеры покрашены одним цветом [3].

в разные моменты времени и различных рынков (например, страны) в один момент времени.

В [5] для построения кластеров фирм разработали адаптацию алгоритма K-Means для транзакционных данных. Используя расстояние Джаккарда, авторы сравнили K-Means с другими методами. Авторы [6] также применили K-Means и K-Means++ по данным географического расположения фирм. Далее, используя специальные метрики, они выявили «локальные» кластеры фирм для последующей экономической оптимизации.

Авторы [7] строят кластеры банков в Австрии в ориентированном взвешенном графе, построенном по транзакцонным данным. Формирование кластера осуществляется на основе метрики между банками. Граф транзакций показан на рис. 3, вершины, образующие кластер, покрашены в один цвет.

2 EM-алгоритм трехматричного низкорангового неотрицательного разложения

Будем описывать граф в виде матрицы смежности Λ размера $|V| \times |V|$. Граф ориентированный, двудольный. Вершинами первого типа являются виды экономической деятельности $u \in V$ продавцов, а вершинами второго типа — виды экономической деятельности $v \in V$ продавцов. Ребрами — взаимодействия между фирмами, кото-

рые занимаются данными видами деятельности. Наибольший интерес представляет квадратная матрица Λ, элементы которой λ_{uv} интерпретируются как вероятности транзакции между покупателем, осуществляющим деятельность u и продавцом, осуществляющим деятельность v. Она является матрицей смежности графа финансовых потоков между видами деятельности. Финансовые потоки движутся от покупателей к продавцам (в противоположном направлении, от продавцов к покупателям, движутся товарные потоки).

Для решения поставленной задачи предполагается применить технику аддитивной регуляризации тематических моделей, которая ранее применялась в основном для анализа текстовых коллекций.

Введём дискретное вероятностное пространство $\Omega = F \times V \times F \times V$ с неизвестным распределением $p(b, u, s, v)$, где $b \in F$ — фирма-покупатель, $u \in V$ — вид деятельности покупателя, $s \in F$ — фирма-продавец, $v \in V$ — вид деятельности продавца. Предполагается, что из этого распределения случайно и независимо порождаются транзакции $p(b_i, u_i, s_i, v_i)$, причём переменные b_i и s_i являются наблюдаемыми, а переменные u_i и v_i — скрытыми.

Альтернативная интерпретация состоит в том, чтобы множество транзакций считать неслучайным, а под вероятностями понимать частоты

$$p(b, u, s, v) = \frac{1}{n} \sum_{i=1}^{n} [b_i = b]\,[u_i = u]\,[s_i = s]\,[v_i = v].$$

Все дальнейшие выкладки справедливы для обеих интерпретаций.

Рассмотрим следующую вероятностную модель порождения данных. С вероятностью $\lambda_{uv} = p(u, v)$ в отрасли возникает потребность осуществить деятельность u, но для этого сначала необходимо осуществить деятельность v. С вероятностью $p_{bu} = p(b \,|\, u)$ для осуществления деятельности u выступает фирма b. С вероятностью $q_{sv} = p(s \,|\, v)$ для осуществления деятельности v выступает фирма s. В результате происходит транзакция (b, s), то есть b покупает у s необходимый товар или услугу.

Введём гипотезу условной независимости $p(b, s \,|\, u, v) = p(b \,|\, u)p(s \,|\, v)$, означающую, что выбор покупателя b зависит только от его деятельности u и не зависит от продавца, а выбор продавца s зависит только от его деятельности v и не зависит от покупателя. Заметим, что эти предположения могут не выполняться из-за регио-

нальных особенностей размещения фирм b и s, но мы пренебрегаем этими эффектами ради упрощения модели.

При сделанных предположениях вероятность транзакции между покупателем $b \in F$ и продавцом $s \in F$ описывается формулой полной вероятности:

$$p(b,s) = \sum_{u \in V} \sum_{v \in V} p(b \mid u) p(s \mid v) p(u,v) = \sum_{u \in V} \sum_{v \in V} p_{bu} q_{sv} \lambda_{uv}.$$

Параметрами модели являются три матрицы: $P = (p_{bu})_{F \times V}$, $Q = (q_{sv})_{F \times V}$, $\Lambda = (\lambda_{uv})_{V \times V}$. Для их определения необходимо решить задачу приближения заданной матрицы N низкоранговым неотрицательным матричным разложением:

$$N \approx nP\Lambda Q^{\intercal},$$

где слева находится известная матрица числа транзакций, справа — три неизвестные матрицы. Для получения матричного разложения воспользуемся принципом максимума правдоподобия при ограничениях нормировки и неотрицательности:

$$\sum_{b \in F} \sum_{s \in F} n_{bs} \log p(b,s) = \sum_{b \in F} \sum_{s \in F} n_{bs} \log \sum_{u \in V} \sum_{v \in V} p_{bu} q_{sv} \lambda_{uv} \to \max_{P,Q,\Lambda}, \tag{2}$$

$$\sum_{b \in F} p_{bu} = 1,\ p_{bu} \geqslant 0, \quad \sum_{s \in F} q_{sv} = 1,\ q_{sv} \geqslant 0, \quad \sum_{u \in V} \sum_{v \in V} \lambda_{uv} = 1,\ \lambda_{uv} \geqslant 0.$$

Заметим, что для каждой фирмы-покупателя $b \in F$ распределение видов деятельности $p(u \mid b)$ может быть получено по теореме Байеса:

$$p(u \mid b) = \frac{p(b \mid u) p(u)}{p(b)}.$$

Аналогично получаем распределение видов деятельности $p(v \mid s)$ для фирмы-продавца $s \in F$:

$$p(v \mid s) = \frac{p(s \mid v) p(v)}{p(s)}.$$

Задача матричного разложения некорректно поставлена и имеет в общем случае бесконечное множество решений. Для получения приемлемого решения введём регуляризатор R, формализующий дополнительные требования к модели:

$$\sum_{b \in F} \sum_{s \in F} n_{bs} \log p(b,s) + R(P,Q,\Lambda) \to \max_{P,Q,\Lambda}.$$

Если ограничений много, то вводится взвешенная сумма k регуляризаторов с коэффициентами τ_j:

$$R(P, Q, \Lambda) = \sum_{j=1}^{k} \tau_j R_j(P, Q, \Lambda).$$

Тогда с учетом регуляризаторов перепишем оптимизационную задачу (2) в следующем виде:

$$\begin{cases} \sum\limits_{b\in F}\sum\limits_{s\in F} n_{bs} \ln \sum\limits_{u\in V}\sum\limits_{v\in V} p_{bu}q_{sv}\lambda_{uv} + R(P, Q, \Lambda) \to \max\limits_{P,Q,\Lambda}, \\ \sum\limits_{b\in F} p_{bu} = 1,\ p_{bu} \geqslant 0, \\ \sum\limits_{s\in F} q_{sv} = 1,\ q_{sv} \geqslant 0, \\ \sum\limits_{u\in V}\sum\limits_{v\in V} \lambda_{uv} = 1,\ \lambda_{uv} \geqslant 0. \end{cases} \tag{3}$$

