Вероятностные тематические модели
Лекция 5.
Классика тематических моделей: PLSA, LDA и EM-алгоритм

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

ВМК МГУ • 14 марта 2019

Содержание

Напоминание. Задача тематического моделирования

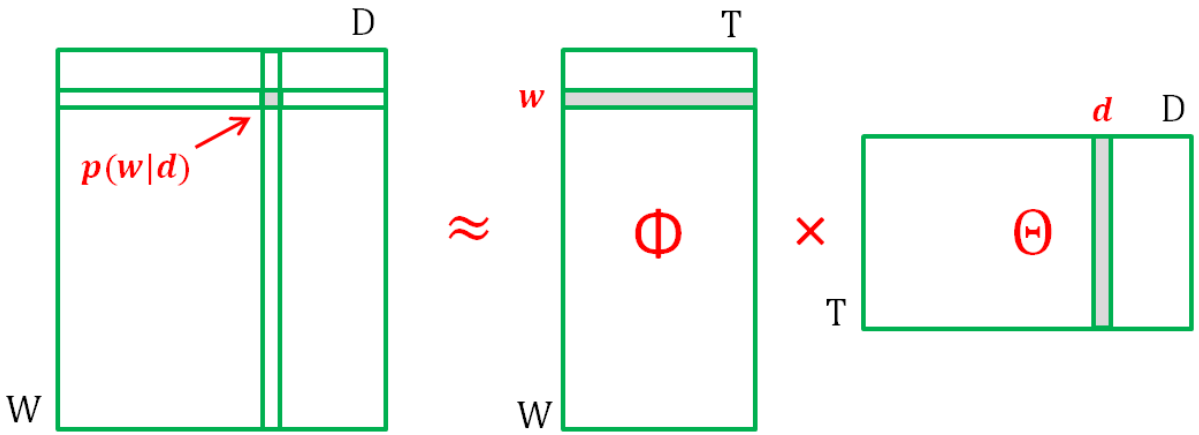
Дано: коллекция текстовых документов

- n_{dw} — частоты терминов 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 терминов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Это задача стохастического матричного разложения:



Напоминание. PLSA (Probabilistic Latent Semantic Analysis)

Задача максимизации логарифма правдоподобия:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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.

Недостатки PLSA (и необходимость его регуляризации)

1. Большая размерность пространства параметров
2. Якобы из-за этого сильное переобучение
3. Якобы невозможность моделирования новых документов
4. Неединственность и неустойчивость решения:
   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 выделения нетематических (фоновых) слов
8. Не ясно, как учитывать дополнительную информацию

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

Гипотеза об априорных распределениях Дирихле

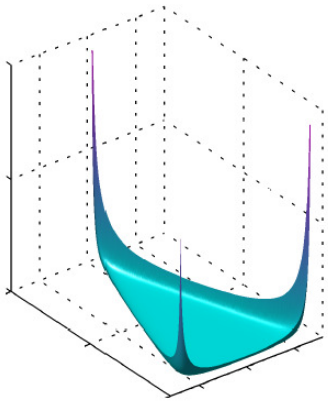
Гипотеза: вектор-столбцы $\phi_t = (\phi_{wt})_{w \in W}$ и $\theta_d = (\theta_{td})_{t \in T}$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

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

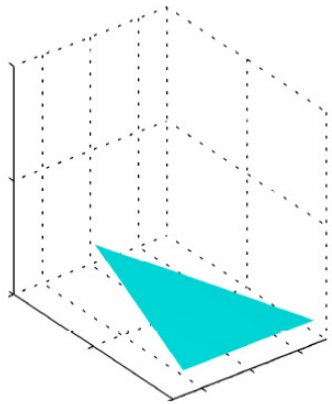
$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \theta_{td} > 0; \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t > 0;$$

Пример:

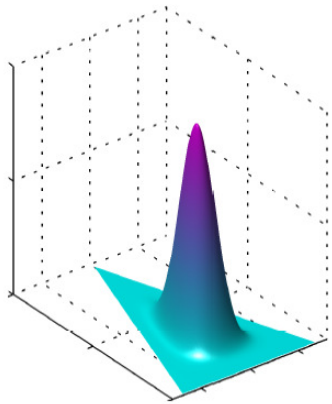
$\mathrm{Dir}(\theta|\alpha)$,
$|T| = 3$,
$\theta, \alpha \in \mathbb{R}^3$



