Байесовская и классическая регуляризация в вероятностном тематическом моделировании

д.ф.-м.н., проф. РАН Воронцов Константин Вячеславович

(Лаборатория машинного интеллекта МФТИ)

научный семинар

«Актуальные проблемы прикладной математики»

НГУ • Математический центр в Академгородке • 19 февраля 2021

Содержание

Что такое «тема» в коллекции текстовых документов?

- *тема* — специальная терминология предметной области
- *тема* — набор часто совместно встречающихся терминов
- *тема* — семантически однородный кластер текстов

Тематическая модель автоматически выявляет латентные темы по наблюдаемым распределениям слов $p(w|d)$ в документах.

Имея коллекцию текстовых документов, хотим узнать:

- из каких тем состоит коллекция,
  $p(t)$ — вероятность (доля) темы t в коллекции;
- из каких тем состоит каждый документ,
  $p(t|d)$ — вероятность (доля) темы t в документе d;
- из каких слов или терминов состоит каждая тема,
  $p(w|t)$ — вероятность (доля) слова w в теме t.

Пример 1. Мультиязычная модель Википедии

216 175 русско-английских пар статей.
Первые 10 слов и их частоты $p(w|t)$ в %:

Тема №68				Тема №79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример 1. Мультиязычная модель Википедии

216 175 русско-английских пар статей.

Первые 10 слов и их частоты $p(w|t)$ в %:

Тема №88				Тема №251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример 2. Биграммная модель научных конференций

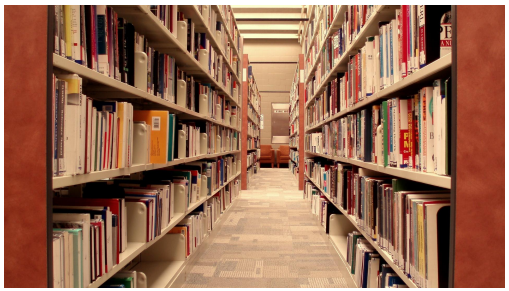
Коллекция 1000 статей конференций ММРО, ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
униграммы	биграммы	униграммы	биграммы
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

Сергей Стенин. Мультиграммные аддитивно регуляризованные тематические модели // Магистерская диссертация, МФТИ, 2015.

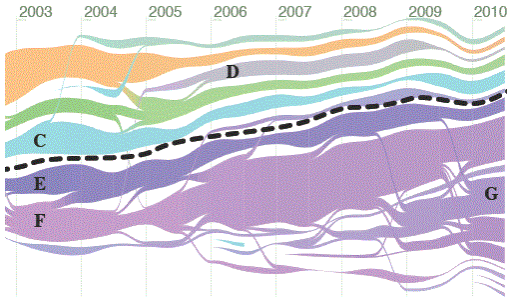
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

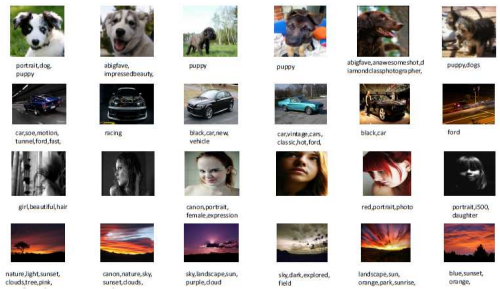


поиск тематического контента в соцсетях

выявление и отслеживание цепочек новостей

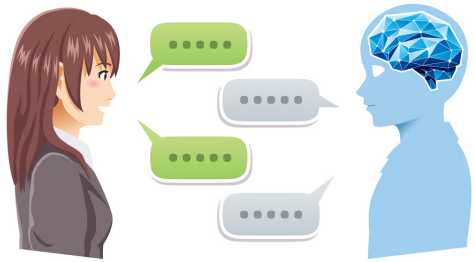


мультимодальный поиск текстов и изображений



анализ банковских транзакционных данных

управлением диалогом в разговорном интеллекте



Пусть

- W — конечное множество *термов* (слов, терминов)
- D — конечное множество текстовых документов
- T — конечное множество тем
- порядок слов в документе не важен (bag of words)
- порядок документов в коллекции не важен
- каждый терм w в документе d связан с некоторой темой t
- $D \times W \times T$ — дискретное вероятностное пространство
- коллекция — это i.i.d. выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- гипотеза условной независимости: $p(w|d, t) = p(w|t)$

Тематическая модель, по формуле полной вероятности:

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

Задача построения тематической модели

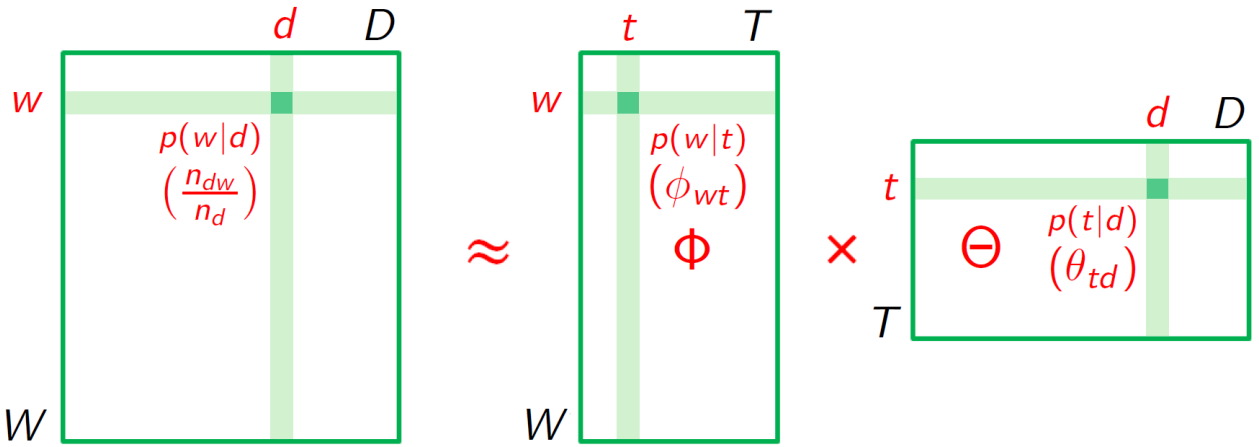
Дано: коллекция текстовых документов

- n_{dw} — частоты термов в документах, $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

Найти: параметры тематической модели $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$

- $\phi_{wt} = p(w|t)$ — вероятности термов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Задача низкорангового стохастического матричного разложения:



Критерий максимума правдоподобия

Правдоподобие — плотность распределения выборки $(d_i, w_i)_{i=1}^n$:

$$\prod_{i=1}^{n} p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(d, w)^{n_{dw}}$$

Максимизация логарифма правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d) \underbrace{p(d)}_{\text{const}} \to \max_{\Phi, \Theta}$$

эквивалентна максимизации функционала

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

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w \in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t \in T} \theta_{td} = 1.$$

Модель PLSA (Probabilistic Latent Semantic Analysis)

Задача максимизации log-правдоподобия с ограничениями:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta}; \quad \sum_w \phi_{wt} = 1; \ \sum_t \theta_{td} = 1; \ \phi_{wt} \geqslant 0; \ \theta_{td} \geqslant 0$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} \equiv p(t|d,w) = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \text{M-шаг:} & \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum\limits_{d\in D} n_{dw} p_{tdw}\right) \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum\limits_{w\in d} n_{dw} p_{tdw}\right) \end{cases}$$

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

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

Недостатки PLSA и необходимость регуляризации

1. Большая размерность пространства параметров
2. Якобы из-за этого сильное переобучение
3. Якобы невозможность моделирования новых документов
4. Неединственность решения — матричного разложения:
   если $\Phi\Theta$ — решение, то $(\Phi S)(S^{-1}\Theta)$ — тоже решение
5. Нет управления *разреженностью* Φ и Θ, т.к.
   (в начале $\phi_{wt} = 0$) $\Leftrightarrow$ (в финале $\phi_{wt} = 0$),
   (в начале $\theta_{td} = 0$) $\Leftrightarrow$ (в финале $\theta_{td} = 0$)
6. Темы не всегда интерпретируемы
7. Нет механизмов учёта дополнит. данных и ограничений
8. Нет выделения нетематических (фоновых) слов

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. JMLR, 2003.

Гипотеза об априорных распределениях Дирихле

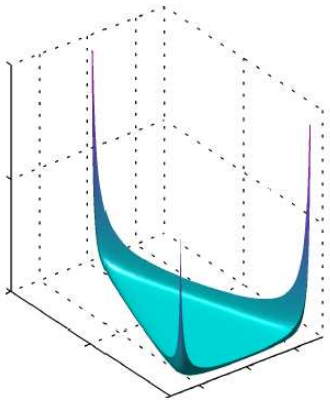
Гипотеза: вектор-столбцы $\phi_t = (\phi_{wt})_{w \in W}$ и $\theta_d = (\theta_{td})_{t \in T}$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \phi_{wt} > 0; \quad \beta_0 = \sum_w \beta_w, \ \beta_w > 0;$$

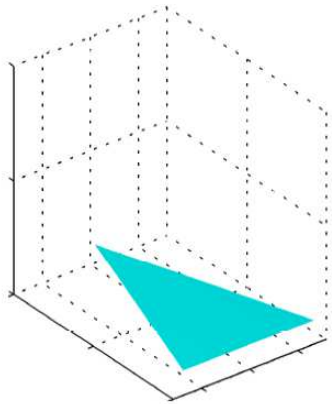
$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \theta_{td} > 0; \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t > 0;$$

Пример:

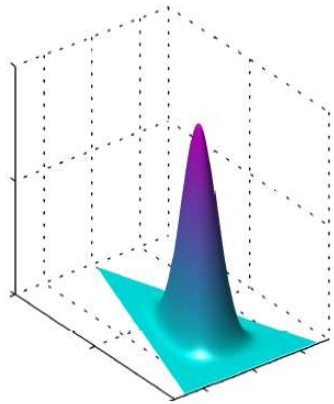
$\mathrm{Dir}(\theta|\alpha)$,
$|T| = 3$,
$\theta, \alpha \in \mathbb{R}^3$