Будем называть вид деятельности $u \in V$ *регулярным*, если $n_{bu} + p_{bu}\frac{\partial R}{\partial p_{bu}} > 0$ для хотя бы одной фирмы-покупателя $b \in F$, иначе будем говорить, что вид деятельности u *перерегуляризирован*. Аналогично будем называть вид деятельности $v \in V$ *регулярным*, если $n_{sv} + q_{sv}\frac{\partial R}{\partial q_{sv}} > 0$ для хотя бы одной фирмы-продавца $s \in F$, иначе будем говорить, что вид деятельности v *перерегуляризирован*.

Теорема 1. *Если $P,\ Q,\ \Lambda$ — решение задачи максимизации регуляризованного правдоподобия (3), функция $R(P, Q, \Lambda)$ непрерывно дифференцируема, то при условиях регулярности она удовлетворяет системе уравнений со вспомогательными переменными $p_{uvbs} = p(u, v \mid b, s)$:*

E-шаг :

$$p_{uvbs} = \underset{(u,v)\in V^2}{\operatorname{norm}} p_{bu}q_{sv}\lambda_{uv};$$

M-шаг :

$$p_{bu} = \underset{b\in F}{\operatorname{norm}}\left(n_{bu} + p_{bu}\frac{\partial R}{\partial p_{bu}}\right); \qquad n_{bu} = \sum_{s\in F}\sum_{v\in V} n_{bs}p_{uvbs};$$

$$q_{sv} = \underset{s\in F}{\operatorname{norm}}\left(n_{sv} + q_{sv}\frac{\partial R}{\partial q_{sv}}\right); \qquad n_{sv} = \sum_{b\in F}\sum_{u\in V} n_{bs}p_{uvbs};$$

$$\lambda_{uv} = \underset{(u,v)\in V^2}{\operatorname{norm}}\left(n_{uv} + \lambda_{uv}\frac{\partial R}{\partial \lambda_{uv}}\right); \qquad n_{uv} = \sum_{b\in F}\sum_{s\in F} n_{bs}p_{uvbs},$$

где $\underset{t\in T}{\operatorname{norm}}(x_t) = \dfrac{\max\{x_t, 0\}}{\sum\limits_{s\in T}\max\{x_s, 0\}}$ *— операция нормирования вектора.*

Доказательство. Запишем Лагранжиан оптимизационной задачи (3):

$$\mathscr{L} = \sum_{b\in F}\sum_{s\in F} n_{bs}\log\sum_{u\in V}\sum_{v\in V} p_{bu}\lambda_{uv}q_{vs} + R(P,Q,\Lambda) - \sum_{u\in V}\theta_u\left(\sum_{b\in F} p_{bu} - 1\right) -$$
$$- \sum_{v\in V}\beta_v\left(\sum_{s\in F} q_{sv} - 1\right) - \alpha\left(\sum_{u\in V}\sum_{v\in V}\lambda_{uv} - 1\right) + \sum_{b\in F}\sum_{u\in V}\theta_{bu}\cdot p_{bu} + \sum_{s\in F}\sum_{v\in V}\beta_{sv}\cdot q_{sv} +$$
$$+ \sum_{u\in V}\sum_{v\in V}\alpha_{uv}\cdot\lambda_{uv}. \quad (4)$$

Доказательство для p_{bu}:

Условия Каруша-Куна-Таккера для p_{bu}:

$$\begin{cases} \theta_u - \theta_{ub} = \sum\limits_{s\in F} n_{bs}\dfrac{\sum\limits_{v\in V}\lambda_{uv}q_{vs}}{\sum\limits_{u\in V}\sum\limits_{v\in V} p_{bu}\lambda_{uv}q_{vs}} + \dfrac{\partial R}{\partial p_{bu}}, \\ \theta_{bu} \geqslant 0, \text{ двойственные ограничения}, \\ \theta_{bu}p_{bu} = 0, \text{ условие дополняющей нежёсткости}. \end{cases} \quad (5)$$

Домножим обе части первого равенства в (5) на p_{bu}:

$$p_{bu}\theta_u - p_{bu}\theta_{ub} = p_{bu}\sum_{s\in F} n_{bs}\frac{\sum\limits_{v\in V}\lambda_{uv}q_{vs}}{p(b,s)} + p_{bu}\frac{\partial R}{\partial p_{bu}}.$$

Учитывая условия дополняющей нежёсткости в (5), получим:

$$p_{bu}\theta_u = \sum_{s\in F} n_{bs}\sum_{v\in V}\underbrace{\frac{p_{bu}\lambda_{uv}q_{vs}}{p(b,s)}}_{p_{ubvs}} + p_{bu}\frac{\partial R}{\partial p_{bu}} = n_{bu} + p_{bu}\frac{\partial R}{\partial p_{bu}}.$$

Условие $\theta_u \leqslant 0$, противоречит условию регулярности u. Это означает, что $p_{bu} \equiv 0 \ \ \forall b$. Следовательно, $\theta_u > 0$, $p_{bu} \geqslant 0$. Левая часть уравнения неотрицательная, значит, правая часть также неотрицательна, и

$$p_{bu}\theta_u = \left(n_{bu} + p_{bu}\frac{\partial R}{\partial p_{bu}}\right)_+. \quad (6)$$

Просуммируем обе части уравнения по всем $b \in F$:

$$\theta_u = \sum_{b\in F}\left(n_{bu} + p_{bu}\frac{\partial R}{\partial p_{bu}}\right)_+. \quad (7)$$

Подставив в (6) выражение (7) для θ_u, получим:

$$p_{bu} = \frac{\left(n_{bu} + p_{bu}\dfrac{\partial R}{\partial p_{bu}}\right)_+}{\sum\limits_{b' \in F}\left(n_{b'u} + p_{b'u}\dfrac{\partial R}{\partial p_{b'u}}\right)_+}.$$

Доказательство для q_{sv}: проводится аналогично.