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

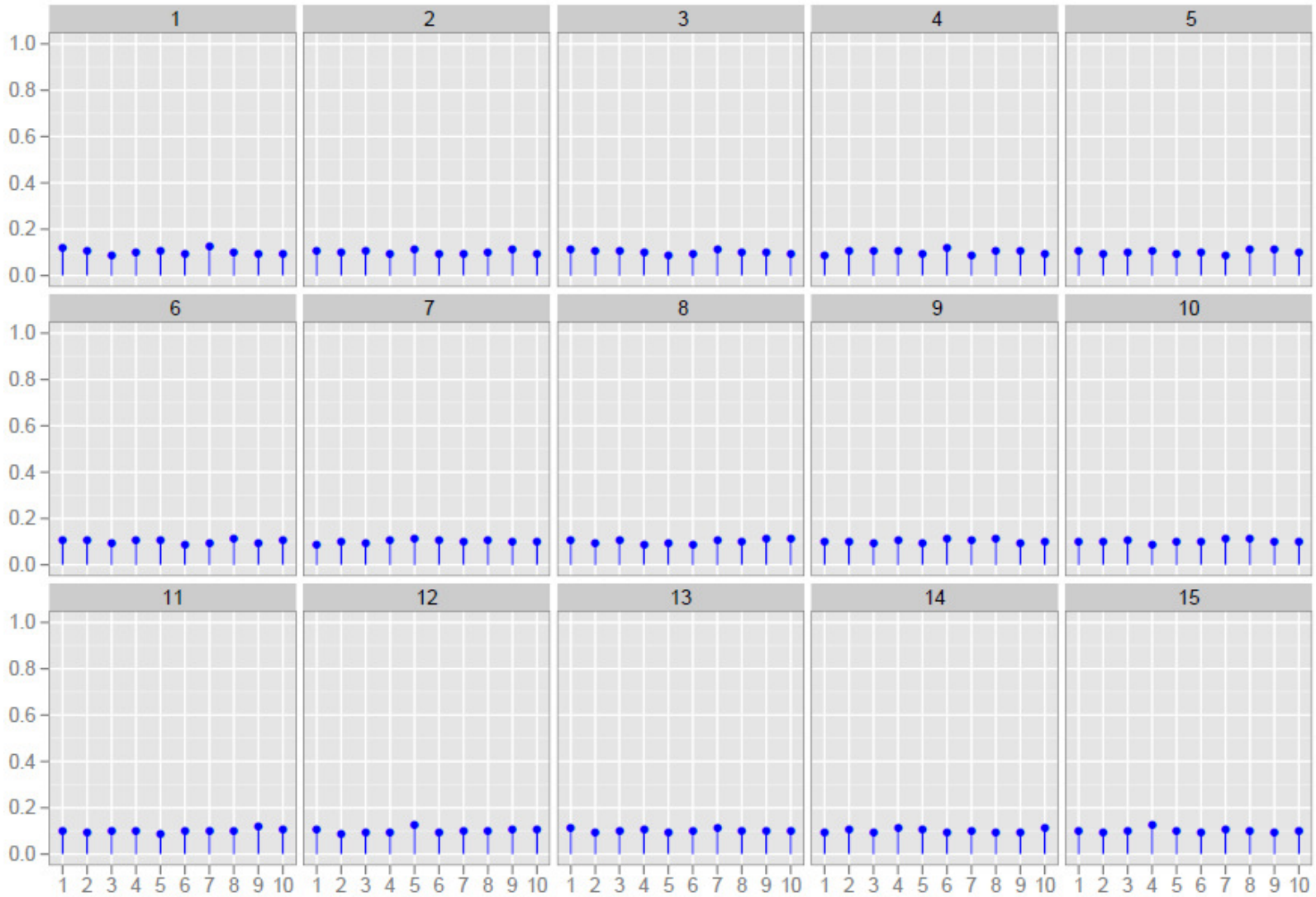


$\alpha_1 = \alpha_2 = \alpha_3 = 1$

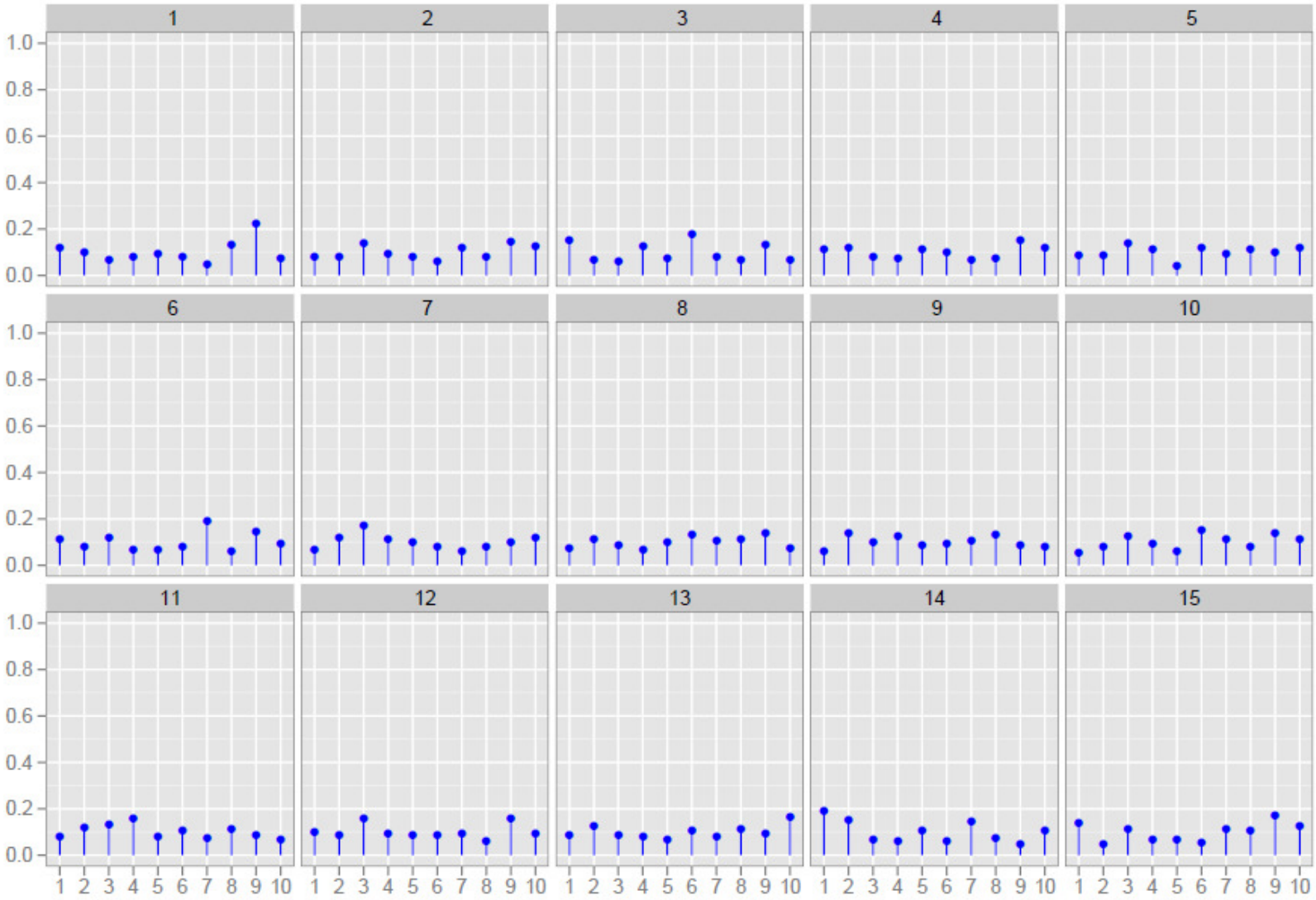


$\alpha_1 = \alpha_2 = \alpha_3 = 10$

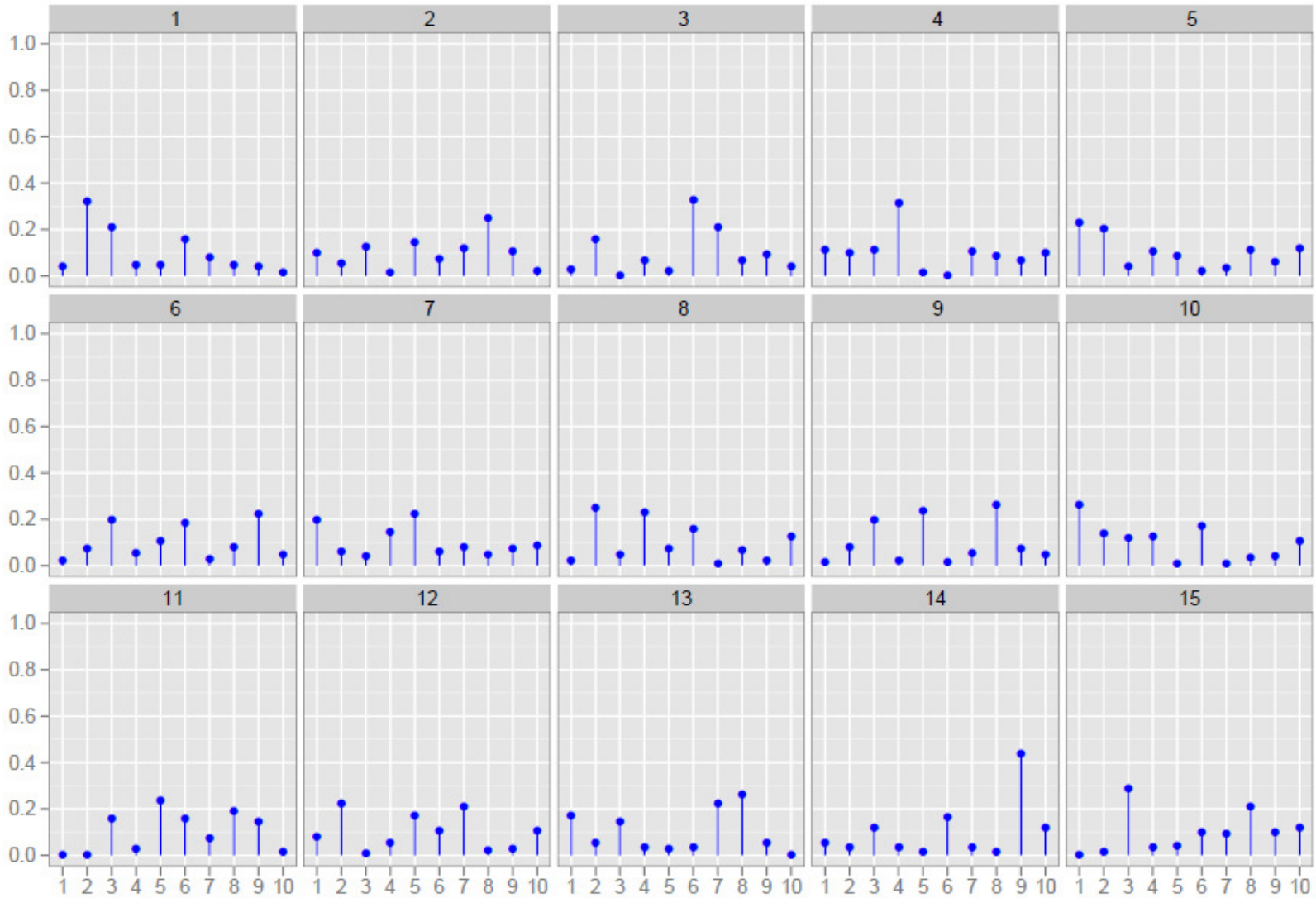
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 100$, 10 тем, 15 документов