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$

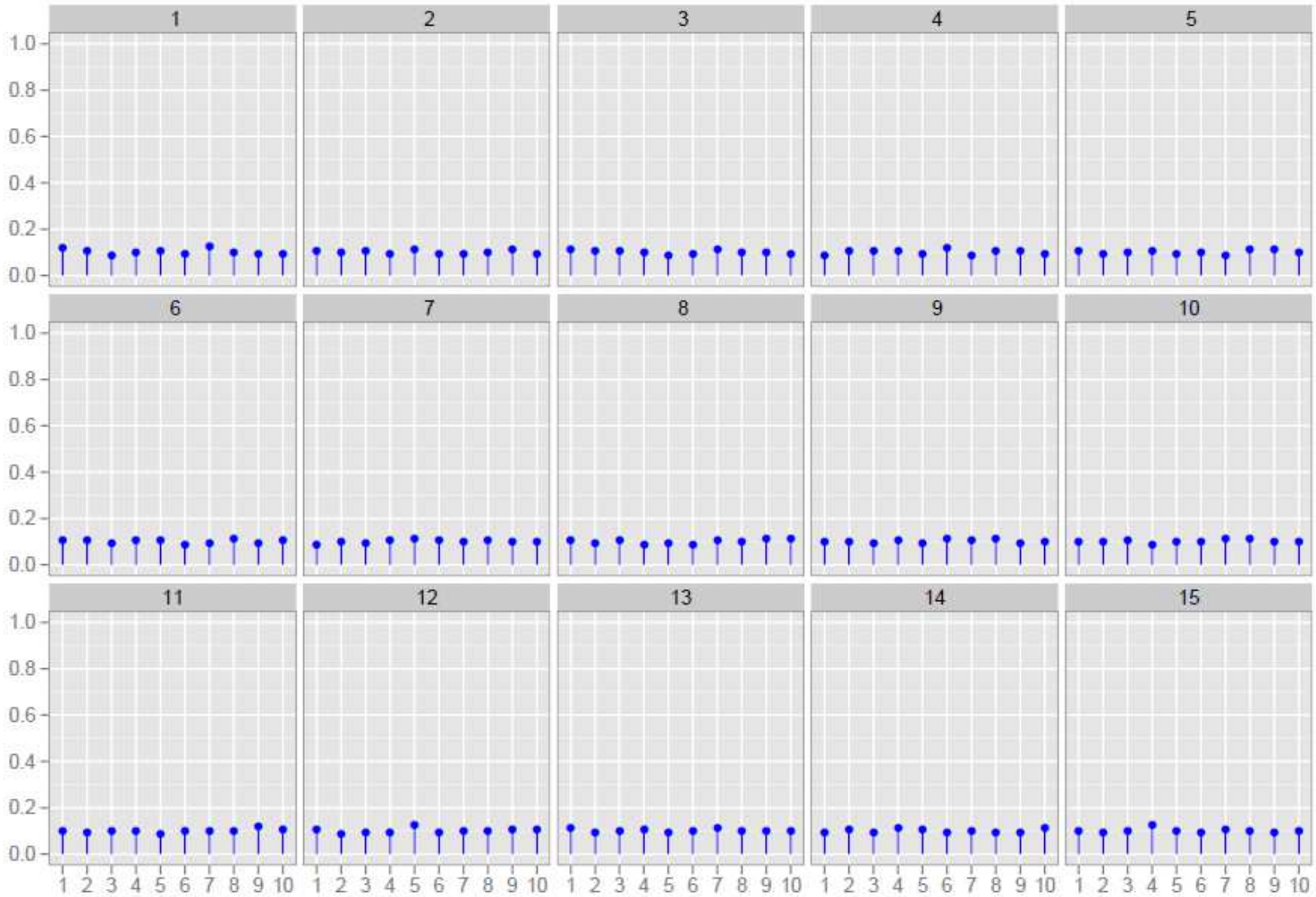


$\alpha_1 = \alpha_2 = \alpha_3 = 1$

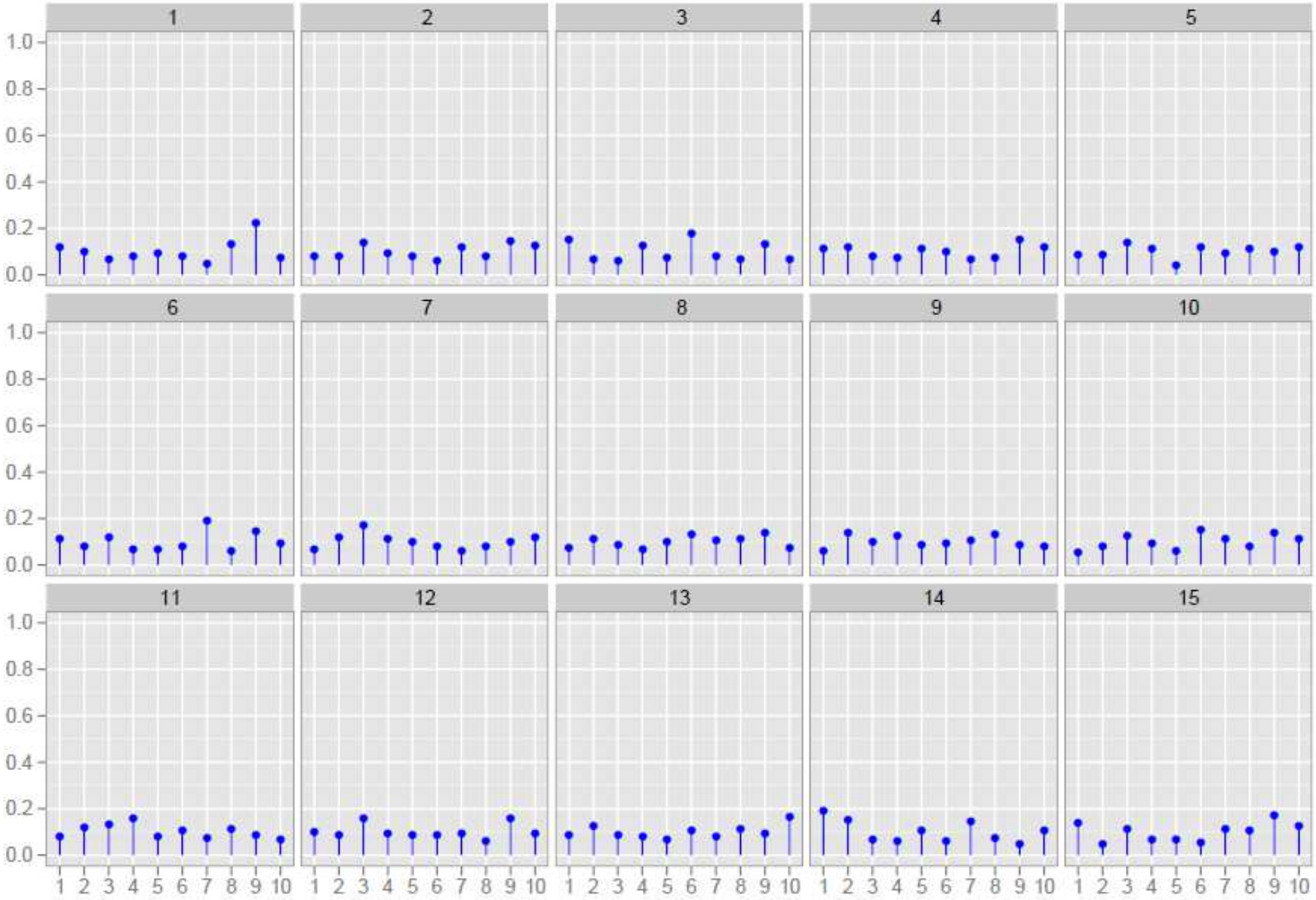


$\alpha_1 = \alpha_2 = \alpha_3 = 10$

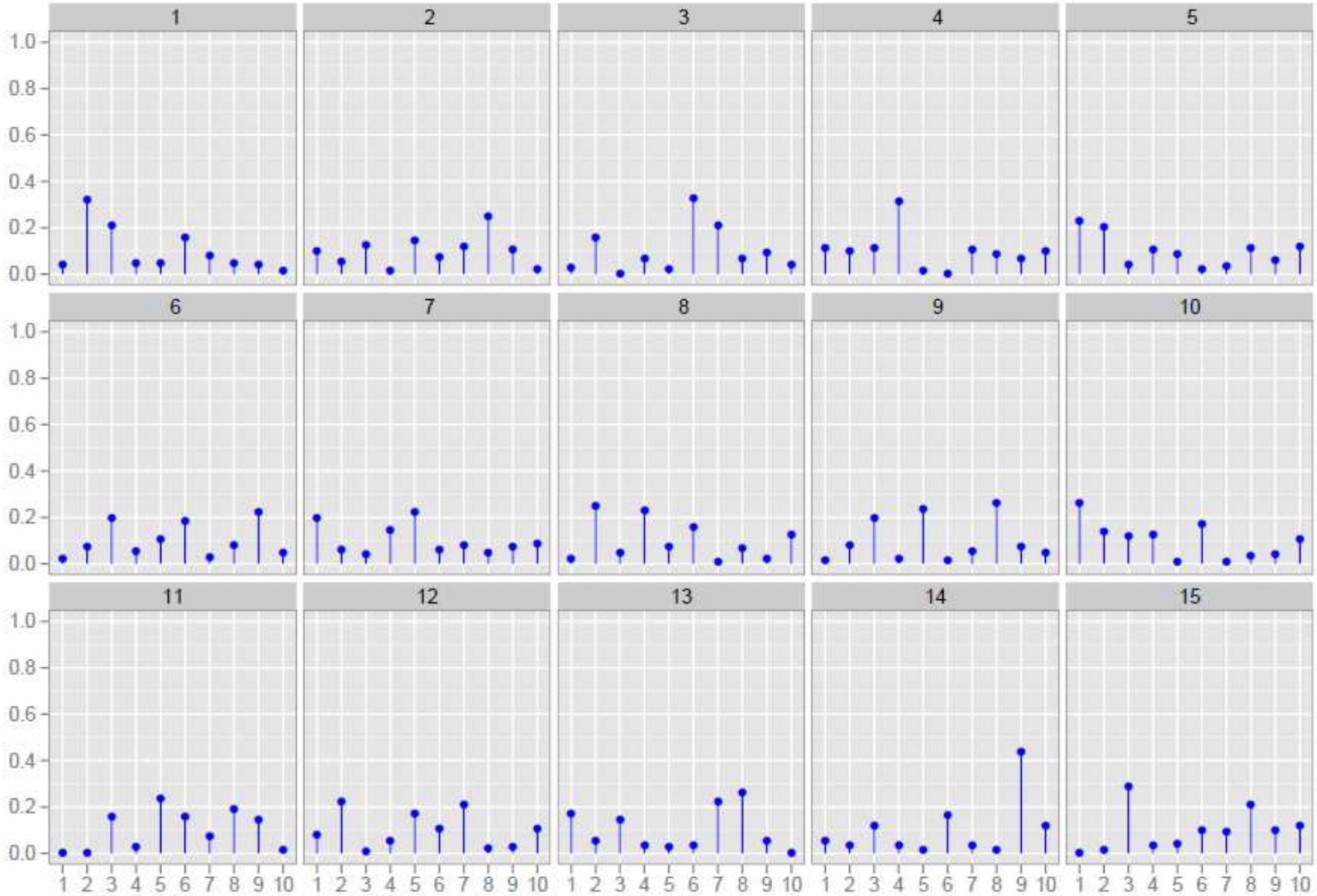
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 100$, 10 тем, 15 документов