Доказательство для λ_{uv}:

Продифференцируем лагранжиан (4) по λ_{uv} и, приравняв нулю производную, получим условия Каруша-Куна-Таккера для λ_{uv}:

$$\begin{cases} \alpha - \alpha_{uv} = \sum\limits_{b \in F}\sum\limits_{s \in F} n_{bs}\dfrac{p_{bu}q_{vs}}{\sum\limits_{u \in V}\sum\limits_{v \in V} p_{bu}\lambda_{uv}q_{vs}} + \dfrac{\partial R}{\partial \lambda_{uv}}, \\ \alpha_{uv} \geqslant 0, \text{ двойственные ограничения,} \\ \alpha_{uv}\lambda_{uv} = 0, \text{ условие дополняющей нежёсткости.} \end{cases} \tag{8}$$

Домножим обе части первого равенства в (8) на λ_{uv}, и, учитывая условие дополняющей нежёсткости, получим:

$$\lambda_{uv}\alpha = \sum_{b \in F}\sum_{s \in F} n_{bs}\underbrace{\frac{p_{bu}\lambda_{uv}q_{vs}}{p(b,s)}}_{p_{ubvs}} + \lambda_{uv}\frac{\partial R}{\partial \lambda_{uv}} = n_{uv} + \lambda_{uv}\frac{\partial R}{\partial \lambda_{uv}}.$$

Условие $\alpha \leqslant 0$, противоречит условию регулярности. Это означает, что $\lambda_{uv} \equiv 0 \ \ \forall u,\ v$. Следовательно, $\alpha > 0$, $\lambda_{uv} \geqslant 0$. Левая часть уравнения неотрицательная, значит, правая часть также неотрицательна, и

$$\lambda_{uv}\alpha = \left(n_{uv} + \lambda_{uv}\frac{\partial R}{\partial \lambda_{uv}}\right)_+. \tag{9}$$

Просуммируем обе части по всем $u, v \in V$:

$$\alpha = \sum_{u \in V}\sum_{v \in V}\left(n_{uv} + \lambda_{uv}\frac{\partial R}{\partial \lambda_{uv}}\right)_+. \tag{10}$$

Подставив выражение (10) для α в (9), получим:

$$\lambda_{uv} = \frac{\left(n_{uv} + \lambda_{uv}\dfrac{\partial R}{\partial \lambda_{uv}}\right)_+}{\sum\limits_{u' \in V}\sum\limits_{v' \in V}\left(n_{u'v'} + \lambda_{u'v'}\dfrac{\partial R}{\partial \lambda_{u'v'}}\right)_+}.$$

Теорема доказана. ■

Рассмотрим формулу полной вероятности, описывающую вероятность транзакции (b, s):

$$p(b, s) = \sum_{u \in V} p(b \mid u) \sum_{v \in V} p(s \mid v) p(u, v).$$

Видно, что если $\sum\limits_{v \in V} p(s \mid v) p(u, v) \equiv 0 \ \ \forall s$ и при этом $p(b \mid u) \neq 0$ для хотя бы одного u, то вероятность транзакции $p(b, s)$ все равно будет равна нулю. Таким образом, может случиться так, что вероятности в P будут ненулевыми для некоторой фирмы b, а транзакций с этой фирмой не могут возникнуть. Аналогичная ситуация возникает, если $\sum\limits_{u \in V} p(b \mid u) p(u, v) \equiv 0 \ \ \forall b$ и $p(s \mid v) \neq 0$, но транзакций с фирмой s не может возникнуть. Если в матрице Λ вид деятельности u ни с кем не взаимодействует, т.е. $\sum\limits_{v \in V} \lambda_{uv} \equiv 0$, то надо исключить этот вид деятельности из матрицы P; аналогично, если $\sum\limits_{u \in V} \lambda_{uv} \equiv 0$, то v надо исключить из Q.

Таким образом, потребуем непротиворечивости в данных транзакций:

- Если $\sum_v p(s \mid v) p(u, v) \equiv 0 \ \ \forall s$, то $p(b \mid u) := 0$;
- Если $\sum_u p(b \mid u) p(u, v) \equiv 0 \ \ \forall b$, то $p(s \mid v) := 0$;
- Если $\sum_v \lambda_{uv} \equiv 0$, то $p(b \mid u) := 0 \ \ \forall b$;
- Если $\sum_u \lambda_{uv} \equiv 0$, то $p(s \mid v) := 0 \ \ \forall s$.

3 EM-алгоритм двухматричного низкорангового неотрицательного разложения

Упростим предыдущую модель, положив $Q \equiv P$. В реальных данных это соответствует тому, что мы сообщаем модели, что каждый вид экономической деятельности должен объснять как покупки, так и продажи фирмы, не разделяя деятельность фирмы как подавца и как покупателя.

В этой модели матрица смежности Λ размера $|V| \times |V|$ будет описывать ориентированный граф, вершинами которого являются виды экономической деятельности $u \in V$ фирм, а ребрами — взаимодействия между фирмами, которые занимаются данными видами деятельности. Вершины этого графа легче интерпретировать, так как нет разрыва между покупательской и продавательской деятельностью фирм.

Рассмотрим следующую вероятностную модель порождения данных. С вероятностью $\lambda_{uv} = p(u, v)$ в отрасли возникает потребность осуществить деятельность u, но для этого сначала необходимо осуществить деятельность v. С вероятностью $p_{bu} = p(b \mid u)$ для осуществления деятельности u выступает фирма b. С вероятностью $p_{sv} = p(s \mid v)$ для осуществления деятельности v выступает фирма s. В результате происходит транзакция (b, s), то есть b покупает у s необходимый товар или услугу.

Как и в предыдущей модели, введём гипотезу условной независимости $p(b, s \mid u, v) = p(b \mid u)p(s \mid v)$. При сделанных предположениях вероятность транзакции между покупателем $b \in F$ и продавцом $s \in F$ описывается формулой полной вероятности:

$$p(b, s) = \sum_{u \in V} \sum_{v \in V} p(b \mid u)p(s \mid v)p(u, v) = \sum_{u \in V} \sum_{v \in V} p_{bu} p_{sv} \lambda_{uv}.$$

Параметрами модели являются две матрицы: $P = (p_{ft})_{F \times V}$, $\Lambda = (\lambda_{uv})_{V \times V}$. Теперь необходимо решить задачу приближения заданной матрицы N низкоранговым неотрицательным матричным разложением:

$$N \approx nP\Lambda P^{\mathrm{т}},$$

где слева находится известная матрица числа транзакций, справа — две неизвестные матрицы.