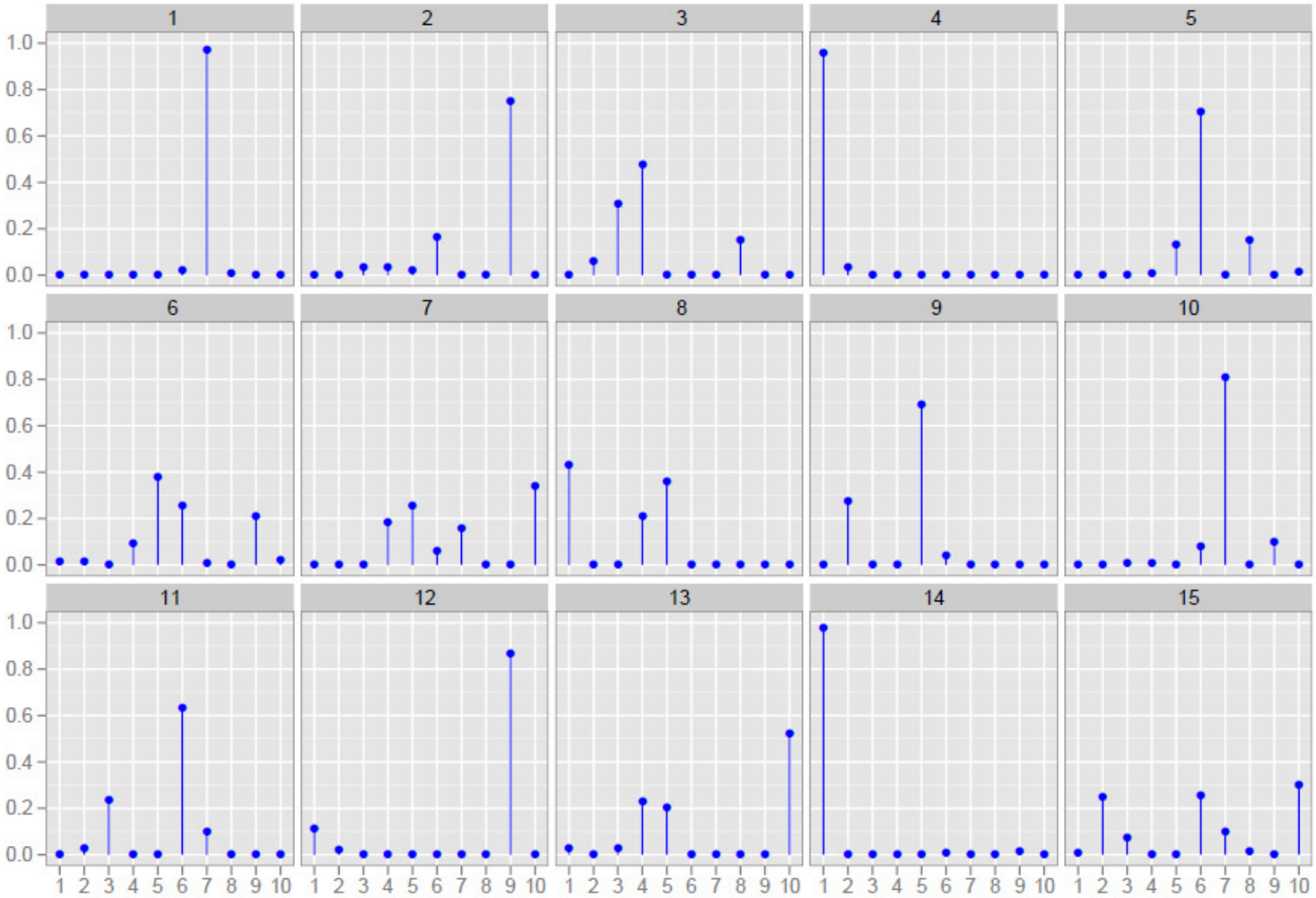
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 10$, 10 тем, 15 документов



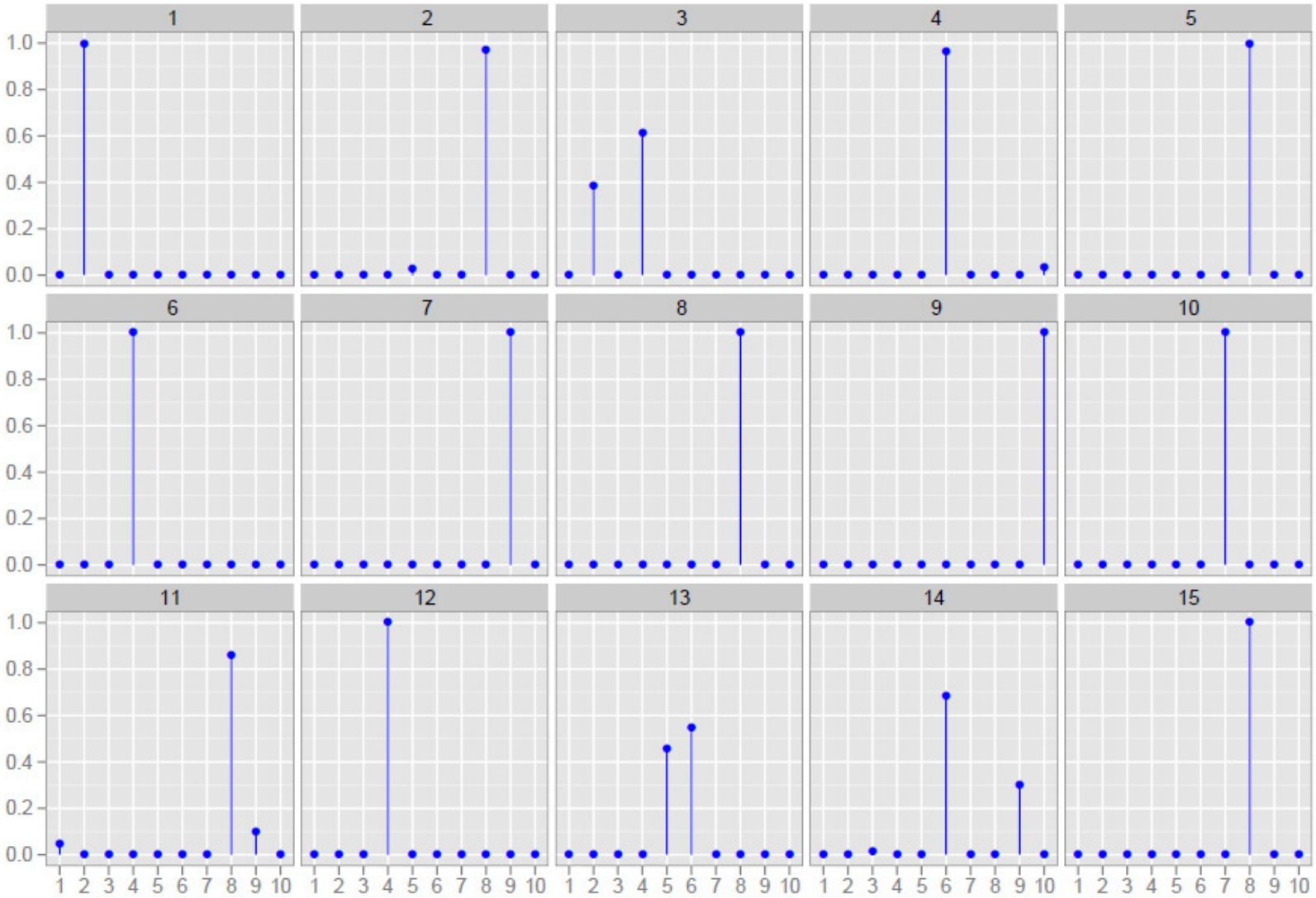
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 1$, 10 тем, 15 документов



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



Вероятностная модель порождения текста

Тематическая модель LDA (Latent Dirichlet Allocation):

$$p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}, \quad \phi_t \sim \mathrm{Dir}(\phi|\beta), \quad \theta_d \sim \mathrm{Dir}(\theta|\alpha).$$

Процесс порождения документов $d = \{w_1 \dots w_{n_d}\}$ коллекции D:

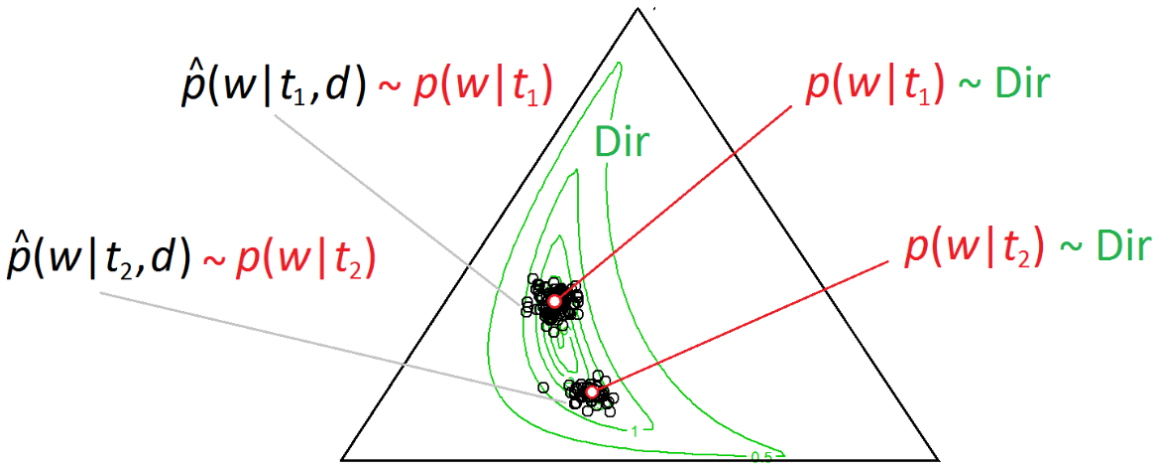
Вход: векторы гиперпараметров β, α;
Выход: коллекция документов;

выбрать вектор ϕ_t из $\mathrm{Dir}(\phi|\beta)$ для каждой темы $t \in T$;
выбрать вектор θ_d из $\mathrm{Dir}(\theta|\alpha)$ для каждого документа $d \in D$;
для всех документов $d \in D$
  для всех позиций слов $i = 1, \dots, n_d$ в документе d
    выбрать тему t_i из $p(t|d) \equiv \theta_{td}$;
    выбрать слово w_i из $p(w|t_i) \equiv \phi_{wt_i}$;

Почему именно распределение Дирихле?

- оно может 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 $\hat{p}(w|t,d)$.



Формула Байеса для апостериорного распределения

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 probability* (MAP):

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

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

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 в ARTM

Максимизация апостериорной вероятности эквивалентна регуляризатору логарифма априорного распределения:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Промежуточный итог

- LDA проще вводить через KL-дивергенцию, как регуляризатор сглаживания/разреживания
- Заодно снимаются ограничения $\beta_w > 0$, $\alpha_t > 0$

- Распределение Дирихле играет особую роль в байесовских методах тематического моделирования
- ARTM — это более простая альтернатива байесовским методам, но в статьях по тематическому моделированию они преобладают, поэтому в них надо разбираться
- Мы рассмотрим байесовские методы в следующей лекции, а сейчас введём несколько полезных для них техник

Постановка задачи со скрытыми переменными

Вернёмся к нашим общим обозначениям:

$X = (d_i, w_i)_{i=1}^n$ — 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 не Ω, а его распределение $p(\Omega|X, \gamma)$.

Байесовский вывод *апостериорного распределения*:

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

Дальнейший план — поэтапно усложнять постановку задачи:

1. Z — наблюдаемые переменные (временное упрощение)
2. Z — скрытые переменные
3. Z, Φ, Θ — скрытые переменные (в следующей лекции)

Функция совместного правдоподобия X и Z

Допустим (временно), что скрытые переменные Z известны.
Тогда известны и все частоты, связанные с темами:

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

Воспользуемся независимостью элементов выборки (d_i, w_i, t_i):

$$p(X, Z|\Omega) = \prod_{i=1}^{n} p(d_i,w_i,t_i|\Omega) = \prod_{d,w,t} p(d,w,t|\Omega)^{n_{dwt}} =$$

$$= \prod_{d,w,t} \big(p(w|t,\Phi)\,p(t|d,\Theta)\,p(d)\big)^{n_{dwt}} =$$

$$= \prod_{d,w,t} \big(\phi_{wt}\theta_{td}p_d\big)^{n_{dwt}} = \prod_{d} p_d^{n_d} \prod_{w,t} \phi_{wt}^{n_{wt}} \prod_{d,t} \theta_{td}^{n_{td}}.$$

В дальнейшем эта функция нам неоднократно понадобится

Случай известных Z и равномерного априорного распределения

Допустим (временно), что априорное распределение равномерно.

Максимизация логарифма правдоподобия

$$\ln p(X, Z|\Phi, \Theta) = \sum_{w,t} n_{wt} \ln \phi_{wt} + \sum_{d,t} n_{td} \ln \theta_{td} \;\to\; \max_{\Phi,\Theta}$$

при ограничениях нормировки и неотрицательности

$$\phi_{wt} \geqslant 0; \quad \sum_w \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_t \theta_{td} = 1$$

Решение — частотные оценки условных вероятностей:

$$\phi_{wt} = \frac{n_{wt}}{n_t} = \underset{w \in W}{\operatorname{norm}}(n_{wt}), \qquad n_t = \sum_w n_{wt};$$

$$\theta_{td} = \frac{n_{td}}{n_d} = \underset{t \in T}{\operatorname{norm}}(n_{td}), \qquad n_d = \sum_t n_{td}.$$

Доказательство

Чтобы применить условия KKT, выписываем лагранжиан:

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

Условия KKT для стационарной точки лагранжиана:

$$\frac{\partial \mathscr{L}}{\partial \phi_{wt}} = n_{wt} \frac{1}{\phi_{wt}} - \lambda_t = 0$$
$$n_{wt} = \lambda_t \phi_{wt}$$
$$n_t = \lambda_t$$
$$\phi_{wt} = \frac{n_{wt}}{n_t}$$

$$\frac{\partial \mathscr{L}}{\partial \theta_{td}} = n_{td} \frac{1}{\theta_{td}} - \mu_d = 0$$
$$n_{td} = \mu_d \theta_{td}$$
$$n_d = \mu_d$$
$$\theta_{td} = \frac{n_{td}}{n_d}$$

Сопряженные распределения

Пусть теперь априорное распределение $p(\Omega|\gamma)$ — Дирихле.
Распределение Дирихле — *сопряжённое* к мультиномиальному.
Поэтому апостериорное $p(\Omega|X, Z, \gamma)$ — тоже Дирихле:

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

$$= \prod_d p_d^{n_d} \prod_{w,t} \phi_{wt}^{n_{wt}} \prod_{d,t} \theta_{td}^{n_{td}} \prod_t \mathrm{Dir}(\phi_t|\beta) \prod_d \mathrm{Dir}(\theta_d|\alpha) \propto$$

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

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

$$= \prod_t \mathrm{Dir}(\phi_t|\tilde{\beta}_t) \prod_d \mathrm{Dir}(\theta_d|\tilde{\alpha}_d),$$

где $\tilde{\beta}_{wt} = n_{wt} + \beta_w - 1$, $\tilde{\alpha}_{td} = n_{td} + \alpha_t - 1$.

Случай известных Z и априорного распределения Дирихле

Пусть априорное распределение $p(\Omega|\gamma)$ — Дирихле.

Максимизация правдоподобия апостериорного распределения:

$$\ln p(X, Z|\Phi, \Theta)\, p(\Phi, \Theta|\beta, \alpha) =$$
$$= \sum_{w,t}(n_{wt} + \beta_w - 1)\ln \phi_{wt} + \sum_{d,t}(n_{td} + \alpha_t - 1)\ln \theta_{td} \to \max_{\Phi,\Theta}$$

при ограничениях нормировки и неотрицательности

$$\phi_{wt} \geqslant 0; \quad \sum_w \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_t \theta_{td} = 1$$

Решение — сглаженные оценки условных вероятностей:

$$\phi_{wt} = \underset{w \in W}{\text{norm}}(n_{wt} + \beta_w - 1), \qquad n_t = \sum_w n_{wt};$$
$$\theta_{td} = \underset{t \in T}{\text{norm}}(n_{td} + \alpha_t - 1), \qquad n_d = \sum_t n_{td}.$$

Максимизация неполного правдоподобия

Проблема — возникает сумма под логарифмом:

$$\ln p(X|\Omega) = \ln \sum_Z p(X, Z|\Omega) \;\to\; \max_\Omega$$

Формула условной вероятности:

$$p(X, Z|\Omega) = p(Z|X, \Omega)p(X|\Omega) \;\Rightarrow\; p(X|\Omega) = \frac{p(X,Z|\Omega)}{p(Z|X,\Omega)}$$

Для произвольного распределения $q(Z)$

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

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

Основная идея EM-алгоритма. Задача E-шага

Максимизировать нижнюю оценку $L(q, \Omega)$ то по q, то по Ω:

$$\text{E-шаг:} \quad L(q, \Omega) \to \max_q$$
$$\text{M-шаг:} \quad L(q, \Omega) \to \max_\Omega$$

Задача E-шага.

Подставим $p(X, Z|\Omega) = p(Z|X, \Omega)p(X|\Omega)$ в формулу $L(q, \Omega)$:

$$\sum_Z q(Z) \ln p(Z|X, \Omega) + \underbrace{\sum_Z q(Z)}_{=1} \underbrace{\ln p(X|\Omega)}_{\text{const по } q} - \sum_Z q(Z) \ln q(Z) \to \max_q$$

$$\mathrm{KL}\big(q(Z) \,\big\|\, p(Z|X, \Omega)\big) \to \min_q$$

Утв. 1. $q(Z) = p(Z|X, \Omega)$ — точное решение задачи E-шага.

Утв. 2. $L(q, \Omega)$ — достигаемая нижняя оценка $\ln p(X|\Omega)$.

EM-алгоритм. Обоснование сходимости

Мы вывели EM-алгоритм для Z и Ω общего вида:

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

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

и доказали его *сходимость в слабом смысле:*

- на каждом шаге правдоподобие $\ln p(X|\Omega)$ увеличивается;
- не гарантируется достижение max с заданной точностью;
- не гарантируется глобальная сходимость, так как задача в общем случае многоэкстремальная (на практике важен выбор начального приближения).

N.B. Если скрытая переменная Z не дискретна, а непрерывна, то суммирование $\sum_Z$ заменяется интегрированием $\int_Z$.

Максимизация регуляризованного правдоподобия

Пусть $p(\Omega)$ — априорное распределение параметров модели

Принцип максимума апостериорной вероятности:

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

Регуляризатор $R(\Omega)$ может даже и не иметь вероятностной интерпретации, тем не менее, все выкладки остаются в силе!

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

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

Регуляризаторы используются для формализации дополнительных требований к вероятностной модели.

Регуляризованный EM-алгоритм для тематической модели

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

E-шаг: в силу независимости элементов выборки

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

M-шаг:

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

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

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

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

Регуляризованный EM-алгоритм для тематической модели

... продолжаем вывод формулы M-шага:

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

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

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

Чтобы применить условия KKT, выписываем лагранжиан:

$$\mathscr{L}(\Phi,\Theta) = \sum_{w,t} n_{wt}\ln\phi_{wt} - \sum_{t}\lambda_t\left(\sum_{w}\phi_{wt} - 1\right) +$$

$$+ \sum_{d,t} n_{td}\ln\theta_{td} - \sum_{d}\mu_d\left(\sum_{t}\theta_{td} - 1\right) + R(\Phi,\Theta)$$

Регуляризованный EM-алгоритм для тематической модели

Условия ККТ для стационарной точки лагранжиана:

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

$$\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+ = \lambda_t \phi_{wt} \qquad \left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+ = \mu_d \theta_{td}$$

$$\phi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \qquad \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$$

Ещё раз вывели формулы ARTM, теперь из общего EM-алгоритма.
Преимущество — есть доказательство (слабой) сходимости.

Частные случаи:

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

LDA: $R(\Phi, \Theta) = \ln \prod_{t \in T} \text{Dir}(\phi_t | \beta) \prod_{d \in D} \text{Dir}(\theta_d | \alpha).$

Промежуточный итог

Мы узнали более общий вариант EM-алгоритма:

- также снабжённый возможностью регуляризации,
- для которого имеется доказательство слабой сходимости,
- используемый в методах байесовского вывода.

Следующая лекция — про *байесовский вывод*, который

- даёт апостериорные распределения $p(\Omega|X)$,
  хотя в BTM используются только точечные оценки Ω.
- намного более громоздкий по сравнению с ARTM,
  хотя в литературе именно он в основном и используется.
- претендует на то, чтобы оценивать меньше параметров,
  хотя на деле оценивает те же Φ и Θ, плюс гиперпараметры.

Способны ли PLSA и LDA восстановить истинные темы?

Матрицы Φ_0 и Θ_0 порождаются распределением Дирихле.
Синтетическая коллекция порождается матрицами Φ_0 и Θ_0.
Размеры: $|D| = 500$, $|W| = 1000$, $|T| = 30$, $n_d \in [100, 600]$.

Цель — сравнить восстановленные распределения $p(i|j)$
с исходными синтетическими распределениями $p_0(i|j)$
по среднему расстоянию Хеллингера:

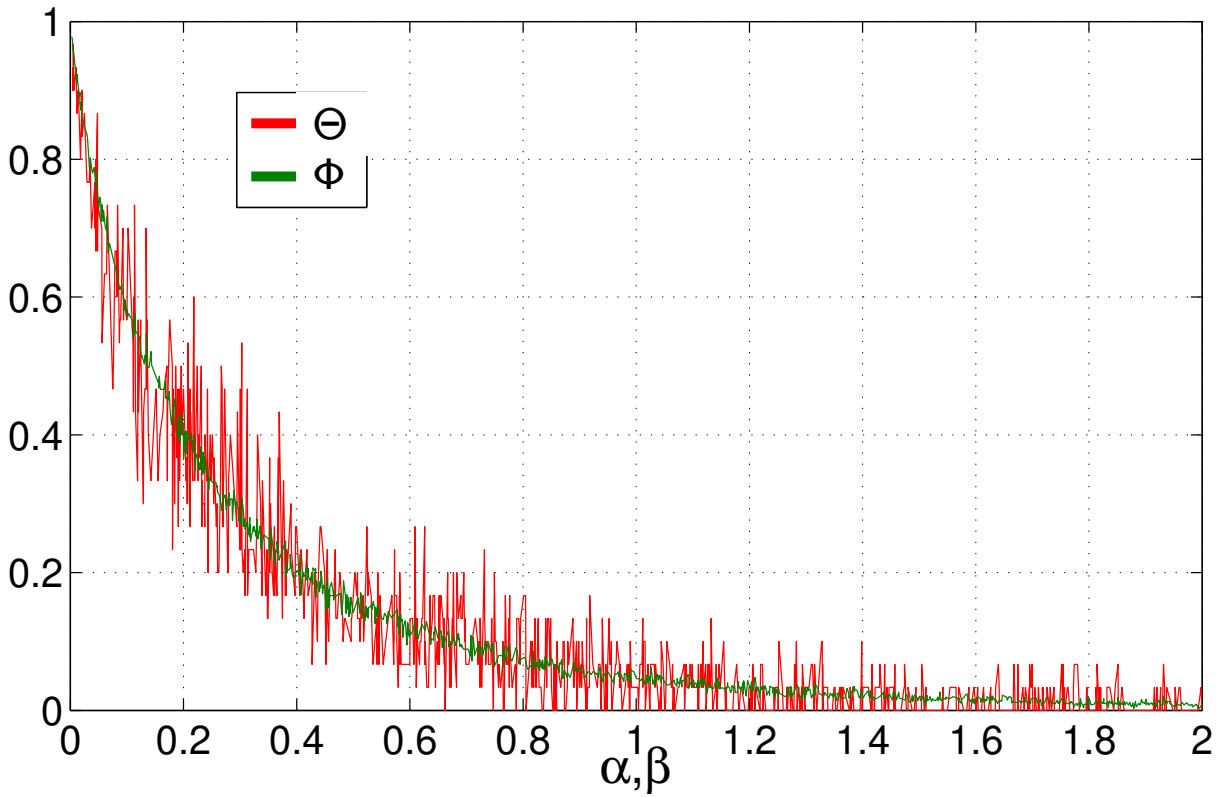
$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{p(i|j)} - \sqrt{p_0(i|j)} \right)^2},$$

как для самих матриц Φ и Θ, так и для их произведения:

$$D_\Phi = H(\Phi, \Phi_0);$$
$$D_\Theta = H(\Theta, \Theta_0);$$
$$D_{\Phi\Theta} = H(\Phi\Theta, \Phi_0\Theta_0).$$

Разреженность векторов, порождаемых распределением Dir

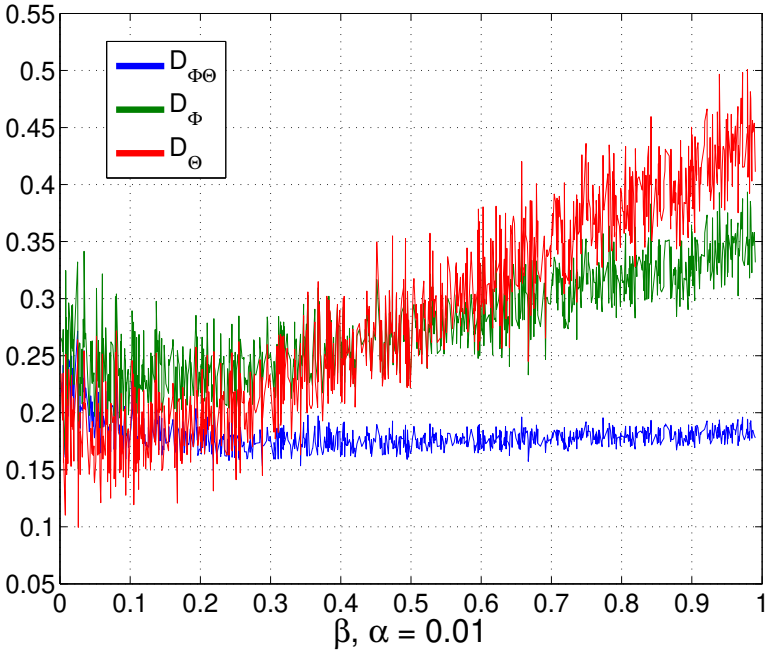
Зависимость разреженности (доли почти нулевых элементов) распределений $\theta_d^0 \sim \mathrm{Dir}(\alpha)$ и $\phi_t^0 \sim \mathrm{Dir}(\beta)$ от параметров α и β симметричного распределения Дирихле:



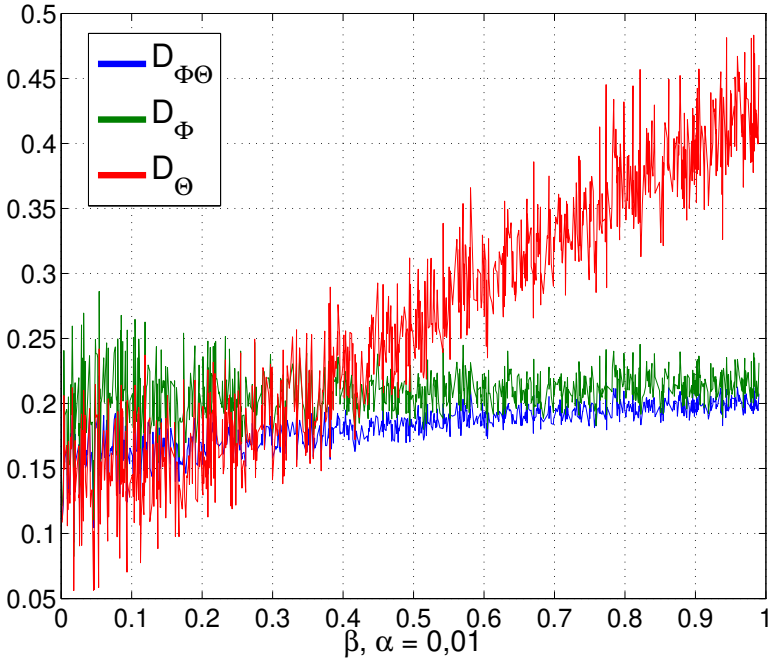
Неустойчивость восстановления матриц Φ и Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Φ_0 при фиксированном $\alpha = 0.01$

PLSA



LDA

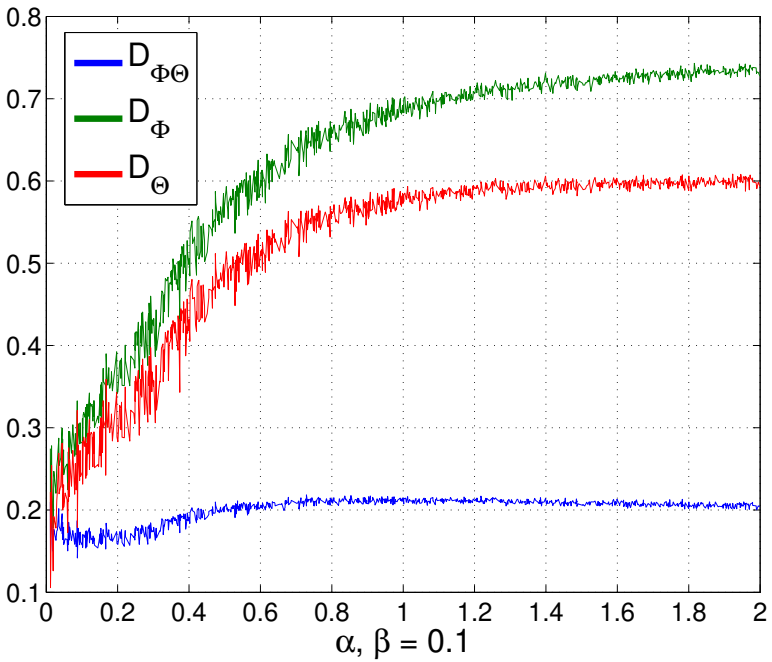


Виталий Глушаченков. Устойчивость матричных разложений в задачах тематического моделирования. Магистерская диссертация, МФТИ, 2013.

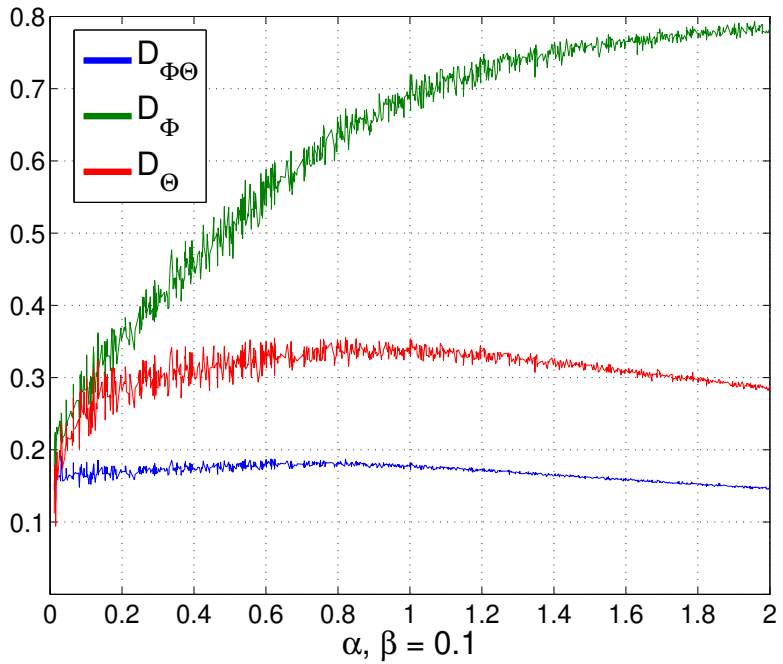
Неустойчивость восстановления матриц Φ и Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Θ_0 при фиксированном $\beta = 0.1$

PLSA



LDA



Виталий Глушаченков. Устойчивость матричных разложений в задачах тематического моделирования. Магистерская диссертация, МФТИ, 2013.

Второй эксперимент — на реальных данных

Посты ЖЖ: $|D| = 300\,\text{K}$, $|W| = 154\,\text{K}$, $n = 35\,\text{M}$, $|T| = 120$.
LDA: симметричное распределение Дирихле, $\beta = 0.1$, $\alpha = 0.5$.

Цель эксперимента — оценить различность тем, получаемых в нескольких запусках алгоритма LDA Gibbs Sampling.

Проблема «проклятия размерности»:
длинные хвосты мешают сравнивать распределения.

Доля существенных терминов в темах (word ratio):

$$WR = \frac{1}{|W|}\frac{1}{|T|}\sum_{w\in W}\sum_{t\in T}\left[\phi_{wt} > \frac{1}{|W|}\right] \quad \text{(в эксперименте } \sim 3.5\%\text{)}$$

Доля существенных тем в документах (document ratio):

$$DR = \frac{1}{|D|}\frac{1}{|T|}\sum_{d\in D}\sum_{t\in T}\left[\theta_{td} > \frac{1}{|T|}\right] \quad \text{(в эксперименте } \sim 11.5\%\text{)}$$

Koltcov S., Koltsova O., Nikolenko S. Latent Dirichlet Allocation: Stability and applications to studies of user-generated content. ACM WebSci, 2014.

Методика эксперимента

Оставлены слова w, имеющие $\phi_{wt} > \frac{1}{|W|}$ хотя бы в одной теме
Сокращение словаря (vocabulary reduction): 154 K $\to$ 8 K.

Дивергенция Кульбака–Лейблера между темами t и s:

$$\mathrm{KL}(t,s) = \sum_{w \in W} p(w|t) \ln \frac{p(w|t)}{p(w|s)}$$

Нормированная KL-близость пар тем t и s:

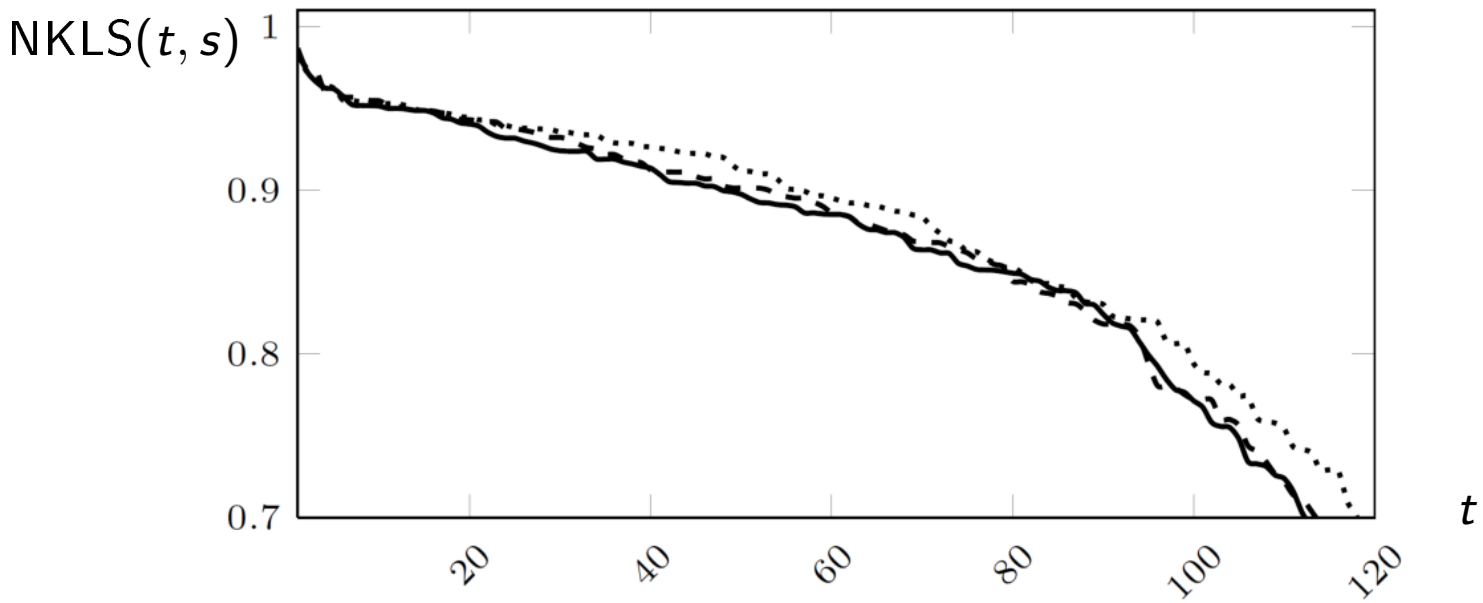
$$\mathrm{NKLS}(t,s) = \left(1 - \frac{\mathrm{KL}(t,s)}{\max\limits_{t',s'} \mathrm{KL}(t',s')}\right)$$

При $\mathrm{NKLS}(t,s) > 0.9$ в темах совпадают 30–50 топовых слов, и эксперты-социологи признают такие темы одинаковыми.

Koltcov S., Koltsova O., Nikolenko S. Latent Dirichlet Allocation: Stability and applications to studies of user-generated content. ACM WebSci, 2014.

Неустойчивость LDA в разных запусках

Результат эксперимента: нормированная KL-близость NKLS между темой t и ближайшей к ней s в другом запуске.



1. Менее 50% тем воспроизводятся от запуска к запуску.
2. Плохо воспроизводятся как мусорные темы, так и хорошие.

Koltcov S., Koltsova O., Nikolenko S. Latent Dirichlet Allocation: Stability and applications to studies of user-generated content. ACM WebSci, 2014.

Третий эксперимент: робастная тематическая модель

Гипотеза: каждое слово в документе (d, w) является
— либо тематическим, связанным с какой-то темой t,
— либо специфичным для данного документа (шум),
— либо общеупотребительным (фон).

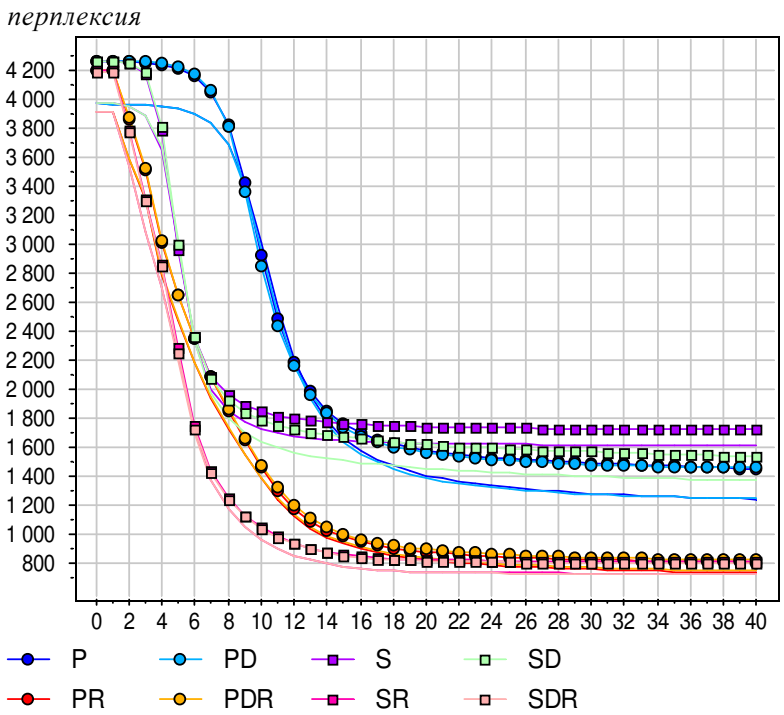
Модель смеси тематической, шумовой и фоновой компонент SWB (Special Words with Background):

$$p(w|d) = \gamma\pi_{dw} + \varepsilon\pi_w + (1 - \gamma - \varepsilon)\sum_{t\in T}\phi_{wt}\theta_{td}$$

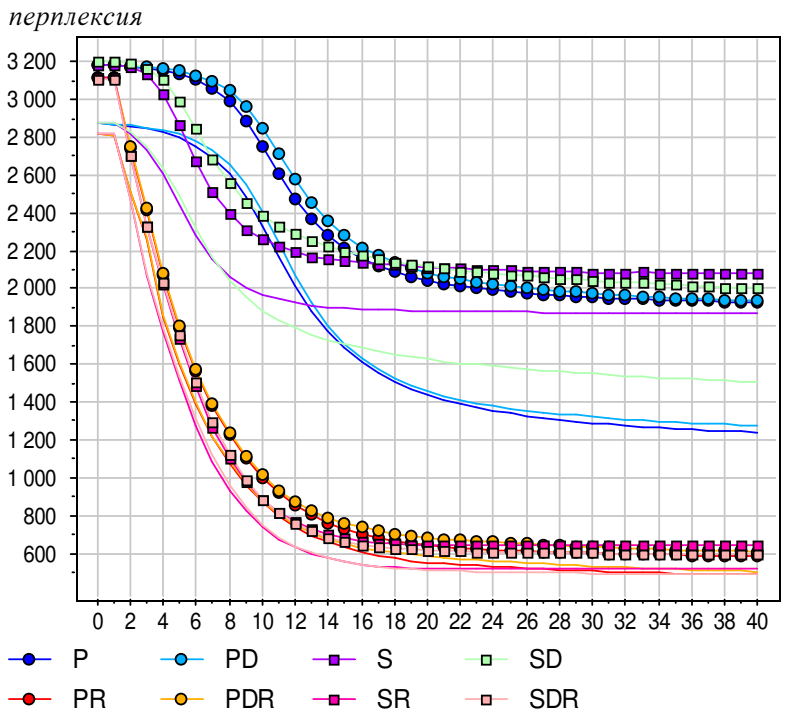
$\pi_{dw} \equiv p_{\text{ш}}(w|d)$ — *шумовая* компонента, γ — параметр;
$\pi_w \equiv p_{\text{ф}}(w)$ — *фоновая* компонента, ε — параметр.

Chemudugunta C., Smyth P., Steyvers M. Modeling general and specific aspects of documents with a probabilistic topic model. NIPS, 2006.

Эксперименты с робастными PLSA и LDA



Коллекция RuDis



Коллекция NIPS

Обозначения: P – PLSA, D – LDA ($\alpha_t = 0.5$, $\beta_w = 0.01$)
S – сэмплирование темы из $p(t|d, w)$ для каждого d, w
R – робастность (шум $\gamma = 0.3$, фон $\varepsilon = 0.01$)

A.Potapenko, K.Vorontsov. Robust PLSA performs better than LDA. ECIR-2013.

Выводы из экспериментов

- Матрицы Φ, Θ устойчиво восстанавливаются только при сильной разреженности Φ_0, Θ_0 (более 90% нулей)
- Произведение $\Phi\Theta$ восстанавливается устойчиво, независимо от разреженности исходных Φ_0, Θ_0
- В разных запусках со случайной инициализацией или сэмплированием строятся существенно различные темы
- PLSA не переобучается, а лишь хуже моделирует малые вероятности редких слов, которые не интересны.
- Распределение Дирихле — слишком слабый регуляризатор

Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models. Machine Learning. Springer, 2015.

Koltcov S., Koltsova O., Nikolenko S. Latent Dirichlet Allocation: Stability and applications to studies of user-generated content. ACM WebSci, 2014.

Вместо резюме. Мифы про LDA

- LDA существенно меньше переобучается, чем PLSA
- LDA строит разреженные тематические модели
- LDA имеет меньше параметров по сравнению с PLSA
- LDA == тематическое моделирование

На самом деле,

- LDA и PLSA почти не отличаются на больших данных
- LDA не максимизирует разреженность моделей
- LDA имеет больше параметров по сравнению с PLSA
- LDA — лишь самая простая базовая модель
- LDA не имеет убедительных лингвистических обоснований

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.