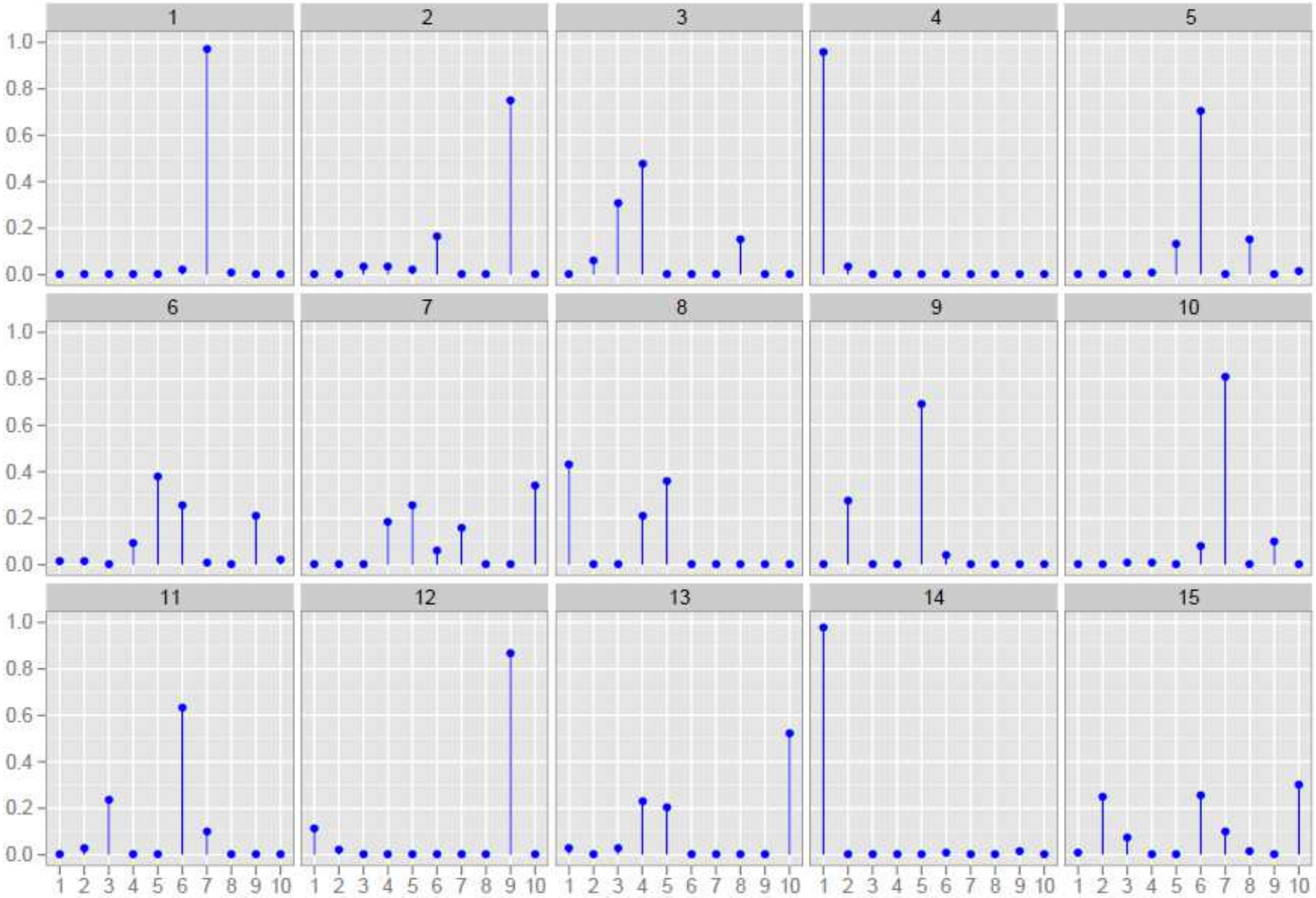
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 10$, 10 тем, 15 документов



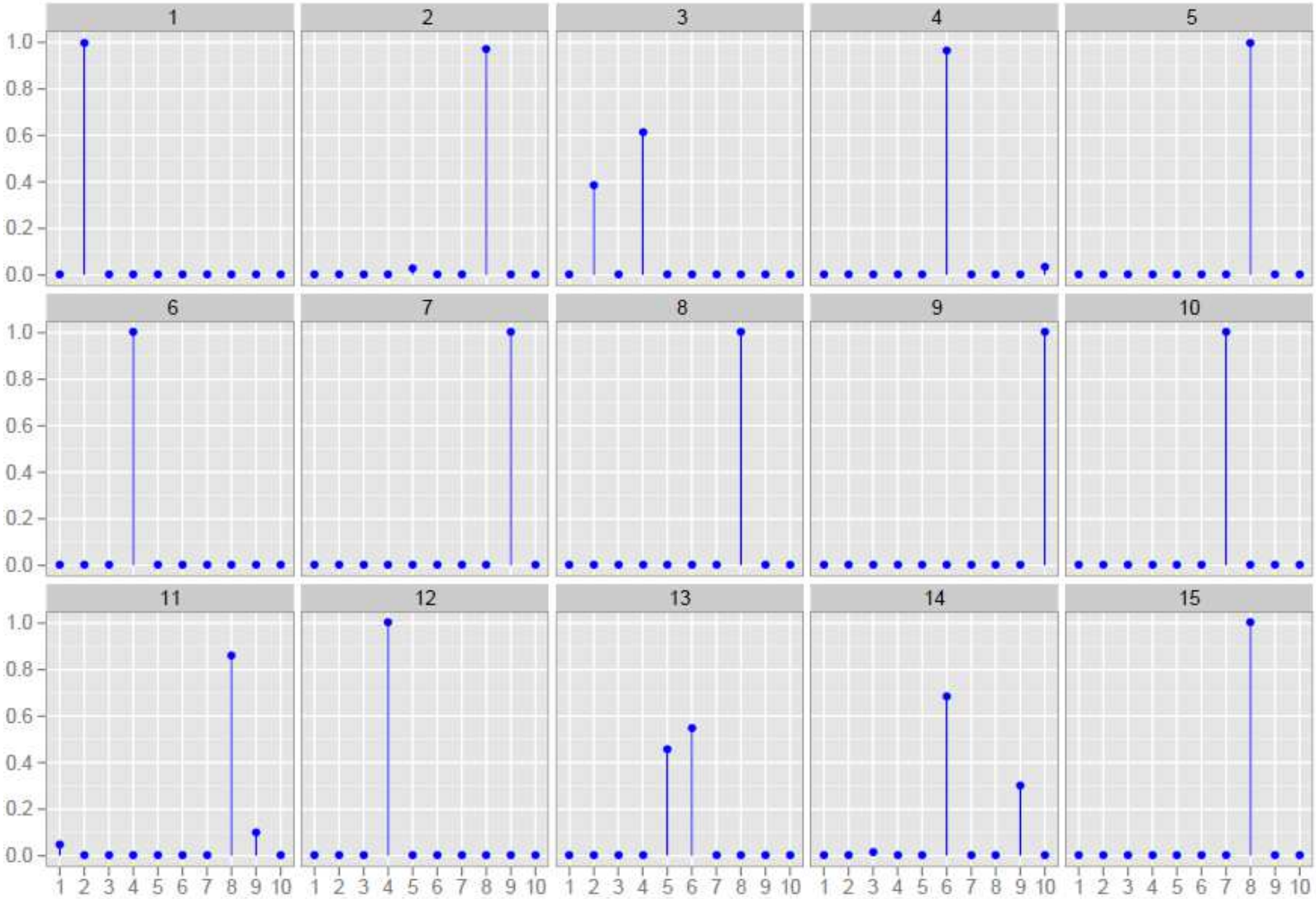
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 1$, 10 тем, 15 документов



Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 0.1$, 10 тем, 15 документов



Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 0.01$, 10 тем, 15 документов



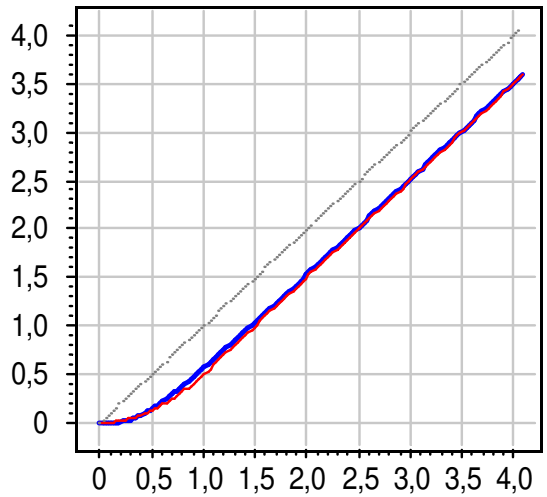
Некоторые свойства распределения Дирихле

1. Матожидание: $\mathrm{E}\theta_t = \int \theta_t \mathrm{Dir}(\theta|\alpha)\, d\theta = \frac{\alpha_t}{\alpha_0} = \underset{t}{\mathrm{norm}}(\alpha_t)$
2. Мода: $\hat{\theta}_t = \frac{\alpha_t - 1}{\alpha_0 - T} = \underset{t}{\mathrm{norm}}(\alpha_t - 1)$
3. Дисперсия: $\mathrm{D}\theta_t = \frac{\alpha_t(\alpha_0 - \alpha_t)}{\alpha_0^2(\alpha_0+1)}$
4. Матожидание ln: $\mathrm{E}\ln\theta_t = \int \ln\theta_t \mathrm{Dir}(\theta|\alpha)\, d\theta = \psi(\alpha_t) - \psi(\alpha_0)$

где $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ — дигамма-функция.

Простая, но очень точная аппроксимация экспоненты от дигамма-функции:

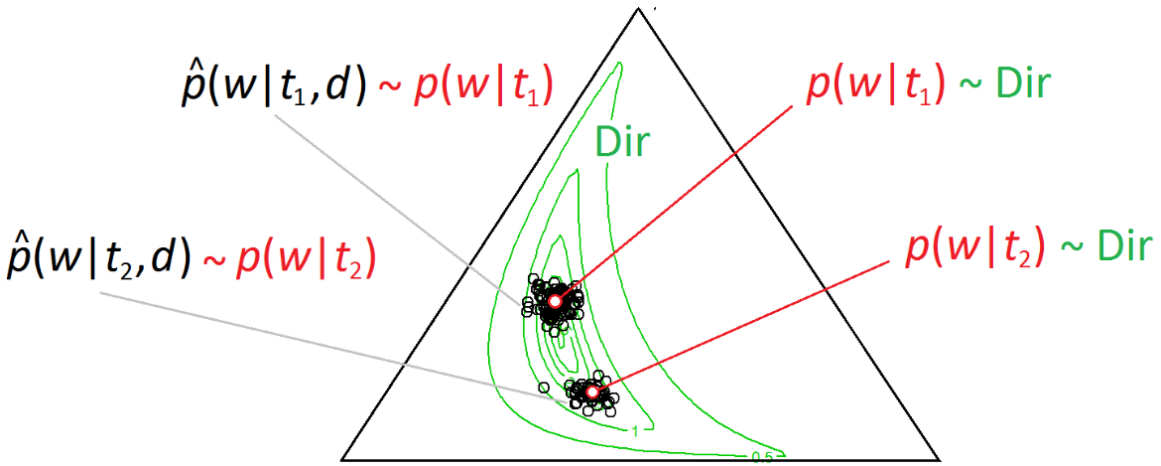
$$E(x) = \exp(\psi(x)) \approx \begin{cases} \frac{x^2}{2}, & 0 \leqslant x \leqslant 1 \\ x - \frac{1}{2}, & 1 \leqslant x \end{cases}$$



Почему именно распределение Дирихле?