Как и в предыдущей модели, для получения матричного разложения воспользуемся принципом максимума правдоподобия с регуляризатором при ограничениях нормировки и неотрицательности:

$$\begin{cases} \sum\limits_{b \in F} \sum\limits_{s \in F} n_{bs} \ln \sum\limits_{u \in V} \sum\limits_{v \in V} p_{bu} \lambda_{uv} p_{sv} + R(P, \Lambda) \to \max\limits_{P, \Lambda}, \\ \sum\limits_{b \in F} p_{bu} = 1, \;\; p_{bu} \geqslant 0, \\ \sum\limits_{u \in V} \sum\limits_{v \in V} \lambda_{uv} = 1, \;\; \lambda_{uv} \geqslant 0. \end{cases} \tag{11}$$

Будем называть вид деятельности $t \in V$ *регулярным*, если $n_{ft} + n'_{ft} + p_{tf} \dfrac{\partial R}{\partial p_{tf}} > 0$ для хотя бы одной фирмы $f \in F$, иначе будем говорить, что вид деятельности t *перерегуляризирован*.

Теорема 2. *Если $P,\ \Lambda$ — решение задачи максимизации регуляризованного правдоподобия* (11)*, функция $R(P,\Lambda)$ непрерывно дифференцируема, то при условиях регулярности она удовлетворяет системе уравнений со вспомогательными переменными $p_{uvbs} = p(u,v \mid b,s)$:*

E-шаг :

$$p_{uvbs} = \underset{(u,v)\in V^2}{\operatorname{norm}} p_{bu} q_{sv} \lambda_{uv};$$

M-шаг :

$$p_{ft} = \underset{b\in F}{\operatorname{norm}}\left(n_{ft} + n'_{ft} + p_{ft}\frac{\partial R}{\partial p_{ft}}\right), \qquad n_{ft} = \sum_{s\in F}\sum_{t\in V} n_{fs} p_{tvfs},$$

$$n'_{ft} = \sum_{b\in F}\sum_{u\in V} n_{bf} p_{utbf};$$

$$\lambda_{uv} = \underset{(u,v)\in V^2}{\operatorname{norm}}\left(n_{uv} + \lambda_{uv}\frac{\partial R}{\partial \lambda_{uv}}\right), \qquad n_{uv} = \sum_{b\in F}\sum_{s\in F} n_{bs} p_{uvbs},$$

где $\underset{t\in T}{\operatorname{norm}}(x_t) = \dfrac{\max\{x_t, 0\}}{\sum\limits_{s\in T}\max\{x_s, 0\}}$ *— операция нормирования вектора.*

Доказательство. Запишем Лагранжиан оптимизационной задачи (11):

$$\mathscr{L} = \sum_{b\in F}\sum_{s\in F} n_{bs} \log \sum_{u\in V}\sum_{v\in V} p_{bu}\lambda_{uv}p_{vs} + R(P,\Lambda) - \sum_{u\in V}\theta_u\left(\sum_{b\in F} p_{bu} - 1\right) -$$
$$- \alpha\left(\sum_{u\in V}\sum_{v\in V}\lambda_{uv} - 1\right) + \sum_{b\in F}\sum_{u\in V}\theta_{bu}p_{bu} + \sum_{u\in V}\sum_{v\in V}\alpha_{uv}\lambda_{uv}. \quad (12)$$

Доказательство для p_{ft}:

Условия Каруша-Куна-Таккера для p_{ft}

$$\begin{cases} \theta_t - \theta_{tf} = \underbrace{\sum\limits_{s\in F} n_{fs}\dfrac{\sum\limits_{v\in V}\lambda_{tv}p_{vs}}{\sum\limits_{u\in V}\sum\limits_{v\in V} p_{fu}\lambda_{uv}p_{vs}}}_{(f,t)\neq(s,v)} + \underbrace{\sum\limits_{b\in F} n_{bf}\dfrac{\sum\limits_{u\in V} p_{bu}\lambda_{ut}}{\sum\limits_{u\in V}\sum\limits_{v\in V} p_{bu}\lambda_{uv}p_{vf}}}_{(b,u)\neq(f,t)} + \underbrace{n_{ff}\dfrac{2\cdot\lambda_{tt}p_{tf}}{\sum\limits_{u\in V}\sum\limits_{v\in V} p_{fu}\lambda_{uv}p_{vfx}}}_{(b,u)=(s,v)=(f,t)} + \dfrac{\partial R}{\partial p_{ft}} \\ \theta_{ft} \geqslant 0, \text{ двойственные ограничения,} \\ \theta_{ft}\cdot p_{ft} = 0, \text{ условие дополняющей нежёсткости.} \end{cases} \quad (13)$$

Заметим, что третье слагаемое в (13) равно 1, когда транзакция происходит внутри фирмы. Такие случаи не рассматриваются в модели, поэтому третье слагаемое можно занулить.

Домножим обе части первого равенства в (13) на p_{ft}, применим условие нормировки вероятностей p_{ft} в левой части и выделим переменную p_{ubvs} в правой части. Получим

$$p_{ft}\theta_t - p_{ft}\theta_{ft} = p_{ft}\sum_{s\in F} n_{fs}\frac{\sum\limits_{v\in V}\lambda_{tv}p_{vs}}{p(f,s)} + p_{ft}\sum_{b\in F} n_{bf}\frac{\sum\limits_{u\in V}p_{bu}\lambda_{ut}}{p(b,f)} + p_{ft}\frac{\partial R}{\partial p_{ft}}$$

Учитывая условия дополняющей нежёсткости в (5), получим:

$$p_{ft}\theta_t = \sum_{s\in F} n_{fs}\sum_{v\in V}\underbrace{\frac{p_{ft}\lambda_{tv}p_{vs}}{p(f,s)}}_{p_{tfvs}} + \sum_{b\in F}\sum_{u\in V} n_{bf}\frac{p_{bu}\lambda_{ut}p_{ft}}{p(b,f)} + p_{ft}\frac{\partial R}{\partial p_{ft}},$$

$$p_{ft}\theta_t = \sum_{s\in F}\sum_{v\in V} n_{fs}p_{tfvs} + \sum_{b\in F}\sum_{u\in V} n_{bf}p_{ubtf} + p_{tf}\frac{\partial R}{\partial p_{tf}} = n_{ft} + n'_{ft} + p_{tf}\frac{\partial R}{\partial p_{tf}}.$$

Условие $\theta_t \leqslant 0$, противоречит условию регулярности t. Это означает, что $p_{ft} \equiv 0 \ \ \forall f$. Следовательно, $\theta_t > 0$, $p_{ft} \geqslant 0$. Левая часть уравнения неотрицательная, значит, правая часть также неотрицательна, и

$$p_{ft}\theta_t = \left(n_{ft} + n'_{ft} + p_{tf}\frac{\partial R}{\partial p_{tf}}\right)_+. \tag{14}$$

Просуммируем обе части уравнения по всем $f \in F$:

$$\theta_t = \sum_{f\in F}\left(n_{ft} + n'_{ft} + p_{tf}\frac{\partial R}{\partial p_{tf}}\right)_+. \tag{15}$$

Подставив в (14) выражение (15) для θ_t, получим:

$$p_{ft} = \frac{\left(n_{ft} + n'_{ft} + p_{tf}\frac{\partial R}{\partial p_{tf}}\right)_+}{\sum\limits_{f\in F}\left(n_{ft} + n'_{ft} + p_{tf}\frac{\partial R}{\partial p_{tf}}\right)_+}. \tag{16}$$

Доказательство для λ_{uv}: в точночти повторяет доказательство в теореме (1).