- оно способно порождать разреженные векторы;
- имеет параметры, управляющие степенью разреженности;
- описывает кластерные структуры на симплексе (см. рис.);
- является сопряжённым с мультиномиальным распределением, что сильно упрощает байесовский вывод (см. далее).

Распределение $\mathrm{Dir}(\phi|\alpha)$ порождает векторы тем $\phi_t = p(w|t)$, которые порождают мультиномиальные распределения $\hat{p}(w|t,d)$:



Основы байесовской регуляризации

Введём более общие обозначения:

$X = (d_i, w_i)_{i=1}^n$ — исходные данные, *наблюдаемые переменные*
$\Omega = (\Phi, \Theta)$ — параметры порождающей модели $p(X|\Omega)$
$\gamma = (\beta, \alpha)$ — гиперпараметры *априорного распределения* $p(\Omega|\gamma)$

Задача: по X найти Ω.

Формула Байеса даёт *апостериорное распределение* $p(\Omega|X,\gamma)$,
где символ $\propto$ означает «равно с точностью до нормировки»:

$$p(\Omega|X,\gamma) = \frac{p(\Omega,X|\gamma)}{p(X|\gamma)} \propto p(\Omega,X|\gamma) \propto p(X|\Omega)\,p(\Omega|\gamma)$$

Далее есть два пути:

- Максимизация правдоподобия: $\Omega = \arg\max_\Omega \ln p(\Omega|X,\gamma)$
- Байесовский вывод: вычисление распределения $p(\Omega|X,\gamma)$

Максимизация апостериорной вероятности для модели LDA

Максимизация *совместного правдоподобия* данных и модели, называется также *Maximum a Posteriori (MAP) estimation*:

$$\ln p(X|\Omega)\,p(\Omega|\gamma) = \ln \prod_{i=1}^{n} p(d_i, w_i|\Phi,\Theta)\,p(\Phi|\beta)\,p(\Theta|\alpha) =$$

$$= \ln \prod_{d\in D}\prod_{w\in d} p(d,w|\Phi,\Theta)^{n_{dw}} \prod_{t\in T} \mathrm{Dir}(\phi_t|\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d|\alpha) \to \max_{\Phi,\Theta}$$

Это задача максимизации регуляризованного log-правдоподобия:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \sum_{t,w} \ln \phi_{wt}^{\beta_w - 1} + \sum_{d,t} \ln \theta_{td}^{\alpha_t - 1} \to \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w\in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t\in T} \theta_{td} = 1.$$

Регуляризованный EM-алгоритм для модели LDA

Максимизация апостериорной вероятности эквивалентна максимизации log-правдоподобия с регуляризатором:

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td}}_{\ln \text{ правдоподобия}} + \underbrace{\sum_{t,w}(\beta_w - 1)\ln \phi_{wt} + \sum_{d,t}(\alpha_t - 1)\ln \theta_{td}}_{\text{регуляризатор } R(\Phi,\Theta) = \ln p(\Phi,\Theta|\alpha,\beta)} \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw} p_{tdw} + \beta_w - 1\right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw} p_{tdw} + \alpha_t - 1\right) \end{cases}$$

Обобщение LDA: регуляризатор сглаживания и разреживания

Общий вид регуляризаторов сглаживания и разреживания:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in T}\sum_{w\in W} \beta_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d\in D}\sum_{t\in T} \alpha_{td} \ln \theta_{td} \to \max,$$

где $\beta_0 > 0$, $\alpha_0 > 0$ — коэффициенты регуляризации,

β_{wt}, α_{td} — параметры, задаваемые пользователем модели:

- $\beta_{wt} > 0$, $\alpha_{td} > 0$ — сглаживание
- $\beta_{wt} < 0$, $\alpha_{td} < 0$ — разреживание
- $\beta_{wt} > -1$, $\alpha_{td} > -1$ — модель LDA

Возможные применения сглаживания и разреживания:

- задать фоновые темы с общей лексикой языка
- задать шумовую тему для нетематичных термов
- задать псевдо-документ с ключевыми термами темы
- скорректировать состав термов и документов темы

Вероятностная модель со скрытыми переменными

Вернёмся к общей задаче $\ln p(\Omega|X,\gamma) \to \max_\Omega$:

$X = (d_i, w_i)_{i=1}^n$ — исходные данные, *наблюдаемые переменные*

$Z = (t_i)_{i=1}^n$ — *скрытые переменные*

$\Omega = (\Phi, \Theta)$ — параметры порождающей модели $p(X|\Omega)$

$\gamma = (\beta, \alpha)$ — гиперпараметры *априорного распределения* $p(\Omega|\gamma)$

Задача: по X найти Ω.

Апостериорное распределение:

$$p(\Omega|X,\gamma) \;\propto\; p(X|\Omega)\,p(\Omega|\gamma) \;=\; \sum_Z p(X,Z|\Omega)\,p(\Omega|\gamma)$$

Принцип максимума апостериорной вероятности:

$$\ln p(X|\Omega) + \underbrace{\ln p(\Omega|\gamma)}_{R(\Omega)} \;\to\; \max_\Omega$$

$R(\Omega)$ может и не иметь вероятностной интерпретации.

Общий EM-алгоритм для задачи со скрытыми переменными

Теорема. Точка Ω локального максимума регуляризованного *маргинализованного правдоподобия* (Marginal log-Likelihood)

$$\ln \sum_Z p(X, Z|\Omega) + R(\Omega) \;\to\; \max_\Omega \qquad \text{(RML)}$$

удовлетворяет системе уравнений, решение которой методом простых итераций сводится к чередованию двух шагов:

$$\text{E-шаг:} \quad q(Z) = p(Z|X, \Omega);$$

$$\text{M-шаг:} \quad \sum_Z q(Z) \ln p(X, Z|\Omega) + R(\Omega) \;\to\; \max_\Omega .$$

Это общий вид EM-алгоритма, используемый не только в тематическом моделировании.

A.P.Dempster, N.M.Laird, D.B.Rubin. Maximum likelihood from incomplete data via the EM algorithm. 1977.

Доказательство теоремы

Необходимые условия локального экстремума:

$$\frac{\partial}{\partial\Omega}\left(\ln\sum_z p(X,Z|\Omega)+R(\Omega)\right)=\frac{1}{p(X|\Omega)}\sum_z\frac{\partial p(X,Z|\Omega)}{\partial\Omega}+\frac{\partial R(\Omega)}{\partial\Omega}=0$$

По формуле условной вероятности $p(X|\Omega)=\frac{p(X,Z|\Omega)}{p(Z|X,\Omega)}$, подставляем:

$$\sum_z\frac{p(Z|X,\Omega)}{p(X,Z|\Omega)}\frac{\partial p(X,Z|\Omega)}{\partial\Omega}+\frac{\partial R(\Omega)}{\partial\Omega}=0$$

$$\sum_z\underbrace{p(Z|X,\Omega)}_{q(Z)}\frac{\partial}{\partial\Omega}\ln p(X,Z|\Omega)+\frac{\partial R(\Omega)}{\partial\Omega}=0$$

Это необходимые условия локального экстремума задачи M-шага, если $q(Z)$ рассматривать как константу, а не как функцию от Ω.

■

Ещё более общий EM-алгоритм и его сходимость

Теорема. Значение маргинализованного правдоподобия

$$\ln \sum_Z p(X, Z|\Omega) + R(\Omega) \;\to\; \max_\Omega \qquad \text{(RML)}$$

не убывает на каждом шаге итерационного процесса

$$\text{E-шаг:} \quad \mathrm{KL}\big(q(Z) \,\big\|\, p(Z|X,\Omega)\big) \;\to\; \min_q;$$

$$\text{M-шаг:} \quad \sum_Z q(Z) \ln p(X, Z|\Omega) + R(\Omega) \;\to\; \max_\Omega.$$

$q(Z) = p(Z|X,\Omega)$ является точным решением задачи E-шага.

Минимизация KL на E-шаге используется в тех случаях, когда $p(Z|X,\Omega)$ не удаётся вычислить в явном виде.

Сходимость *в слабом смысле*: глобальный max не гарантируется.

Доказательство теоремы

По формуле условной вероятности $p(X|\Omega) = \frac{p(X,Z|\Omega)}{p(Z|X,\Omega)}$.

Для произвольного распределения $q(Z)$

$$\ln p(X|\Omega) = \sum_Z q(Z) \ln p(X|\Omega) = \sum_Z q(Z) \ln \frac{p(X,Z|\Omega)}{p(Z|X,\Omega)} =$$

$$= \underbrace{\sum_Z q(Z) \ln \frac{p(X,Z|\Omega)}{q(Z)}}_{L(q,\Omega)} + \underbrace{\sum_Z q(Z) \ln \frac{q(Z)}{p(Z|X,\Omega)}}_{\mathrm{KL}\left(q(Z)\middle\|p(Z|X,\Omega)\right)\geqslant 0}$$

Максимизируем достижимую нижнюю оценку RML то по q, то по Ω:

$$\text{E-шаг:}\quad L(q,\Omega) + \cancel{R(\Omega)} \to \max_q \quad\Leftrightarrow\quad \mathrm{KL}\big(q(Z) \,\big\|\, p(Z|X,\Omega)\big) \to \min_q$$

$$\text{M-шаг:}\quad L(q,\Omega) + R(\Omega) \to \max_\Omega \quad\Leftrightarrow\quad \sum_Z q(Z) \ln p(X,Z|\Omega) + R(\Omega) \to \max_\Omega$$

На каждом шаге значение функционала может только возрастать.

■

Регуляризованный EM-алгоритм для тематической модели

Для тематической модели: $X=(d_i,w_i)_{i=1}^n$, $Z=(t_i)_{i=1}^n$, $\Omega=(\Phi,\Theta)$

Лемма. Точка (Φ,Θ) локального максимума RML (регуляризованного маргинализованного log-правдоподобия)

$$\ln\sum_Z p(X,Z|\Omega)+R(\Omega)=\sum_{d,w} n_{dw}\ln\sum_t \phi_{wt}\theta_{td}+R(\Phi,\Theta)$$

удовлетворяет системе уравнений, решение которой методом простых итераций сводится к чередованию двух шагов:

E-шаг: $$p(t|d,w)=\underset{t\in T}{\operatorname{norm}}\big(\phi_{wt}\theta_{td}\big),\quad \forall(d\in D, w\in d, t\in T)$$

M-шаг: $$\sum_{d,w,t} n_{dw}p(t|d,w)\ln\big(\phi_{wt}\theta_{td}\big)+R(\Phi,\Theta)\ \to\ \max_{\Phi,\Theta}$$

Доказательство леммы

E-шаг: в силу независимости элементов выборки и формулы Байеса

$$q(Z) = p(Z|X,\Omega) = \prod_{i=1}^{n} p(t_i|d_i,w_i) = \prod_{i=1}^{n} \underset{t_i\in T}{\operatorname{norm}}\left(\phi_{w_i t_i}\theta_{t_i d_i}\right)$$

M-шаг: подставим $q(Z)$ и $p(X,Z|\Omega)$ в общую формулу M-шага:

$$\sum_{Z\in T^n} q(Z)\ln p(X,Z|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

$$\sum_{t_1\in T}\cdots\sum_{t_n\in T}\prod_{k=1}^{n} p(t_k|d_k,w_k)\sum_{i=1}^{n}\ln p(d_i,w_i,t_i|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

$$\sum_{i=1}^{n}\sum_{t_1\in T}\cdots\sum_{t_n\in T}\prod_{k=1}^{n} p(t_k|d_k,w_k)\ln p(d_i,w_i,t_i|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

$$\sum_{i=1}^{n}\sum_{t\in T} p(t|d_i,w_i)\ln p(d_i,w_i,t|\Omega) + R(\Omega) \;\to\; \max_{\Omega}$$

$$\sum_{d\in D}\sum_{w\in W}\sum_{t\in T} n_{dw}\,p(t|d,w)\ln\left(\phi_{wt}\theta_{td}\right) + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}$$

■

Регуляризованный EM-алгоритм для тематической модели

Теорема. Точка (Φ, Θ) локального максимума задачи

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi,\Theta}$$

удовлетворяет системе уравнений

$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}} \right), & n_{wt} = \sum\limits_{d \in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}} \right), & n_{td} = \sum\limits_{w \in d} n_{dw} p_{tdw} \end{cases}$$

PLSA: $R(\Phi, \Theta) = 0$

LDA: $R(\Phi, \Theta) = \sum\limits_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum\limits_{d,t} (\alpha_t - 1) \ln \theta_{td}$

Доказательство. Вывод формул M-шага с регуляризатором

Преобразуем задачу M-шага из леммы, положив $p_{tdw} = p(t|d,w)$:

$$\sum_{w,t} \underbrace{\sum_d n_{dw} p_{tdw}}_{n_{wt}} \ln \phi_{wt} + \sum_{d,t} \underbrace{\sum_w n_{dw} p_{tdw}}_{n_{td}} \ln \theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

Чтобы применить условия KKT, выписываем лагранжиан:

$$\mathscr{L}(\Phi,\Theta) = \sum_{w,t} n_{wt} \ln \phi_{wt} - \sum_t \lambda_t \Big(\sum_w \phi_{wt} - 1\Big) + \\ + \sum_{d,t} n_{td} \ln \theta_{td} - \sum_d \mu_d \Big(\sum_t \theta_{td} - 1\Big) + R(\Phi,\Theta)$$

Условия KKT для стационарной точки лагранжиана:

$$\frac{\partial \mathscr{L}}{\partial \phi_{wt}} = \frac{n_{wt}}{\phi_{wt}} + \frac{\partial R}{\partial \phi_{wt}} - \lambda_t = 0$$

$$\Big(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}\Big)_+ = \lambda_t \phi_{wt}$$

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}} \Big(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}\Big)$$

$$\frac{\partial \mathscr{L}}{\partial \theta_{td}} = \frac{n_{td}}{\theta_{td}} + \frac{\partial R}{\partial \theta_{td}} - \mu_d = 0$$

$$\Big(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\Big)_+ = \mu_d \theta_{td}$$

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}} \Big(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\Big)$$

■

Байесовский вывод — основной подход в Topic Modeling

$X = (d_i, w_i)_{i=1}^n$ — наблюдаемые переменные, коллекция длины n
$Z = (t_i)_{i=1}^n$ — скрытые переменные
$\Omega = (\Phi, \Theta)$ — искомые параметры модели
$\gamma = (\beta, \alpha)$ — гиперпараметры априорных распределений

Задача байесовского вывода — получить не Ω, а $p(\Omega|X, \gamma)$

Вариационный байесовский вывод:

$$\text{вывести } p(Z, \Omega|X, \gamma) \propto p(X, Z|\Omega, \gamma)\, p(\Omega|\gamma)$$

Сэмплирование Гиббса:

$$\text{вывести } p(Z|X, \gamma) \text{ и сэмплировать } Z \sim p(Z|X, \gamma)$$
$$\text{вывести } p(\Omega|X, Z, \gamma) \propto p(X, Z|\Omega, \gamma)\, p(\Omega|\gamma)$$

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. JMLR, 2003.
Griffiths T., Steyvers M. Finding scientific topics. 2004.

Основная идея Variational Bayesian inference

$X = (d_i, w_i)_{i=1}^n$ — исходные данные, *наблюдаемые переменные*
$Z = (t_i)_{i=1}^n$ — *скрытые переменные*

$p(\Phi|\beta) = \prod_{t\in T} \mathrm{Dir}(\phi_t|\beta)$ — априорное распределение на Φ
$p(\Theta|\alpha) = \prod_{d\in D} \mathrm{Dir}(\theta_d|\alpha)$ — априорное распределение на Θ

Задача: найти апостериорное распределение $p(Z, \Phi, \Theta|X, \beta, \alpha)$.

Основная идея: найти его приближение в виде произведения $n + |T| + |D|$ распределений по блокам переменных t_i, ϕ_t, θ_d:

$$q(Z, \Phi, \Theta) = \prod_{i=1}^{n} q_i(t_i) \prod_{t\in T} q_t(\phi_t) \prod_{d\in D} q_d(\theta_d)$$

Обозначив $(Z, \Phi, \Theta) = Y$, $(\beta, \alpha) = \gamma$, перейдём к общей задаче

Основная теорема вариационного байесовского вывода

Теорема. Решение задачи $\mathrm{KL}\big(q(Y) \,\|\, p(Y|X,\gamma)\big) \to \min_q$ в семействе факторизованных распределений $q(Y) = \prod_j q_j(Y_j)$ по переменным Y_j, $j \in J$, удовлетворяет системе уравнений

$$\ln q_j(Y_j) = \mathsf{E}_{q\setminus j} \ln p(X, Y|\gamma) + \text{const},$$

где $\mathsf{E}_{q\setminus j}$ — матожидание по всем переменным кроме Y_j,
const — ln нормировочного множителя распределения q_j.

Для решения этой системы используют метод простой итерации.

Идея доказательства: расписываем $\mathrm{KL}(\cdot\|\cdot)$ и сводим задачу к

$$\sum_{Y_j} q_j(Y_j) \underbrace{\sum_{Y\setminus Y_j} \prod_{i\neq j} q_i(Y_i) \ln p(X, Y|\gamma)}_{\mathsf{E}_{q\setminus j} \ln p(X,Y|\gamma)} - \sum_{Y_j} q_j(Y_j) \ln q_j(Y_j) \;\to\; \min_q$$

Доказательство

1. В оптимизационной задаче можно перекидывать X через условную черту:

$$\sum_Y q(Y) \ln \frac{p(Y|X,\gamma)}{q(Y)} \to \max_q \;\Leftrightarrow\; \sum_Y q(Y) \ln \frac{p(X,Y|\gamma)}{q(Y)} - \sum_Y q(Y) \ln p(X|\gamma) \to \max_q$$

2. Будем минимизировать KL-дивергенцию поочерёдно по всем Y_j.
Применим факторизацию и вынесем слагаемое с $q_j(Y_j)$ вперёд:

$$\sum_{Y_j} q_j(Y_j) \underbrace{\sum_{Y \setminus Y_j} \prod_{i \neq j} q_i(Y_i) \ln p(X, Y|\gamma)}_{\mathsf{E}_{q \setminus j} \ln p(X,Y|\gamma)} - \sum_{Y_j} q_j(Y_j) \underbrace{\sum_{Y \setminus Y_j} \prod_{i \neq j} q_i(Y_i) \sum_{k \in J} \ln q_k(Y_k)}_{\ln q_j(Y_j) + \text{const}} \to \max_{q_j}$$

3. Почему вторую фигурную скобку можно заменить на $\ln q_j(Y_j)$:

$$\underbrace{\sum_{Y \setminus Y_j} \prod_{i \neq j} q_i(Y_i) \sum_{k \neq j} \ln q_k(Y_k)}_{\text{не зависит от } q_j} + \underbrace{\sum_{Y \setminus Y_j} \prod_{i \neq j} q_i(Y_i)}_{1} \ln q_j(Y_j)$$

4. Введём $r(Y_j) \propto \exp\big(\mathsf{E}_{q \setminus j} \ln p(X, Y|\gamma)\big)$, тогда $\mathrm{KL}\big(q_j(Y_j) \big\| r(Y_j)\big) \to \min_{q_j}$
5. Точное решение данной задачи $q_j(Y_j) = r(Y_j)$, следовательно,

$$\ln q_j(Y_j) = \mathsf{E}_{q \setminus j} \ln p(X, Y|\gamma) + \text{const}.$$

■

Вариационный байесовский вывод для модели LDA

Обозначим $Y = (Z, \Phi, \Theta)$, $\gamma = (\beta, \alpha)$, $J = \{1, \ldots, n\} \sqcup T \sqcup D$:

$$\ln q_j = \mathsf{E}_{q \setminus j} \ln p(X, Z, \Phi, \Theta | \beta, \alpha) + \text{const}$$

Нам предстоит брать матожидания $\mathsf{E}_{q \setminus j}$ по всем (кроме одного) распределениям $q_t(\phi_t)$, $q_d(\theta_d)$, $q_i(t_i)$ от

$$\ln p(X, Z, \Phi, \Theta | \beta, \alpha) = \ln p(X, Z | \Phi, \Theta)\, p(\Phi | \beta)\, p(\Theta | \alpha) =$$

$$= \ln \prod_{i=1}^{n} p(d_i, w_i, t_i | \Phi, \Theta) + \ln \prod_{t \in T} \mathrm{Dir}(\phi_t | \beta) + \ln \prod_{d \in D} \mathrm{Dir}(\theta_d | \alpha) =$$

$$= \sum_{i=1}^{n} \ln \phi_{w_i t_i} \theta_{t_i d_i} + \sum_{t,w} (\beta_w - 1) \ln \phi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td} + \text{const}.$$

Замечание, сильно упрощающее выкладки:
если слагаемое S не зависит от j-й переменной, то $\mathsf{E}_{q_j} S = \text{const}$.

Распределения для блока переменных $q_t(\phi_t)$

Уравнение для распределения переменной $\phi_t \in \mathbb{R}^W$:

$$\begin{aligned}
\ln q_t(\phi_t) &= \sum_{i=1}^{n} \mathsf{E}_{q_i(t_i)}[t_i = t] \ln \phi_{w_i t_i} + \sum_{w \in W} (\beta_w - 1) \ln \phi_{wt} + \text{const} = \\
&= \sum_{i=1}^{n} \sum_{w \in W} [w_i = w] q_i(t) \ln \phi_{wt} + \sum_{w \in W} (\beta_w - 1) \ln \phi_{wt} + \text{const} = \\
&= \sum_{w \in W} \Bigg(\underbrace{\sum_{i=1}^{n} [w_i = w] q_i(t)}_{n_{wt}} + \beta_w - 1 \Bigg) \ln \phi_{wt} + \text{const} = \\
&= \ln \mathrm{Dir}(\phi_t | \tilde{\beta}_t).
\end{aligned}$$

Это распределение Дирихле с параметрами $\tilde{\beta}_{wt} = n_{wt} + \beta_w$,
n_{wt} — оценка числа генераций терма w из темы t.
При больших n_{wt} оно сконцентрировано в точке $\phi_{wt} = \underset{w}{\mathrm{norm}}(\tilde{\beta}_{wt})$.