Теорема доказана. ■

Аналогично трехматричной модели, в двухматричной модели также потребуем непротиворечивости в данных транзакций.

4 Регуляризаторы

Рассмотрим несколько требований и соответствующих им регуляризаторов.

1. Предполагается, что основной объём транзакций возможно описать разреженным графом, то есть в матрице Λ много нулевых элементов. Введём регуляризатор разреживания матрицы Λ, максимизирующий KL-дивергенцию между равномерным распределением и $\lambda_{uv} = p(u, v)$:

$$R(\Lambda) = -\sum_{u \in V} \sum_{v \in V} \gamma_{uv}^{\Lambda} \log \lambda_{uv} \to \max,$$

где γ_{uv}^{Λ} — матрица размера $|V| \times |V|$, которая является маской, показывающей, для каких элементов Λ применять регуляризатор.

2. Предполагается, что фирма не может иметь слишком много видов деятельности, поэтому в матрицах P и Q также много нулевых элементов. Введём регуляризаторы разреживания матриц P и Q, максимизирующие KL-дивергенции между распределениями $p(b \mid u)$, $p(s \mid v)$ и равномерным распределением на множестве фирм F:

$$R(P) = -\sum_{b \in F} \sum_{u \in V} \gamma_{bu}^{P} \log p_{bu} \to \max;$$

$$R(Q) = -\sum_{s \in F} \sum_{v \in V} \gamma_{sv}^{Q} \log q_{sv} \to \max,$$

где $\gamma_{bu}^{P},\ \gamma_{sv}^{Q}$ — матрицы размера $|F| \times |V|$, которые являются масками, показывающими, для каких элементов $P,\ Q$ соответственно применять регуляризатор.

3. Для каждой фирмы известен перечень кодов ОКВЭД (общероссийского классификатора видов экономической деятельности). Обозначим через W множество всех кодов ОКВЭД, а через $W_f \subset W$ — перечень фирмы f из F. На практике эти перечни могут соответствовать реальной деятельности фирм весьма приблизительно. Поэтому будем предполагать существование вероятностной взаимосвязи между условными распределениями $p(v \mid s)$, $v \in V$ и $p(w \mid s)$, $w \in W$ для каждой фирмы $s \in F$ как продавца товаров и услуг. Эту взаимосвязь будем описывать неизвестными условными вероятностями $\psi_{wv} = p(w \mid v)$:

$$p(w \mid s) = \sum_{v \in V} \psi_{wv} p(v \mid s).$$

Предполагается, что условное распределение $p(w \mid s)$ приблизительно соответствует перечню кодов ОКВЭД фирмы s. Последний представим в виде равномерного

распределения на подмножестве W_s:

$$\hat{p}(w \mid s) = \frac{1}{|W_s|}[w \in W_s].$$

Для оценивания матрицы параметров модели $\Psi = (\psi_{wv})_{W \times V}$ применим принцип максимума правдоподобия:

$$R(Q, \Psi) = \sum_{s \in F} \sum_{w \in W_s} \log \sum_{v \in V} \psi_{wv} q_{sv} \to \max_{Q, \Psi}.$$

Ещё один регуляризатор на матрицу Ψ получим, если потребуем её разреженности:

$$R(\Psi) = -\sum_{v \in V} \sum_{w \in W} \log \psi_{wv} \to \max_{\Psi}.$$

Аналогичный регуляризатор получается для матрицы P.

5 Эксперименты на модельных данных

5.1 Исходные данные и метрики качества восстановления матриц

Целью серий экспериментов на модельных данных является исследование устойчивости модели, ее способности восстанавливать как матрицу $N = (n_{bs})_{F \times F}$, так и матрицы $P,\ Q,\ \Lambda$.

Далее модельные матрицы будем обозначать $P^*,\ Q^*,\ \Lambda^*$, матрицы для инициализации EM-алгоритма — $P^0,\ Q^0,\ \Lambda^0$, а восстановленные матрицы — $P^1,\ Q^1,\ \Lambda^1$ соответственно.

Генерация модельных матриц P^*, Q^*:

Введем параметр для генерации матриц P^*, Q^*: k — количство видов деятельности для каждой фирмы. Является общим для обеих матриц P^*, Q^*. В каждой строке этих матриц будет ровно k ненулевых элементов на случайных позициях. Ненулевые элементы матриц генерируются из равномерного распределения на $[0, 1]$.

Генерация модельной матрицы Λ^*:

Матрица Λ^* также генерируется из равномерного распределения на $[0, 1]$. Она имеет верхнетреугольный вид, с нулями на диагонали. Требование верхней треугольности матрицы Λ^* выступает в роли дополнительного регуляризатора: в графе нет циклов, товарный поток движется от сырья к конечному продукту, а финансовый поток движется ему навстречу Введем параметр генерации матрицы Λ^*: $\alpha \in \left[\frac{|V|(|V|+1)}{2}, 1 - \frac{2}{|V|-1}\right]$ — разреженность верхней треугольной части матрицы, $\alpha = \dfrac{2\sum\limits_{u,v \in V}[\lambda^*_{uv} = 0]}{|V|(|V|-1)}$.

Инициализация матриц $P^0,\ Q^0,\ \Lambda^0$:

В каждом эксперименте вид матриц для инициализации EM-алгоритма прописан отдельно. Общим для всех экспериментов является то, что ненулевые элементы этих матриц генерируются из равномерного распределения на $[\varepsilon, 1], \varepsilon = 1/2$.

Также для матриц $P^0,\ Q^0$ вводится параметр разреживания $\beta \in \left[0, \frac{|V|}{2}\right]$ — среднее количество ненулевых элементов в каждой строке этих матриц. Этот параметр имеет следующую интерпретацию: для каждой фирмы из внешних источников мы можем узнать, какими видами деятельности она точно не занимается. Тогда β обозначает количество видов деятельности фирмы. При этом справделиво полагать, что для разных фирм это количество разное, поэтому при генерации матриц $P^0,\ Q^0$ этот параметр варьируется от 1 до 2β. При этом инициализации все элементы, которые были при генерации ненулевыми, также ненулевые. Это значит, что параметр β задает, насколько мы информированы о видах деятельности фирм.

Метрики качества восстановления матриц:

Для определения качества восстановления матриц использовалось расстояние Хеллингера, которое, как известно, используется для измерения похожести двух вероятностных распределений:

$$H(A, B) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} \left(\sqrt{a_i} - \sqrt{b_i}\right)^2},$$

где $A = (a_1, \ldots, a_k), B = (b_1, \ldots, b_k)$ — два дискретных распределения, $H(A, B)$ — расстояние Хеллингера между ними.

Так как в столбцах матриц P, Q находятся распределения (нормированные, неотрицательные вектора), то расстоянием Хеллингера между восстановленной и мо-

дельной матрицей будем считать среднее расстояние между их столбцами. В случае матрицы Λ расстояние Хеллингера измеряется между вытянутыми в вектор восстановленной и модельной матрицами.

Так же при проведении экспериментов строится поведение логарифма правдоподобия (2).

5.2 Эксперимент 1: инициализация модельными матрицами

Целью данного тривиального эксперимента является проверка того, что модель способна восстановить модельные матрицы, если подать их в качестве инициализации.

Условия эксперимента: $|F| = 100,\ |V| = 30,\ \alpha = 0.9,\ k = 3,\ \beta = 0.$

Инициализация $P^0,\ Q^0,\ \Lambda^0$: модельные матрицы.

Вывод: в результате расстояния Хеллингера после первой же итерации для всех трех матриц равны вычислительным погрешностям ($\approx 10^{-7}$), поэтому графики этих расстояний не приводятся.

5.3 Эксперимент 2: инициализация случайными матрицами со структурой разреженности модельных матриц

Целью данного эксперимента является проверка того, что модель способна восстановить модельные матрицы, если известна точная структура их разреженности.

Условия эксперимента: $|F| = 100,\ |V| = 30,\ \alpha = 0.85,\ k = 5,\ \beta = 0.$

Инициализация $P^0,\ Q^0$: случайные матрицы со структурой разреженности $P^*,\ Q^*$.

Вывод: на графике 4 видно, что модель сходится до точного решения.

5.4 Эксперимент 3.1: зависимость качества восстановления от β

Целью данного эксперимента является проверка гипотезы о том, что при уменьшении априорных знаний о том, какими видами деятельности может заниматься фирма, качество восстановления падает.

Условия эксперимента: $|F| = 100,\ |V| = 30,\ \alpha = 0.9,\ k = [3, 10],$
$\beta = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13].$

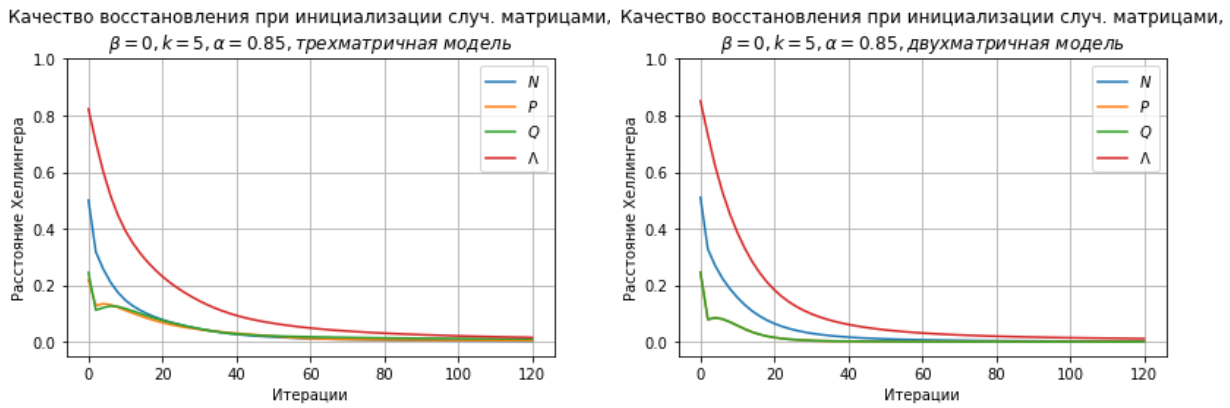

Рис. 4: Эксперимент 2: инициализация случайными матрицами со структурой разреженности модельных матриц.

Вывод: на графике 5 видно, что для трехматричной модели при увеличении β качество восстановления падает, в то время как для двуматричной модели с теми же параметрами модель полностью восстанавливает три матрицы. При увеличении плотности модельных матриц $P^*,\ Q^*$ в трехматричной модели матрица N^1 восстанавливается хорошо, а то время как $P^1,\ Q^1,\ \Lambda^1$ существенно хуже, чем при разреженных $P^*,\ Q^*$. Для двуматричной модели в случае более плотных матриц P^* матрица P^1 восстанавливается более устойчиво, и качество восстановления лучше, чем для трехматричной модели.

5.5 Эксперимент 3.2: зависимость качества восстановления от α

Целью данного эксперимента является проверка гипотезы о том, что при увеличении разреженности Λ качество восстановления должно увеличиваться.

Условия эксперимента: $|F| = 100,\ |V| = 30,\ \beta = [2, 6],\ k = [3, 10],$
$\alpha = [0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9]$.

Вывод: на графике 6 видно, что для трехматричной модели поведение не устойчиво и при увеличении плотности $P^0,\ Q^0$ качество восстановления падает. Для двуматричной модели ситуация обратная: ее поведение более устойчиво и при увеличении плотности модельных матриц качество не ухудшается.

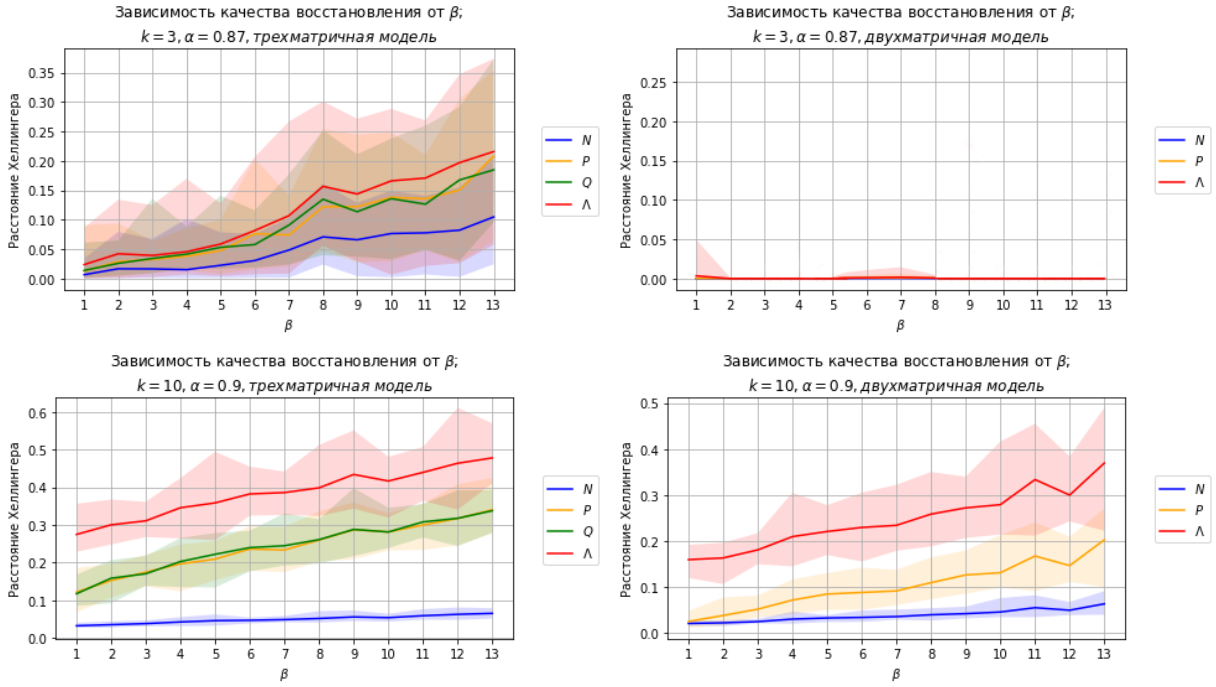

Рис. 5: Эксперимент 3.1: зависимость качества восстановления от β.

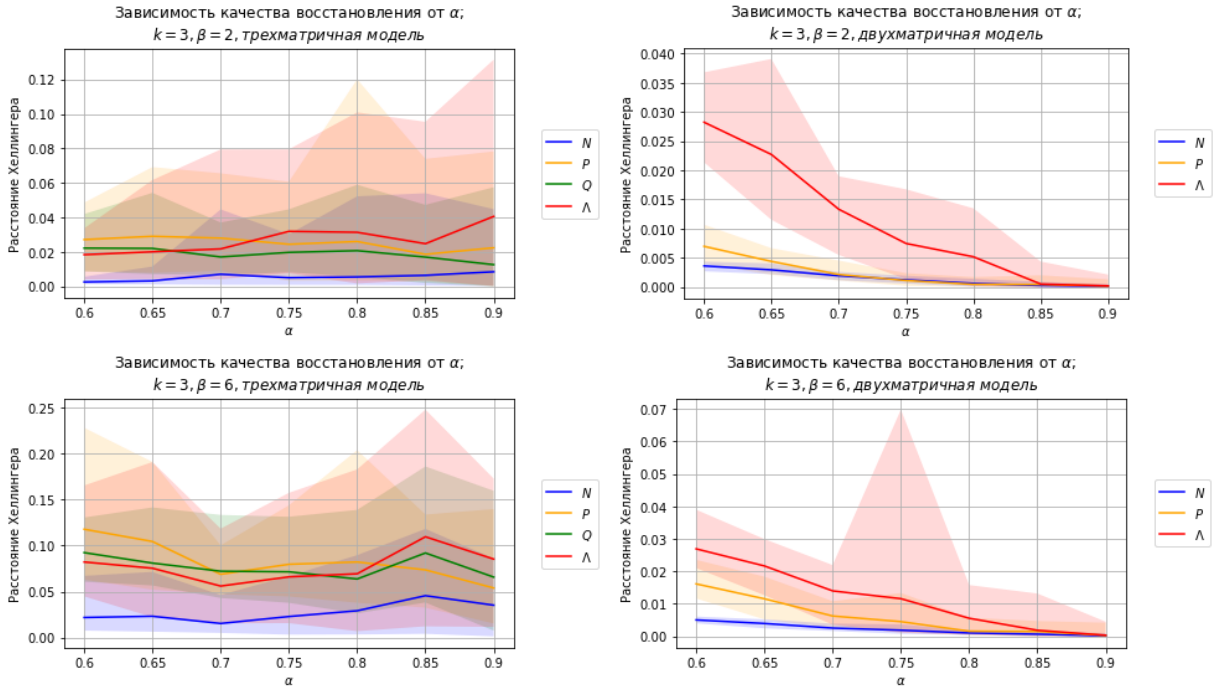

Рис. 6: Эксперимент 3.1: зависимость качества восстановления от α.

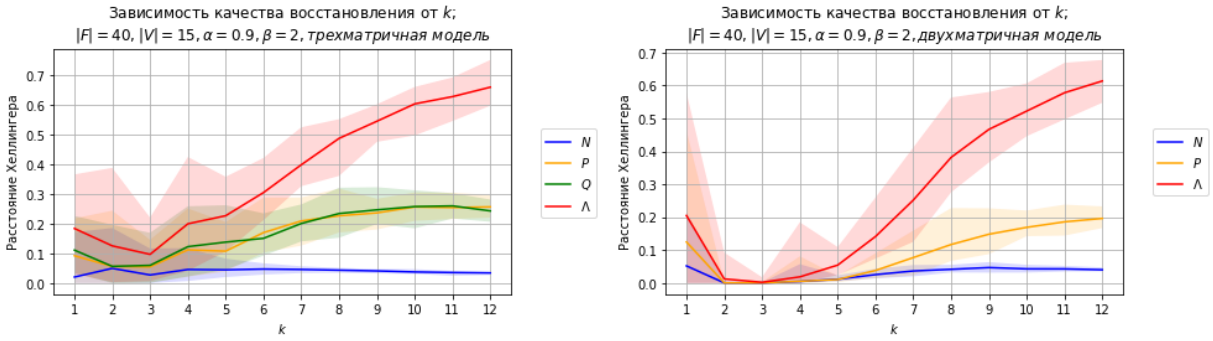

Рис. 7: Эксперимент 4: зависимость качества восстановления от k.

5.6 Эксперимент 4: зависимость качества восстановления от k

Целью данного эксперимента является проверка гипотезы о том, что при увеличении количества реальных видов деятельности фирм при сильно разреженной матрице Λ качество восстановления в двухматричной модели должно быть выше, чем в трехматричной.

Условия эксперимента: $|F| = 40,\ |V| = 15,\ \beta = 2,\ \alpha = 0.9,$
$k = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13].$

Вывод: на графике 7 видно, что чем плотнее исходные $P^*,\ Q^*$, тем хуже восстанавливаются матрицы $P^0,\ Q^0,\ \Lambda^0$ в обеих моделях; двухматричная модель намного более устойчивая, и качество ее восстановления выше.

5.7 Эксперимент 5: зависимость качества восстановления от коэффициента разреживания τ_Λ

Целью данного эксперимента проверить гипотезу о том, что модели способны восстанавливать разреженные модельные матрицы при включении регуляризатора разреживания Λ, и найти границы коэффициентов разреженности, когда модель уже не способна восстановить модельные матрицы.

Условия эксперимента: $|F| = 50,\ |V| = 15,\ \beta = 1,\ \alpha = 0.85,\ k = 4.$

Вывод: на графике 8 видно, что регуляризация позволяет снизить объем априорной информации. При $k = 6, \beta = 0$, то есть когда известна точная структура разреженности, а матрица Λ^* разрежена, использование регулризатора позволяет значительно повысить качество восстановления. Это верно вплоть до ситуации, когда $k = 8, \beta = 3,$

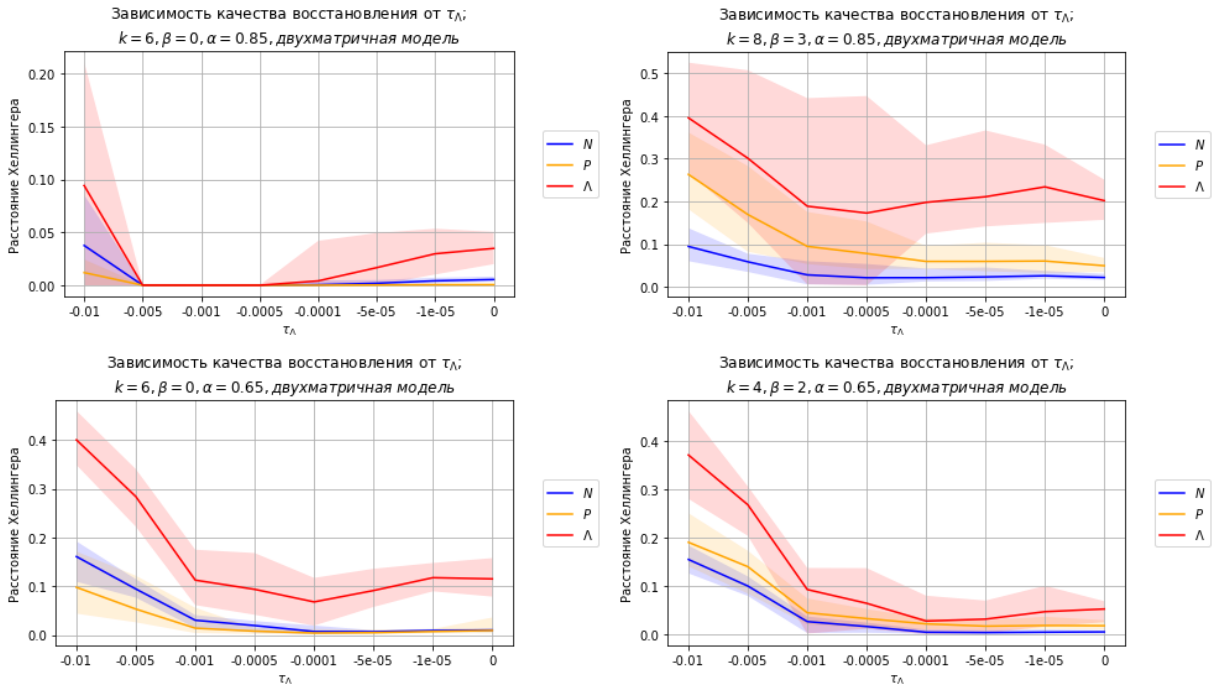

Рис. 8: Эксперимент 5: зависимость качества восстановления от коэффициента регуляризатора разреживания τ_Λ.

то есть когда ничего не известно о матрице P^*. В случае максимально плотной Λ ситуация похожая: при большом объеме данных регуляризация позволяет повысить качество восстановления, а при малом объеме априорной информации модель тоже можно улучшить, но прирост качества будет ниже, чем при разреженной Λ.

6 Заключение

В данной работе были предложены две латентные вероятностные модели, трехматричная и двухматричная, для выявления видов экономической деятельности фирм по банковским транзакционным данным. Также были предложены способы регуляризации этих моделей.

В экспериментах на синтетических данных выявлены условия устойчивого восстановления латентной информации о фирмах. Эксперименты показали, что при увеличении разреженности матриц качество восстановления растет в обеих моделях. При этом исходная матрица транзакций хорошо восстанавливается независимо от разреженности данных. Двухматричная модель более устойчивая, чем трехматричная, и качество ее восстановления выше. Также было показано, что использование регуляризаторов позволяет отказаться от априорной информации о фирмах в двухматричной модели.

Список литературы

[1] *Vorontsov KV.* Additive regularization for topic models of text collections // Doklady Mathematics / Pleiades Publishing. — Vol. 89. — 2014. — Pp. 301–304.

[2] The architecture of transaction networks: a comparative analysis of hierarchy in two sectors / Jianxi Luo, Carliss Y Baldwin, Daniel E Whitney, Christopher L Magee // *Industrial and Corporate Change*. — 2012.

[3] *Baldwin Carliss Y.* Where do transactions come from? Modularity, transactions, and the boundaries of firms // *Industrial and corporate change*. — 2008. — Vol. 17, no. 1. — Pp. 155–195.

[4] Dynamic visualization of large transaction networks: the daily Dutch overnight money market / Ronald Heijmans, Richard Heuver, Clement Levallois, Iman van Lelyveld. — 2014.

[5] *Giannotti Fosca, Gozzi Cristian, Manco Giuseppe.* Clustering Transactional Data // Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery. — PKDD '02. — Springer-Verlag, 2002. — Pp. 175–187.

[6] *Akeyama Yuki, Akiyama Yuki, Shibasaki Ryosuke.* A New Method of Estimating Locality of Industry Cluster Regions Using Large-scale Business Transaction Data // *Proceedings of CUPUM*. — 2015. — Vol. 347. — Pp. 1–13.

[7] Network topology of the interbank market / Michael Boss, Helmut Elsinger, Martin Summer, Stefan Thurner 4 // *Quantitative Finance*. — 2004. — Vol. 4, no. 6. — Pp. 677–684.