Распределения для блока переменных $q_d(\theta_d)$

Уравнение для распределения переменной $\theta_d \in \mathbb{R}^T$:

$$\begin{aligned}
\ln q_d(\theta_d) &= \sum_{i=1}^{n} \mathrm{E}_{q_i(t_i)}[d_i = d] \ln \theta_{t_i d_i} + \sum_{t \in T} (\alpha_t - 1) \ln \theta_{td} + \text{const} = \\
&= \sum_{i=1}^{n} [d_i = d] \sum_{t \in T} q_i(t) \ln \theta_{td} + \sum_{t \in T} (\alpha_t - 1) \ln \theta_{td} + \text{const} = \\
&= \sum_{t \in T} \Bigg(\underbrace{\sum_{i=1}^{n} [d_i = d] q_i(t)}_{n_{td}} + \alpha_t - 1 \Bigg) \ln \theta_{td} + \text{const} = \\
&= \ln \mathrm{Dir}(\theta_d | \tilde{\alpha}_d).
\end{aligned}$$

Это распределение Дирихле с параметрами $\tilde{\alpha}_{td} = n_{td} + \alpha_t$,
n_{td} — оценка числа термов темы t в документе d.
При больших n_{td} оно сконцентрировано в точке $\theta_{td} = \underset{t}{\mathrm{norm}}(\tilde{\alpha}_{td})$.

Распределения для блока переменных $q_i(t_i)$

Уравнение для распределения переменной $t_i \in T$:

$$\ln q_i(t) = \mathsf{E}_{q \setminus i}\big(\ln \phi_{w_i t_i} + \ln \theta_{t_i d_i}\big) + \text{const} =$$
$$= \mathsf{E}_{q_t(\phi_t)} \ln \phi_{w_i t} + \mathsf{E}_{q_d(\theta_d)} \ln \theta_{t_i d} + \text{const} =$$

воспользуемся тем, что $q_t(\phi_t)$ и $q_d(\theta_d)$ уже найдены:

$$= \psi\big(n_{w_i t} + \beta_{w_i}\big) - \psi\Big(\sum_w (n_{wt} + \beta_w)\Big) +$$
$$+ \psi\big(n_{td_i} + \alpha_t\big) - \psi\Big(\sum_t (n_{td_i} + \alpha_t)\Big) + \text{const}$$

Воспользуемся приближением $\exp(\psi(x)) \approx x - \frac{1}{2}$:

$$q_i(t) = \underset{t \in T}{\text{norm}} \left(\frac{n_{w_i t} + \beta_{w_i} - \frac{1}{2}}{\sum_w (n_{wt} + \beta_w) - \frac{1}{2}} \cdot \frac{n_{td_i} + \alpha_t - \frac{1}{2}}{\sum_t (n_{td_i} + \alpha_t) - \frac{1}{2}} \right)$$

Похоже на обычную формулу E-шага $p(t|d_i, w_i) = \underset{t \in T}{\text{norm}}(\phi_{w_i t} \theta_{t d_i})$

Собираем всё воедино

В итерационном процессе чередуются два шага:

1) распределение термов (d_i, w_i) по темам, $E(x) = \exp(\psi(x))$:

$$q_i(t) = \underset{t \in T}{\text{norm}} \left(\frac{E\left(n_{w_i t} + \beta_{w_i}\right)}{E\left(\sum_w (n_{wt} + \beta_w)\right)} \cdot \frac{E\left(n_{td_i} + \alpha_t\right)}{E\left(\sum_t (n_{td_i} + \alpha_t)\right)} \right)$$

2) аккумулирование счётчиков n_{wt} и n_{td}:

$$n_{wt} = \sum_{i=1}^{n} [w_i = w]\, q_i(t) \qquad n_{td} = \sum_{i=1}^{n} [d_i = d]\, q_i(t)$$

Точечные оценки параметров по матожиданию или моде:

$$\mathrm{E}\phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \beta_w\right) \qquad \mathrm{E}\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \alpha_t\right)$$

$$\hat{\phi}_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \beta_w - 1\right) \qquad \hat{\theta}_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \alpha_t - 1\right)$$

Резюме по вариационному байесовскому выводу

- Из-за факторизации вариационный байесовский вывод даёт лишь приближённое решение, тем не менее,
- формулы для MAP и VB очень похожи [Asuncion]:
  — при $n_{wt}, n_{td} \gg 1$ различия неощутимы,
  — при $n_{wt}, n_{td} \lesssim 1$ тема t незначима для w или d.
- Можно добавить M-шаг для оптимизации β, α [Wallach].
- Некуда добавлять регуляризаторы $R(\Phi, \Theta)$.
- Нужны матрицы Φ, Θ, а не распределения $p(\Phi, \Theta|X)$.
- Начинает смущать разнообразие оценок... какая лучше?

Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models. Int'l conf. on Uncertainty in Artificial Intelligence, 2009.

Hanna Wallach, David Mimno, Andrew McCallum. Rethinking LDA: why priors matter. Neural Information Processing Systems, 2009.

Сэмплирование Гиббса (Gibbs Sampling) для модели LDA

Основная идея:

- $Z \sim p(Z|X,\gamma)$ — сэмплировать скрытые переменные
- $p(\Phi,\Theta|X,Z,\gamma)$ — найти апостериорное распределение параметров модели при известных X, Z и $\gamma = (\beta,\alpha)$

Основная теорема о сходимости сэмплирования Гиббса

Процесс сэмплирования одномерных случайных величин

$$t_i^{(k+1)} \sim p(t_i|X,Z_{\setminus i},\gamma) = \frac{p(X,Z|\gamma)}{p(X,Z_{\setminus i}|\gamma)},$$

где k — номер итерации, $Z_{\setminus i} = (t_1^{(k+1)},\ldots,t_{i-1}^{(k+1)},t_{i+1}^{(k)},\ldots,t_n^{(k)})$, сходится к многомерному распределению $Z \sim p(Z|X,\gamma)$

Griffiths T., Steyvers M. Finding scientific topics. 2004.

Распределение Дирихле — сопряжённое к мультиномиальному

$p(\Phi, \Theta|\beta, \alpha)$ — априорное распределение Дирихле
$p(\Phi, \Theta|X, Z, \beta, \alpha)$ — апостериорное распределение тоже Дирихле

Вывод апостериорного распределения Φ, Θ при известных X, Z:

$$p(\Phi, \Theta|X, Z, \beta, \alpha) \propto p(\Phi, \Theta, X, Z|\beta, \alpha) \propto p(X, Z|\Phi, \Theta)p(\Phi, \Theta|\beta, \alpha)$$

$$\propto \prod_{d,w,t} (\phi_{wt}\theta_{td})^{n_{dwt}} \prod_{t\in T} \mathrm{Dir}(\phi_t|\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d|\alpha)$$

$$\propto \prod_{t\in T}\prod_{d,w} \phi_{wt}^{n_{dwt}}\phi_{wt}^{\beta_w-1} \prod_{d\in D}\prod_{w,t} \theta_{td}^{n_{dwt}}\theta_{td}^{\alpha_t-1}$$

$$\propto \prod_{t\in T}\prod_{w} \phi_{wt}^{n_{wt}+\beta_w-1} \prod_{d\in D}\prod_{t} \theta_{td}^{n_{td}+\alpha_t-1}$$

$$\propto \prod_{t\in T} \mathrm{Dir}(\phi_t|\tilde{\beta}_t) \prod_{d\in D} \mathrm{Dir}(\theta_d|\tilde{\alpha}_d), \quad \tilde{\beta}_{wt} = n_{wt}+\beta_w, \quad \tilde{\alpha}_{td} = n_{td}+\alpha_t.$$

Распределение $p(X, Z|\beta, \alpha)$ для схемы сэмплирования Гиббса

Подынтегральное распределение мы только что вывели,
но теперь будем аккуратнее с нормировочными множителями:

$$p(X, Z|\beta, \alpha) = \int_\Phi \int_\Theta p(X, Z|\Phi, \Theta) p(\Phi, \Theta|\beta, \alpha) d\Phi d\Theta =$$

$$= \int_\Phi \int_\Theta \prod_{w,t} \phi_{wt}^{n_{wt}} \prod_{t,d} \theta_{td}^{n_{td}} \prod_d p_d^{n_d} \prod_{t \in T} \mathrm{Dir}(\phi_t|\beta) \prod_{d \in D} \mathrm{Dir}(\theta_d|\alpha) d\Phi d\Theta =$$

$$= \prod_{t \in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \int_{\phi_t} \underbrace{\prod_w \phi_{wt}^{\tilde{\beta}_{wt}-1} d\phi_t}_{\propto \mathrm{Dir}(\phi_t|\tilde{\beta}_t)} \prod_{d \in D} p_d^{n_d} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \int_{\theta_d} \underbrace{\prod_t \theta_{td}^{\tilde{\alpha}_{td}-1} d\theta_d}_{\propto \mathrm{Dir}(\theta_d|\tilde{\alpha}_d)} =$$

$$= \prod_{t \in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{\prod_w \Gamma(\tilde{\beta}_{wt})}{\Gamma(\sum_w \tilde{\beta}_{wt})} \prod_{d \in D} p_d^{n_d} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \frac{\prod_t \Gamma(\tilde{\alpha}_{td})}{\Gamma(\sum_t \tilde{\alpha}_{td})}$$

Распределение $p(X, Z_{\setminus i}|\beta, \alpha)$ для схемы сэмплирования Гиббса

Итак, мы только что получили распределение

$$p(X, Z|\beta, \alpha) =$$
$$= \prod_{t\in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{\prod_w \Gamma(\tilde{\beta}_{wt})}{\Gamma(\sum_w \tilde{\beta}_{wt})} \prod_{d\in D} p_d^{n_d} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \frac{\prod_t \Gamma(\tilde{\alpha}_{td})}{\Gamma(\sum_t \tilde{\alpha}_{td})}$$

Распределение $p(X, Z_{\setminus i}|\beta, \alpha)$ отличается от него лишь тем, что оно построено по выборке без одной i-й точки (d_i, w_i, t_i):

$$p(X, Z_{\setminus i}|\beta, \alpha) =$$
$$\prod_{t\in T} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{\prod_w \Gamma(\tilde{\beta}_{wt} - \delta^i_{wt})}{\Gamma(\sum_w (\tilde{\beta}_{wt} - \delta^i_{wt}))} \prod_{d\in D} p_d^{n_d} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \frac{\prod_t \Gamma(\tilde{\alpha}_{td} - \delta^i_{td})}{\Gamma(\sum_t (\tilde{\alpha}_{td} - \delta^i_{td}))}$$

где $\delta^i_{wt} = [w = w_i][t = t_i]$, $\delta^i_{td} = [t = t_i][d = d_i]$

Ещё чуть-чуть... осталось поделить одно на другое

Для сэмплирования Гиббса нужно одномерное распределение

$$p(t_i|X, Z_{\setminus i}, \beta, \alpha) = \frac{p(X, Z|\beta, \alpha)}{p(X, Z_{\setminus i}|\beta, \alpha)} =$$

В числителе и знаменателе сократятся все множители кроме i-х:

$$= \frac{\Gamma(n_{w_i t_i} + \beta_{w_i})\,\Gamma\big(\sum_w (n_{wt_i} + \beta_w) - 1\big)}{\Gamma(n_{w_i t_i} + \beta_{w_i} - 1)\,\Gamma\big(\sum_w (n_{wt_i} + \beta_w)\big)} \frac{\Gamma(n_{t_i d_i} + \alpha_{t_i})\,\Gamma\big(\sum_t (n_{td_i} + \alpha_t) - 1\big)}{\Gamma(n_{t_i d_i} + \alpha_{t_i} - 1)\,\Gamma\big(\sum_t (n_{td_i} + \alpha_t)\big)}$$

Воспользуемся свойством гамма-функции $\frac{\Gamma(x)}{\Gamma(x-1)} = x - 1$:

$$p(t|X, Z_{\setminus i}, \beta, \alpha) = \underset{t \in T}{\text{norm}} \left(\frac{n_{w_i t} + \beta_{w_i} - 1}{\sum_w (n_{wt} + \beta_w) - 1} \cdot \frac{n_{td_i} + \alpha_t - 1}{\sum_t (n_{td_i} + \alpha_t) - 1} \right)$$

Похоже на обычную формулу E-шага $p(t|d_i, w_i) = \underset{t \in T}{\text{norm}}(\phi_{w_i t} \theta_{t d_i})$

Собираем всё воедино

Выделены отличия от вариационного алгоритма

1) для каждого (d_i, w_i), $i = 1, \ldots, n$, сэмплирование темы t_i:

$$t_i \sim p_i(t) = \underset{t \in T}{\text{norm}} \left(\frac{n_{w_i t} + \beta_{w_i} - 1}{\sum_w (n_{wt} + \beta_w) - 1} \cdot \frac{n_{td_i} + \alpha_t - 1}{\sum_t (n_{td_i} + \alpha_t) - 1} \right)$$

2) аккумулирование счётчиков n_{wt} и n_{td}:

$$n_{wt} = \sum_{i=1}^{n} [w_i = w][t_i = t] \qquad n_{td} = \sum_{i=1}^{n} [d_i = d][t_i = t]$$

Точечные оценки параметров по матожиданию или моде:

$$\mathrm{E}\phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \beta_w\right) \qquad \mathrm{E}\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \alpha_t\right)$$

$$\hat{\phi}_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \beta_w - 1\right) \qquad \hat{\theta}_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \alpha_t - 1\right)$$

Резюме по GS и методам байесовского вывода

- Это тоже «EM-подобный алгоритм», если на E-шаге вместо $p(t|d,w)$ взять $\hat{p}(t|d,w) = [t = t_i]$, $t_i \sim p(t|d,w)$.
- Формулы для MAP, VB и GS очень похожи [Asuncion]:
  — при $n_{wt}, n_{td} \gg 1$ различия неощутимы,
  — при $n_{wt}, n_{td} \lesssim 1$ тема t незначима для w или d.
- Необходимость задания априорных распределений:
  — сопряжённые — только распределения Дирихле,
  — не сопряжённые — сильно усложняют задачу.
- VB и GS не имеют удобных механизмов регуляризации, т.к. нет, собственно, и задачи оптимизации по (Φ, Θ)
- Проблема неустойчивости решения даже не ставится.

Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models. Int'l conf. on Uncertainty in Artificial Intelligence, 2009.

Некорректные задачи и классическая регуляризация

Задача *корректно поставлена по Адамару*, если её решение

- существует,
- единственно,
- устойчиво.

Жак Саломон Адамар
(1865–1963)

Наша задача матричного разложения *некорректно поставлена*:
если Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta)$, $\operatorname{rank} S = |T|$
- $L(\Phi', \Theta') = L(\Phi, \Theta)$
- $L(\Phi', \Theta') \leqslant L(\Phi, \Theta) + \varepsilon$ — приближённые решения

Регуляризация — классический приём доопределения решения с помощью дополнительных критериев.

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризаторами:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \qquad R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta)$$

EM-алгоритм — метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdw} = p(t|d,w)$:

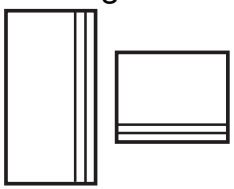
$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}} \right), & n_{wt} = \sum\limits_{d \in D} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}} \right), & n_{td} = \sum\limits_{w \in d} n_{dw} p_{tdw} \end{cases}$$

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014.

Регуляризаторы для улучшения интерпретируемости тем

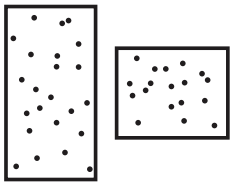
background



Сглаживание фоновых тем $B \subset T$:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in B}\sum_{w} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d}\sum_{t\in B} \alpha_t \ln \theta_{td}$$

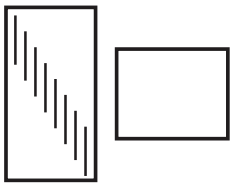
sparse



Разреживание предметных тем $S = T \setminus B$:

$$R(\Phi,\Theta) = -\beta_0 \sum_{t\in S}\sum_{w} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d}\sum_{t\in S} \alpha_t \ln \theta_{td}$$

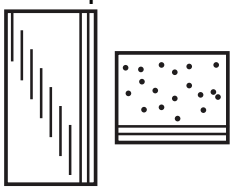
decorrelated



Декоррелирование для повышения различности тем:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s}\sum_{w} \phi_{wt}\phi_{ws}$$

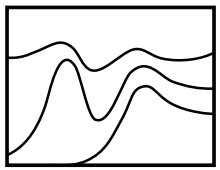
interpretable



Сглаживание + разреживание + декоррелирование для улучшения интерпретируемости тем

Регуляризаторы для учёта дополнительной информации

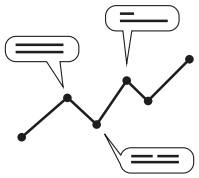
temporal



Темпоральные модели с модальностью времени i:

$$R(\Phi) = -\tau \sum_{i \in I} \sum_{t \in T} \left| \phi_{it} - \phi_{i-1,t} \right|$$

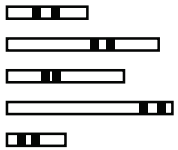
regression



Линейная модель регрессии $\hat{y}_d = \langle v, \theta_d \rangle$ документов:

$$R(\Theta, v) = -\tau \sum_{d \in D} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right)^2$$

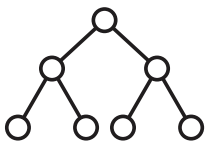
coherence



Модели сочетаемости слов (n_{uv} — частота битерма):

$$R(\Phi) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{vt}$$

hierarchy

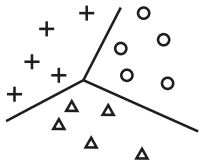


Связь родительских тем t с дочерними подтемами s:

$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \phi_{ws} \psi_{st}$$

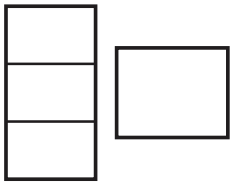
Регуляризаторы для мультимодальных тематических моделей

supervised



Модальности меток классов, категорий или тегов для классификации/категоризации/тегирования текстов

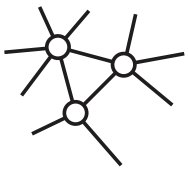
multilanguage



Модальность языков и регуляризация со словарём $\pi_{uwt} = p(u|w,t)$ переводов с языка k на ℓ:

$$R(\Phi,\Pi) = \tau \sum_{u\in W^k} \sum_{t\in T} n_{ut} \ln \sum_{w\in W^\ell} \pi_{uwt}\phi_{wt}$$

graph



Модальность вершин графа v, содержащих D_v:

$$R(\Phi) = -\frac{\tau}{2} \sum_{(u,v)\in E} S_{uv} \sum_{t\in T} n_t^2 \left(\frac{\phi_{vt}}{|D_v|} - \frac{\phi_{ut}}{|D_u|}\right)^2.$$

geospatial

Модальность геолокаций g с близостью $S_{gg'}$:

$$R(\Phi) = -\frac{\tau}{2} \sum_{g,g'\in G} S_{gg'} \sum_{t\in T} n_t^2 \left(\frac{\phi_{gt}}{n_g} - \frac{\phi_{g't}}{n_{g'}}\right)^2$$

Комбинирование регуляризаторов в прикладных задачах

Выявления этнорелевантного дискурса в социальных сетях:

$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{interpretable}\right) + R\left(\text{n-gram}\right) + R\left(\text{seed words}\right) \to \max$$

Тематический поиск научных и научно-популярных статей:

$$\mathscr{L}\left(\text{multimodal}\right) + R\left(\text{interpretable}\right) + R\left(\text{n-gram}\right) + R\left(\text{hierarchy}\right) \to \max$$

Выявление и прослеживание событий в новостном потоке:

$$\mathscr{L}\left(\text{multimodal}\right) + R\left(\text{interpretable}\right) + R\left(\text{temporal}\right) + R\left(\text{sentiment}\right) \to \max$$

Apishev M. et al. Mining ethnic content online with additively regularized topic models, 2016.
Ianina A., Vorontsov K. Hierarchical interpretable topical embeddings for exploratory search and real-time document tracking, 2020.
Feldman D., Sadekova T., Vorontsov K. Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining. 2020

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Встроенная библиотека регуляризаторов и мер качества
- Большие данные: коллекция не хранится в памяти
- Самый быстрый онлайновый параллельный ARTM

Сообщество:

- Открытый код https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Документация http://bigartm.org

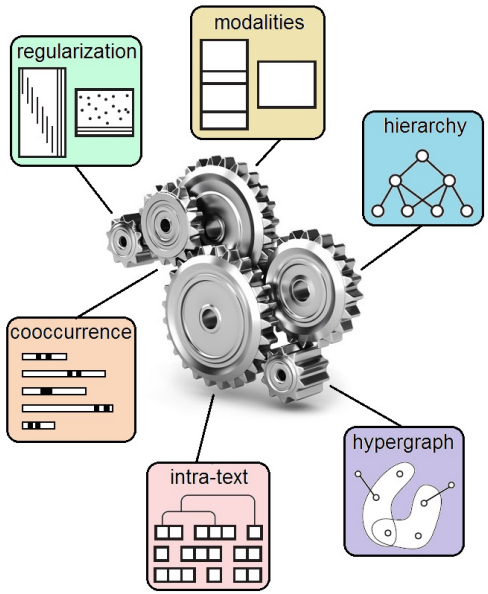
Лицензия и среда разработки:

- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Linux, MacOS, Windows (32/64 bit)
- Интерфейсы API: C++, Python, командная строка

Ключевые возможности библиотек BigARTM и TopicNet

BigARTM (с 2014 г.)

- библиотека регуляризаторов
- мультимодальные модели
- иерархические модели
- гиперграфовые модели
- модели связности текста



TopicNet (с 2020 г.)

- Перебор сценариев регуляризации для выбора моделей
- Автоматическое протоколирование экспериментов
- Построение «банка тем» из множества моделей
- Визуализация тематических моделей

V.Bulatov, E.Egorov, E.Veselova, D.Polyudova, V.Alekseev, A.Goncharov, K.Vorontsov.
TopicNet: making additive regularization for topic modelling accessible. LREC-2020

Преимущества ARTM

- Модели, ранее предложенные в рамках байесовского подхода, допускают переформулировку в терминах ARTM
- Вывод EM-алгоритма простой и общий для всех моделей: достаточно найти частные производные $\frac{\partial R}{\partial \phi_{wt}}$, $\frac{\partial R}{\partial \theta_{td}}$
- Свойство аддитивности регуляризаторов позволяет комбинировать модели в любых сочетаниях
- Распределение Дирихле перестаёт играть «особую роль» и переходит в регуляризатор сглаживания/разреживания

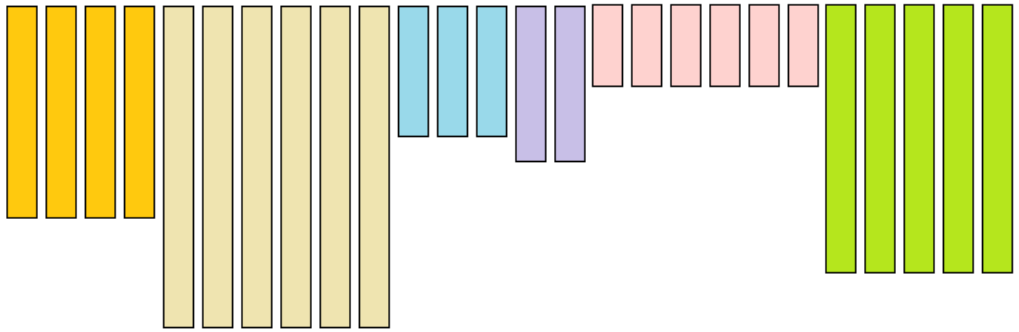
Более того (см. далее),

- Итерационный процесс и доказательство его сходимости обобщается на более широкий класс задач
- Теория вероятностного тематического моделирования упрощается до нескольких основных теорем

Задача максимизации функции на единичных симплексах

Пусть $\Omega = (\omega_j)_{j \in J}$ — набор нормированных неотрицательных векторов $\omega_j = (\omega_{ij})_{i \in I_j}$ различных размерностей $|I_j|$:

$$\Omega = \left(\quad \right)$$



Задача максимизации функции $f(\Omega)$ на единичных симплексах:

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad j \in J; \\ \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J. \end{cases}$$

Теорема о максимизации функции на единичных симплексах

Операция нормировки вектора: $p_i = \underset{i \in I}{\operatorname{norm}}(x_i) = \dfrac{\max\{x_i, 0\}}{\sum\limits_{k \in I} \max\{x_k, 0\}}$

Теорема. Пусть $f(\Omega)$ непрерывно дифференцируема по Ω. Тогда векторы ω_j локального экстремума задачи $f(\Omega) \to \max$ удовлетворяют системе уравнений

$$\omega_{ij} = \underset{i \in I_j}{\operatorname{norm}}\left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}}\right), \quad \text{если } \exists i : \omega_{ij} \frac{\partial f}{\partial \omega_{ij}} > 0$$

$$\omega_{ij} = \underset{i \in I_j}{\operatorname{norm}}\left(-\omega_{ij} \frac{\partial f}{\partial \omega_{ij}}\right), \quad \text{иначе, если } \exists i : \omega_{ij} \frac{\partial f}{\partial \omega_{ij}} < 0$$

$$\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} = 0, \quad \text{иначе}$$

Замечания к теореме о максимизации на симплексах

- Теорема применима для широкого класса моделей, параметрами которых являются дискретные распределения вероятности (нормированные неотрицательные векторы)
- Численное решение системы — методом простых итераций
- Существование стационарной точки Ω гарантировано
- Первый из трёх случаев является основным:

$$\omega_{ij} := \underset{i \in I_j}{\text{norm}} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right)$$

- В остальных случаях нормирующий знаменатель нулевой; такие векторы будем удалять из модели как вырожденные
- Итерации похожи на градиентную оптимизацию, но учитывают ограничения и не требуют подбора шага η:

$$\omega_{ij} := \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}}$$

Доказательство Теоремы

Запишем условия Каруша–Куна–Таккера для $\omega_j = (\omega_{ij} : i \in I_j)$:

$$\frac{\partial f}{\partial \omega_{ij}} = \lambda_j - \mu_{ij}; \quad \mu_{ij}\omega_{ij} = 0.$$

Предполагая $\omega_{ij} > 0$, умножим обе части равенства на ω_{ij}:

$$A_{ij} \equiv \omega_{ij}\frac{\partial f}{\partial \omega_{ij}} = \omega_{ij}\lambda_j.$$

Возможны три случая:

- Если $\lambda_j > 0$, то либо $A_{ij} > 0$, либо $\omega_{ij} = 0$. Тогда $\omega_{ij}\lambda_j = (A_{ij})_+$; $\lambda_j = \sum_i (A_{ij})_+ \;\Rightarrow\; \omega_{ij} = \operatorname{norm}_i(A_{ij})$.
- Если $\lambda_j < 0$ и $(\exists i)\ A_{ij} < 0$, то $(\forall i)\ A_{ij} \leqslant 0$. Тогда $\omega_{ij}\lambda_j = -(-A_{ij})_+$; $\lambda_j = -\sum_i (-A_{ij})_+ \;\Rightarrow\; \omega_{ij} = \operatorname{norm}_i(-A_{ij})$.
- Иначе $\lambda_j = 0$ и ω_j находится из уравнений $\omega_{ij}\frac{\partial f}{\partial \omega_{ij}} = 0$.

■

Теорема о сходимости

Рассматривается итерационный процесс

$$\omega_{ij}^{t+1} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij}^{t} \frac{\partial f(\Omega^t)}{\partial \omega_{ij}^{t}} \right)$$

Теорема. Пусть $f(\Omega)$ — ограниченная сверху, непрерывно дифференцируемая функция, и все Ω^t, начиная с некоторой итерации t^0 обладают свойствами:

- $\forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^{t} = 0 \to \omega_{ij}^{t+1} = 0$ (сохранение нулей)
- $\exists \varepsilon > 0 \;\; \forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^{t} \notin (0, \varepsilon)$ (отделимость от нуля)
- $\exists \delta > 0 \;\; \forall j \in J \;\; \exists i \in I_j \;\; \omega_{ij}^{t} \frac{\partial f(\Omega^t)}{\partial \omega_{ij}} \geqslant \delta$ (невырожденность)

Тогда $f(\Omega^{t+1}) > f(\Omega^t)$ и $\left|\omega_{ij}^{t+1} - \omega_{ij}^{t}\right| \to 0$ при $t \to \infty$.

Теперь вывод EM-алгоритма для ARTM — в две строки!

Применим Теорему к log-правдоподобию с регуляризатором R:

$$f(\Phi,\Theta) = \sum_{d,w} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

$$\phi_{wt} = \operatorname*{norm}_{w\in W}\left(\phi_{wt}\frac{\partial f}{\partial \phi_{wt}}\right) = \operatorname*{norm}_{w\in W}\left(\phi_{wt}\sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) =$$

$$= \operatorname*{norm}_{w\in W}\left(\sum_{d\in D} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right);$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\left(\theta_{td}\frac{\partial f}{\partial \theta_{td}}\right) = \operatorname*{norm}_{t\in T}\left(\theta_{td}\sum_{w\in W} n_{dw}\frac{\phi_{wt}}{p(w|d)} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) =$$

$$= \operatorname*{norm}_{t\in T}\left(\sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right).$$

E-шаг — это вычисление вспомогательных переменных p_{tdw}.

Условия вырожденности модели для тем и документов

Решение может быть вырожденным для некоторых тем (столбцов матриц Φ) и документов (столбцов матрицы Θ).

Тема t вырождена, если для всех термов $w \in W$

$$n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \leqslant 0.$$

Если тема t вырождена, то $p(w|t) = \phi_{wt} \equiv 0$; это означает, что тема исключается из модели (происходит отбор тем).

Документ d вырожден, если для всех тем $t \in T$

$$n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \leqslant 0.$$

Если документ d вырожден, то $p(t|d) = \theta_{td} \equiv 0$; это означает, что модель не в состоянии описать данный документ.

Общий взгляд на байесовское обучение, MAP и ARTM

Байесовский вывод апостериорного распределения $p(\Omega|X)$ (обычно приближённый) ради получения точечной оценки Ω:

$$\text{Posterior}(\Omega|X,\gamma) \;\propto\; p(X|\Omega)\,\text{Prior}(\Omega|\gamma)$$

$$\Omega := \arg\max_{\Omega} \text{Posterior}(\Omega|X,\gamma)$$

Максимизация апостериорной вероятности (MAP) даёт точечную оценку Ω напрямую, без вывода Posterior:

$$\Omega := \arg\max_{\Omega}\Big(\ln p(X|\Omega) + \underbrace{\ln \text{Prior}(\Omega|\gamma)}_{R(\Omega)}\Big)$$

Многокритериальная аддитивная регуляризация (ARTM) обобщает MAP на любые регуляризаторы и их комбинации:

$$\Omega := \arg\max_{\Omega}\Big(\ln p(X|\Omega) + \sum_{i=1} \tau_i R_i(\Omega)\Big)$$

Модульный подход ARTM: сравнение с байесовским подходом

Для построения композитных моделей в ARTM не нужны ни математические выкладки, ни программирование «с нуля».

Этапы моделирования	Bayesian TM	ARTM	
	Анализ требований	Анализ требований	
Формализация:	Вероятностная модель порождения данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Единый регуляризованный EM-алгоритм для любых моделей и их композиций	
Реализация:	Исследовательский код (Matlab, Python, R)	Промышленный код BigARTM (C++, Python API)	
Оценивание:	Исследовательские метрики, исследовательский код	Стандартные метрики	Свои метрики
	Внедрение	Внедрение	

-- нестандартизуемые этапы, уникальная разработка для каждой задачи

-- стандартизуемые этапы

Резюме

«Представляется важной задача освобождения всюду, где это возможно, от излишних вероятностных допущений»

[А. Н. Колмогоров, 1987]

В результате такого отказа в тематическом моделировании мы

- **потеряли:**
  - — возможность оценивания апостериорных распределений (которая практически никогда и не использовалась)
- **приобрели:**
  - — более общие условия сходимости
  - — более удобную формализацию широкого класса моделей
  - — возможность комбинирования моделей (ARTM)
  - — возможность модульной реализации (BigARTM)

В теории ВТМ переход к байесовской регуляризации минуя классическую переусложнил технику и (на годы) закрыл возможности комбинирования моделей

Воронцов К.В. Вероятностное тематическое моделирование. 2020, http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf