Вероятностные тематические модели
Лекция 3. Аддитивная регуляризация тематических моделей (ARTM)

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 • весна 2017

Содержание

Напоминание. Задача тематического моделирования

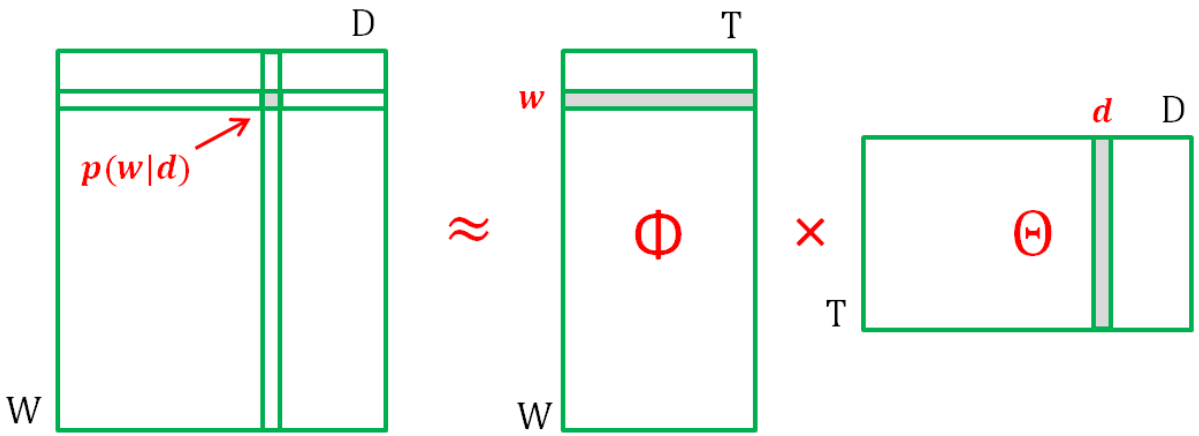
Дано: коллекция текстовых документов, $p(w|d) = \frac{n_{dw}}{n_d}$

Вероятностная тематическая модель:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

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

Это задача стохастического матричного разложения:



Напоминание. PLSA (Probabilistic Latent Semantic Analysis)

Задача: найти максимум правдоподобия

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

при ограничениях неотрицательности и нормировки

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

E-шаг:
$$n_{dwt} = n_{dw} \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}};$$

M-шаг:
$$\phi_{wt} = \frac{n_{wt}}{n_t}; \quad n_{wt} = \sum_{d\in D} n_{dwt}; \quad n_t = \sum_w n_{wt}$$

$$\theta_{td} = \frac{n_{td}}{n_d}; \quad n_{td} = \sum_{w\in d} n_{dwt}; \quad n_d = \sum_t n_{td}$$

Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена*, если её решение

- существует,
- единственно,
- устойчиво.

Жак Саломон Адамар
(1865–1963)

Задача стохастического матричного разложения *некорректно поставлена*, так как имеется бесконечное множество решений:

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

для невырожденных $S_{T\times T}$ таких, что Φ', Θ' — стохастические.

Регуляризация — стандартный приём, введение новых ограничений или критериев, доопределяющих решение.

ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

Условия невырожденности решения

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

Тема t невырождена, если хотя бы для одного термина $w \in W$

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

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

Документ d невырожден, если хотя бы для одной темы $t \in T$

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

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

Напоминание. Условия Каруша–Куна–Таккера

Задача математического программирования:

$$\begin{cases} f(x) \to \min\limits_x; \\ g_i(x) \leqslant 0, \quad i = 1, \ldots, m; \\ h_j(x) = 0, \quad j = 1, \ldots, k. \end{cases}$$

Необходимые условия. Если x — точка локального минимума, то существуют множители μ_i, $i = 1, \ldots, m$, λ_j, $j = 1, \ldots, k$:

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{j=1}^{k} \lambda_j h_j(x); \\ g_i(x) \leqslant 0; \ h_j(x) = 0; \ \text{(исходные ограничения)} \\ \mu_i \geqslant 0; \ \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; \ \text{(условие дополняющей нежёсткости)} \end{cases}$$

Вывод системы уравнений из условий Каруша–Куна–Таккера

1. Условия ККТ для ϕ_{wt} (для θ_{td} всё аналогично):

$$\sum_d n_{dw} \frac{\theta_{td}}{p(w|d)} + \frac{\partial R}{\partial \phi_{wt}} = \lambda_t - \mu_{wt}; \quad \mu_{wt} \geqslant 0; \quad \mu_{wt}\phi_{wt} = 0.$$

2. Умножим обе части равенства на ϕ_{wt} и выделим p_{tdw}:

$$\phi_{wt}\lambda_t = \sum_d n_{dw} \frac{\phi_{wt}\theta_{td}}{p(w|d)} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} = n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}.$$

3. Если $\lambda_t \leqslant 0$, то тема t вырождена, $\phi_{wt} \equiv 0$ для всех w.
4. Если $\lambda_t > 0$, то либо $\phi_{wt} = 0$, либо $n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} > 0$:

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

5. Суммируем обе части равенства по $w \in W$:

$$\lambda_t = \sum_{w \in W} \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+.$$

6. Подставим λ_t из (5) в (4), получим требуемое. ■

Комбинирование регуляризаторов

Максимизация log правдоподобия с k регуляризаторами R_i:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta},$$

где τ_i — коэффициенты регуляризации.

EM-алгоритм: метод простой итерации для системы уравнений

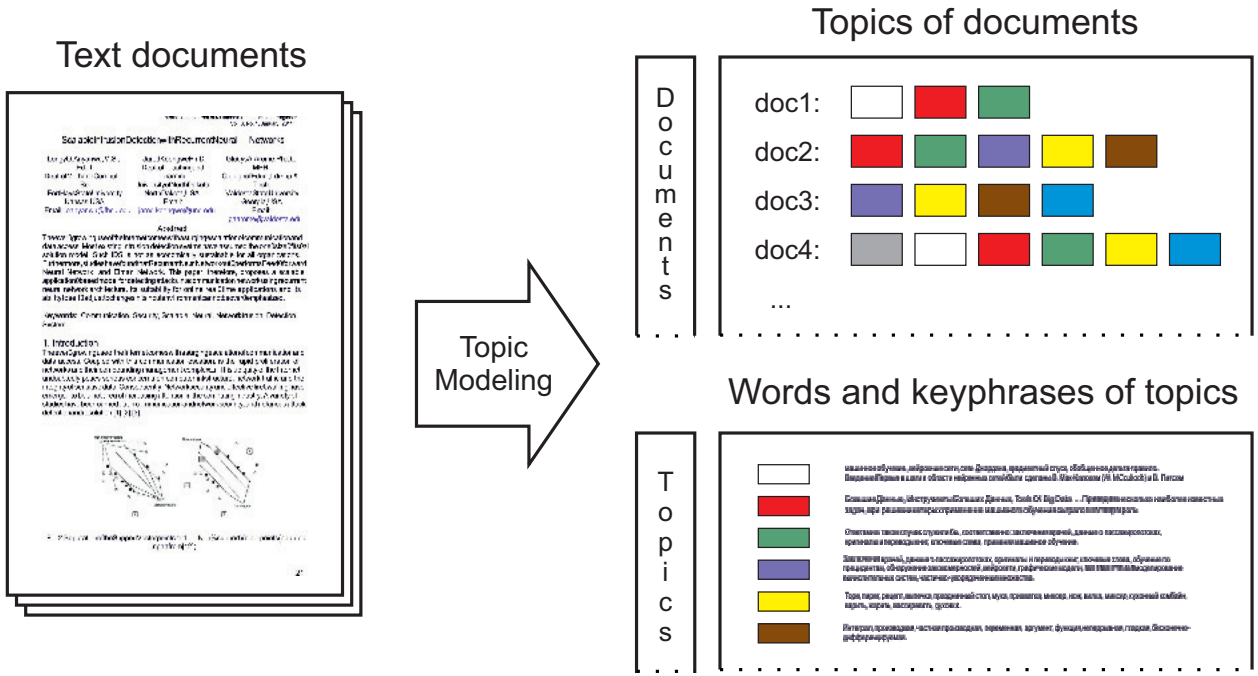
$$\text{E-шаг:} \quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum\limits_{d\in D} n_{dw} p_{tdw} + \phi_{wt} \sum\limits_{i=1}^{k} \tau_i \frac{\partial R_i}{\partial \phi_{wt}} \right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum\limits_{w\in d} n_{dw} p_{tdw} + \theta_{td} \sum\limits_{i=1}^{k} \tau_i \frac{\partial R_i}{\partial \theta_{td}} \right) \end{cases}$$

Тематика (тематический вектор, тематический профиль)

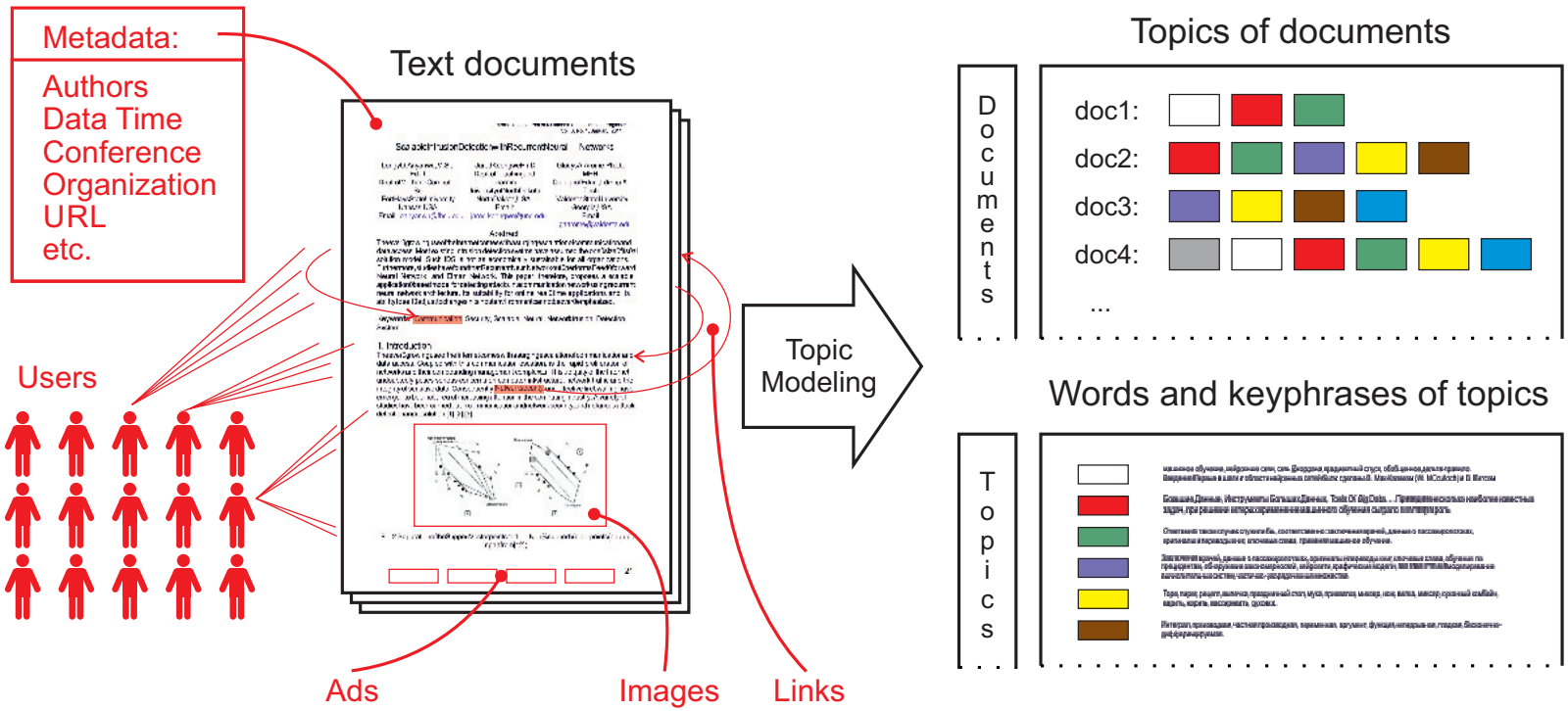
Тематика документа: $p(t|d) = \theta_{td}$

Тематика термина: $p(t|w) = p(w|t)p(t)/p(w) = \phi_{wt} n_t / n_w$



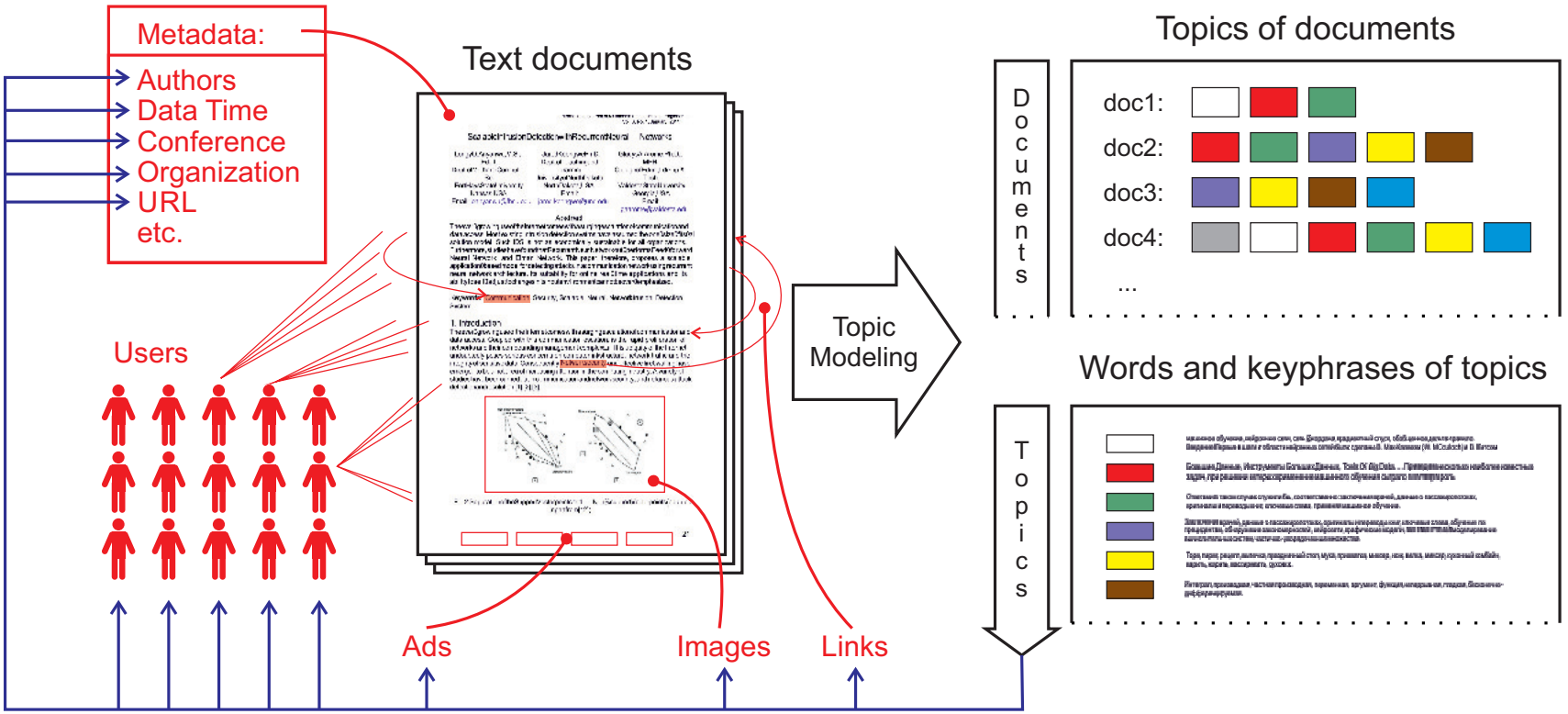
Мультимодальная тематическая модель

Документ — универсальный контейнер не только терминов, но и токенов других *модальностей*: $p(t|\text{автор})$, $p(t|\text{время})$, $p(t|\text{ссылка})$, $p(t|\text{баннер})$, $p(t|\text{изображение})$, $p(t|\text{пользователь})$



Мультимодальная тематическая модель

Документы могут содержать *токены* разных модальностей.
Каждая *модальность* $m \in M$ описывается своим словарём W^m.
Каждая тема имеет своё распределение $\phi_{wt} = p(w|t)$ на W^m.



Мультимодальная ARTM

W^m — словарь токенов m-й модальности, $m \in M$

$W = W^1 \sqcup \cdots \sqcup W^M$ — объединённый словарь всех модальностей

Максимизация суммы log правдоподобий с регуляризацией:

$$\sum_{m\in M} \tau_m \sum_{d\in D} \sum_{w\in W^m} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Оффлайновый EM-алгоритм для ARTM

Идея: E-шаг встраивается внутрь M-шага

Вход: коллекция D, число тем $|T|$, число итераций $i_{\max}$;
Выход: матрицы терминов тем Θ и тем документов Φ;

инициализация ϕ_{wt}, θ_{td} для всех $d \in D$, $w \in W$, $t \in T$;
для всех итераций $i = 1, \ldots, i_{\max}$
  $n_{wt}, n_{td}, n_t, n_d := 0$ для всех $d \in D$, $w \in W$, $t \in T$;
  для всех $d \in D$, $m \in M$, $w \in d \cap W^m$
    $p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$ для всех $t \in T$;
    $n_{wt}, n_{td} \mathrel{+}= \tau_m n_{dw} p_{tdw}$ для всех $t \in T$;
  $\phi_{wt} := \underset{w \in W^m}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)$ для всех $m \in M$, $w \in W^m$, $t \in T$;
  $\theta_{td} := \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$ для всех $d \in D$, $t \in T$;

Онлайновый EM-алгоритм для ARTM

Идея: коллекция D разбивается на пакеты D_b, $b = 1, \ldots, B$

Вход: коллекция D, $\delta \equiv \texttt{decay\_weight}$, $\alpha \equiv \texttt{apply\_weight}$;
Выход: матрица Φ;

инициализировать ϕ_{wt} для всех $w \in W$, $t \in T$;
$n_{wt} := 0$, $\tilde{n}_{wt} := 0$ для всех $w \in W$, $t \in T$;
для всех пакетов D_b, $b = 1, \ldots, B$
  $(\tilde{n}_{wt}) := (\tilde{n}_{wt}) + \textbf{ProcessBatch}(D_b, \Phi)$;
  если пора обновить матрицу Φ **то**
    $n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}$ для всех $w \in W$, $t \in T$;
    $\phi_{wt} := \underset{w \in W^m}{\text{norm}} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ для всех $m \in M$, $w \in W^m$, $t \in T$;
    $\tilde{n}_{wt} := 0$ для всех $w \in W$, $t \in T$;

ProcessBatch: обработка пакета в онлайновом EM-алгоритме

Идея: обработать пакет, не меняя матрицу Φ

Вход: пакет D_b, матрица $\Phi = (\phi_{wt})$;
Выход: матрица $(\tilde{n}_{wt})$;

$\tilde{n}_{wt} := 0$ для всех $w \in W,\ t \in T$;
для всех $d \in D_b$
  инициализировать $\theta_{td} := \frac{1}{|T|}$ для всех $t \in T$;
  повторять
    $p_{tdw} := \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$ для всех $w \in d,\ t \in T$;
    пост-обработка матрицы $(p_{tdw})_{T \times n_d}$ при необходимости;
    $\theta_{td} := \underset{t \in T}{\text{norm}}\left(\sum_{m \in M} \tau_m \sum_{w \in W^m} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\right)$ для всех $t \in T$;
  пока θ_d не сойдётся;
  $\tilde{n}_{wt} := \tilde{n}_{wt} + \tau_m n_{dw} p_{tdw}$ для всех $m \in M,\ w \in W^m,\ t \in T$;

Сравнение оффлайнового и онлайнового алгоритмов

Оффлайн EM-алгоритм:

1. многократное итерирование по коллекции
2. однократный проход по документу
3. хранение матрицы Θ
4. обновление Φ в конце каждого прохода по коллекции
5. применяется при обработке небольших коллекций

Онлайн EM-алгоритм:

1. однократный проход по коллекции
2. многократное итерирование по каждому документу
3. нет необходимости хранить матрицу Θ
4. обновление Φ через заданное число пакетов
5. применяется при потоковой обработке больших коллекций

Гипотеза об априорных распределениях Дирихле

Вероятностная тематическая модель: $p(w|d) = \sum_{t \in T} \underbrace{p(w|t)}_{\phi_{wt}} \underbrace{p(t|d)}_{\theta_{td}}$

1. Пусть $\theta_d = (\theta_{td})_{t \in T} \in \mathbb{R}^{|T|}$ — случайные векторы из распределения Дирихле с параметром $\alpha \in \mathbb{R}^{|T|}$:

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

2. Пусть $\phi_t = (\phi_{wt})_{w \in W} \in \mathbb{R}^{|W|}$ — случайные векторы из распределения Дирихле с параметром $\beta \in \mathbb{R}^{|W|}$:

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

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

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:

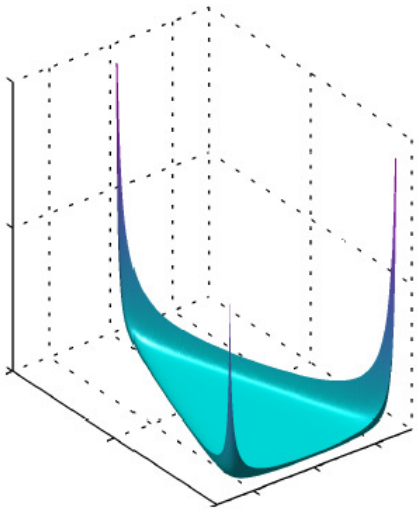
Вход: векторы гиперпараметров β, α;
Выход: коллекция документов;

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 $p(w|t_i) \equiv \phi_{wt}$;

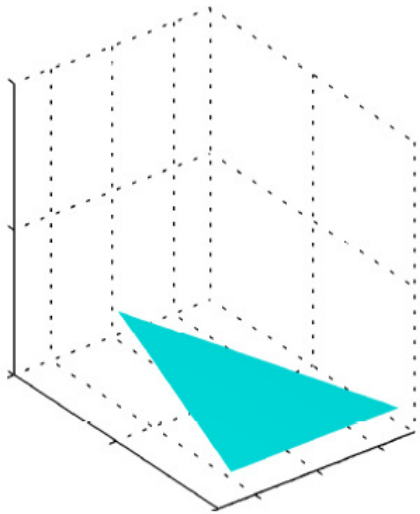
Почему именно распределение Дирихле?

– Может порождать сглаженные или разреженные векторы
– Имеет параметры, управляющие степенью разреженности
– Неплохо описывает кластерные структуры на симплексе
– Является сопряжённым к мультиномиальному распределению

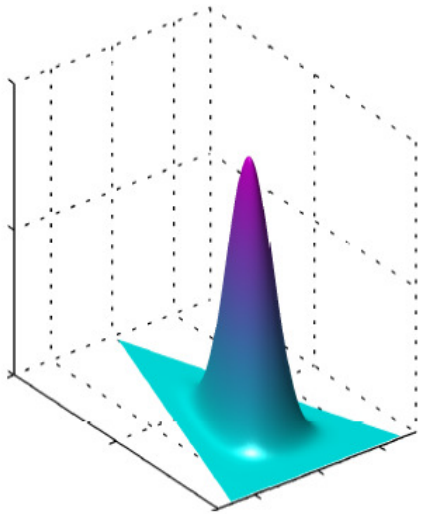
Пример. $\mathrm{Dir}(\theta|\alpha)$ при $|T| = 3$, $\theta, \alpha \in \mathbb{R}^3$:



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

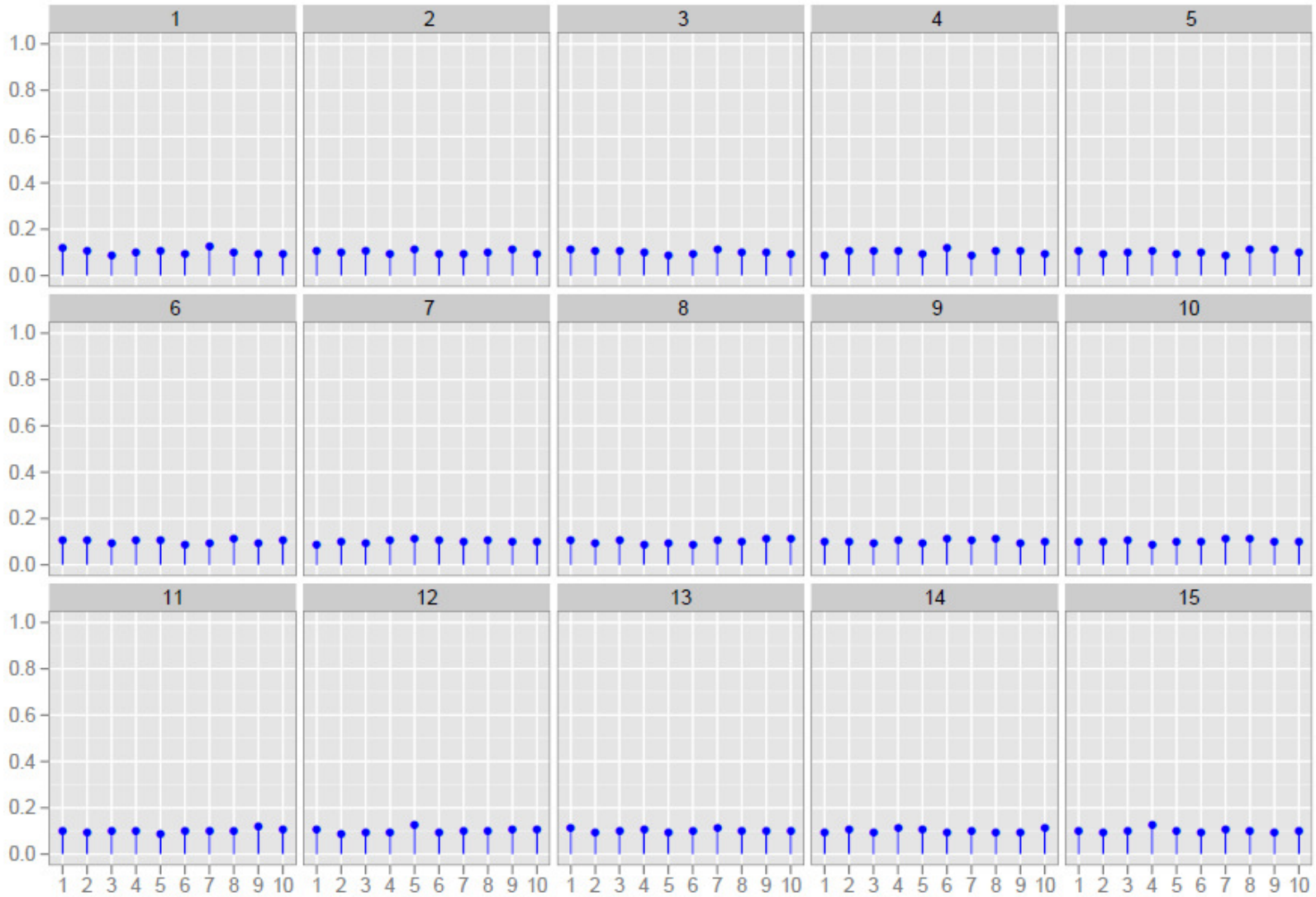


$\alpha_1 = \alpha_2 = \alpha_3 = 1$

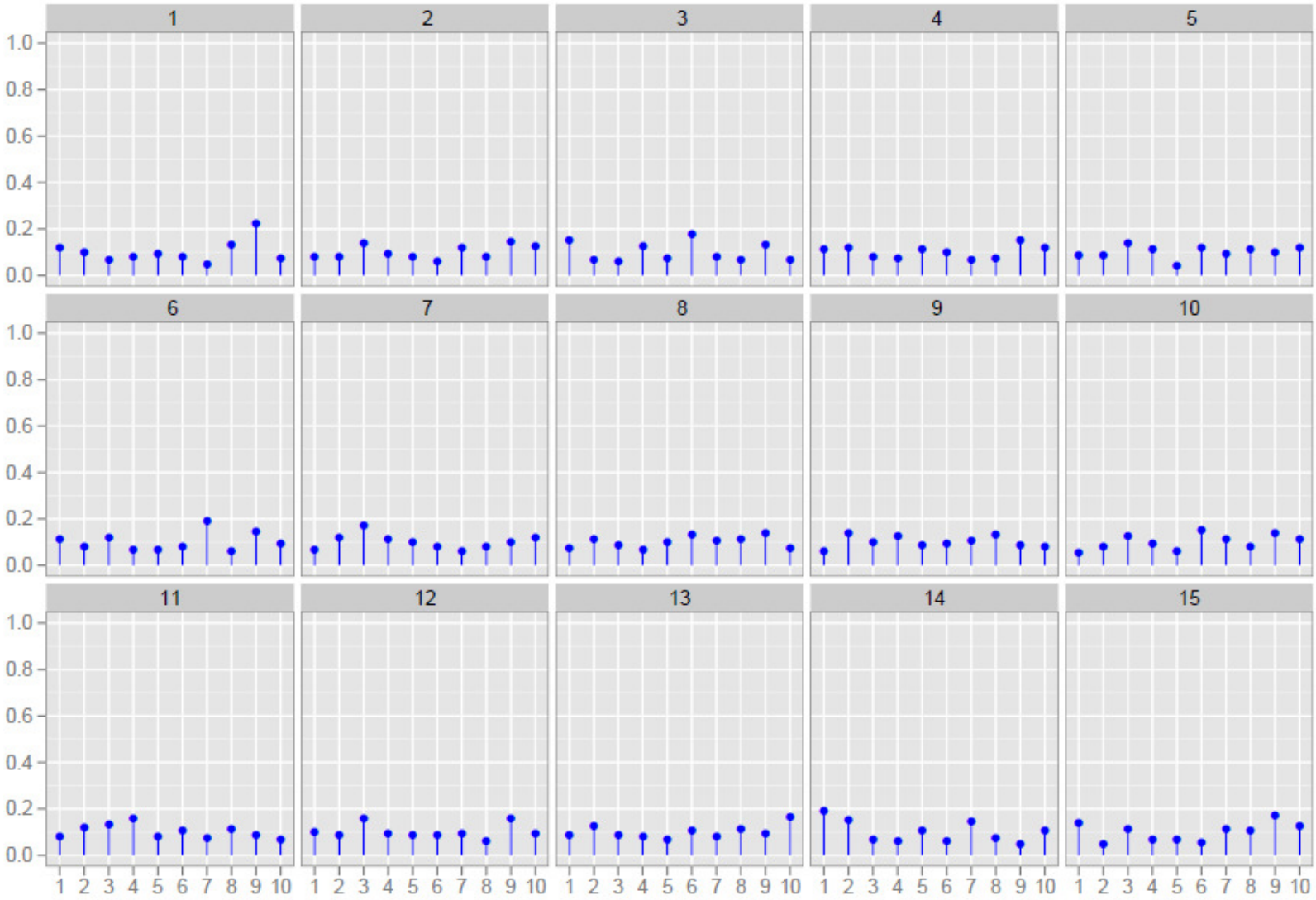


$\alpha_1 = \alpha_2 = \alpha_3 = 10$

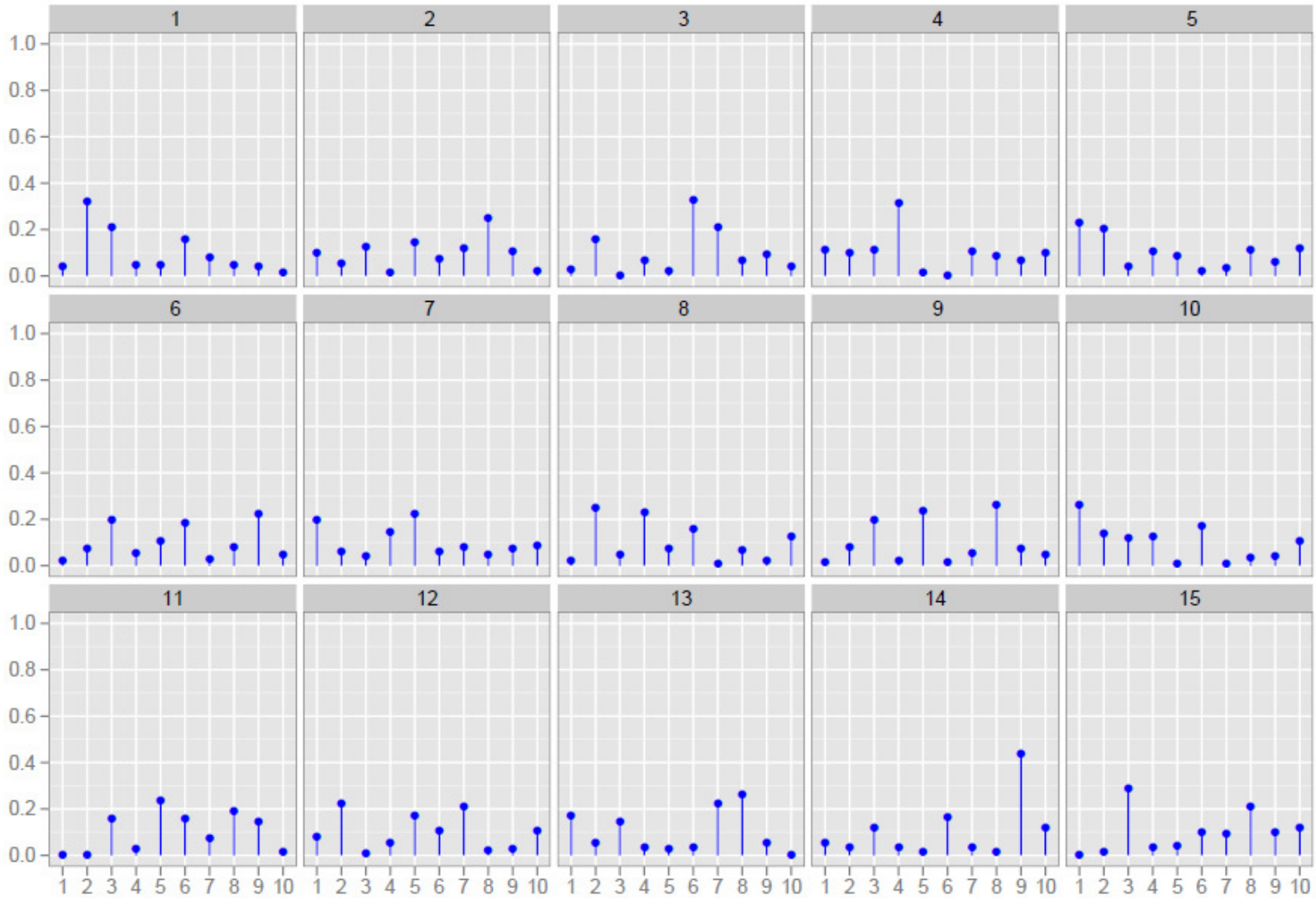
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 100$, 10 тем, 15 документов



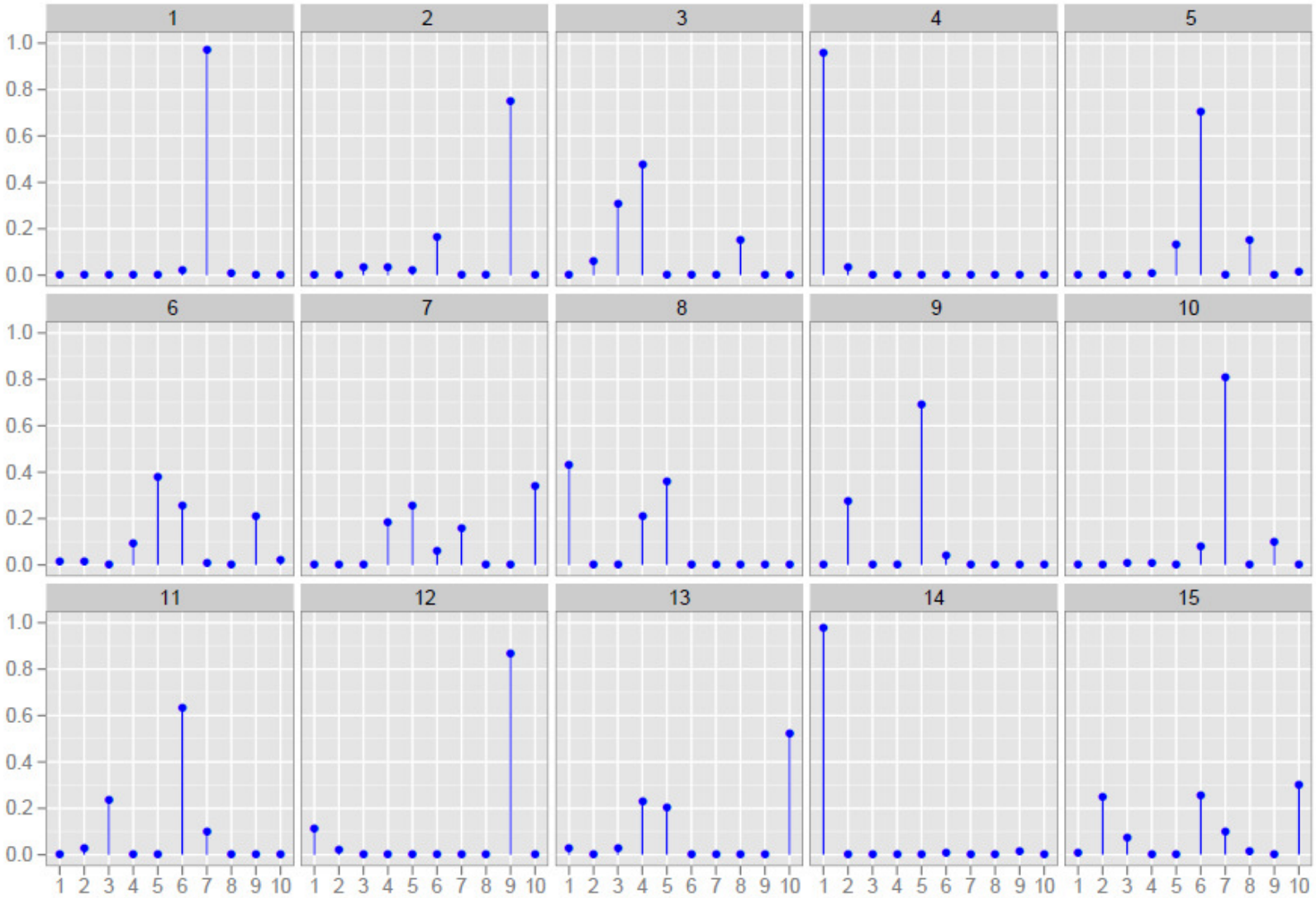
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 10$, 10 тем, 15 документов



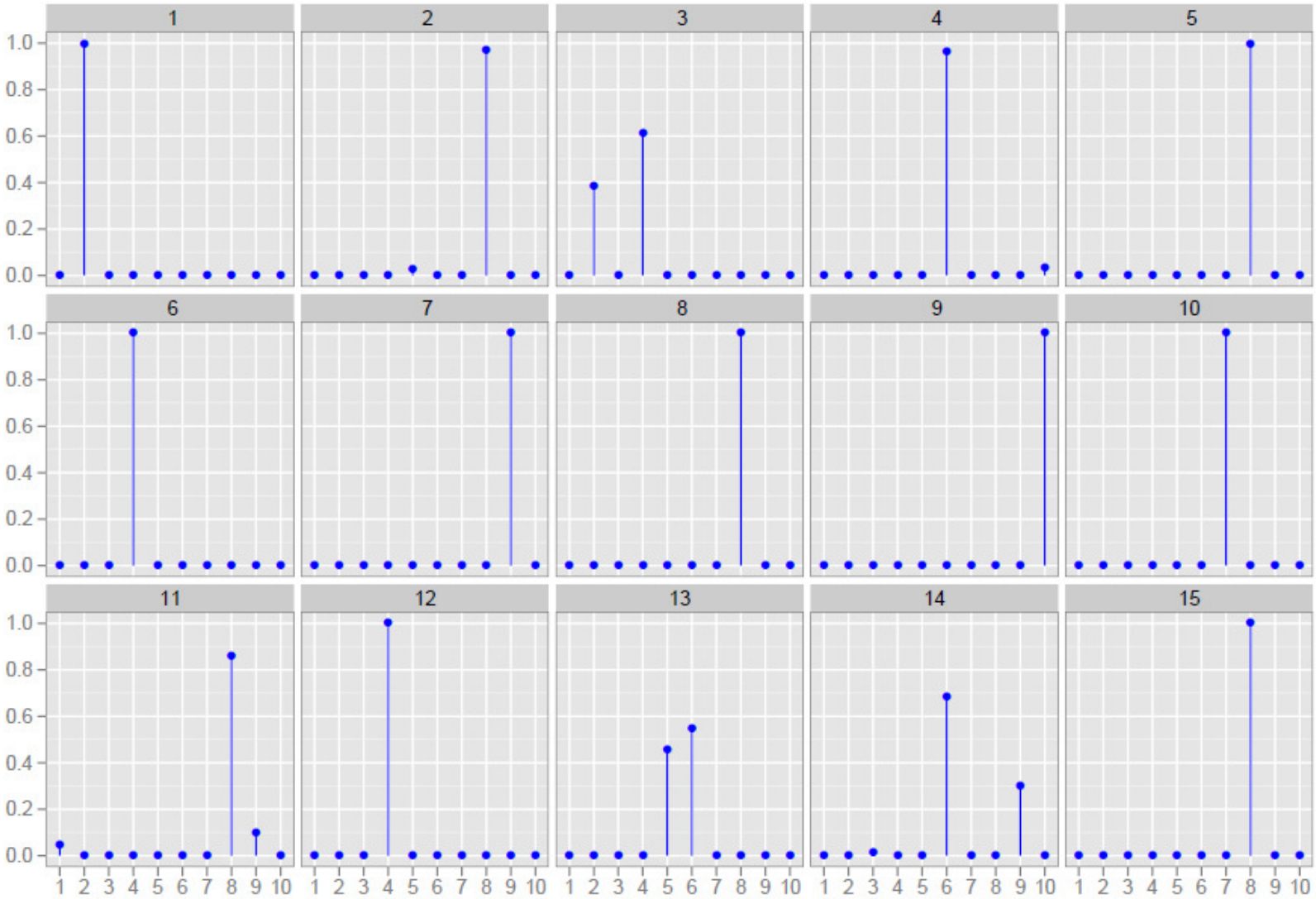
Распределение $\mathrm{Dir}(\theta_d|\alpha)$ при $\alpha_t \equiv 1$, 10 тем, 15 документов



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



Максимизация апостериорной вероятности для модели LDA

Совместное правдоподобие данных и модели:

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

Принцип MAP (maximum a posteriori probability)

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

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

Максимум апостериорной вероятности:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Мифы про LDA

- LDA существенно меньше переобучается, чем PLSA
- LDA строит более разреженную тематическую модель
- LDA имеет меньше параметров по сравнению с PLSA
- LDA == тематическое моделирование

На самом деле,

- LDA отличается от PLSA только на малых n_{wt}, n_{td}
- LDA имеет больше параметров по сравнению с PLSA
- LDA не имеет убедительных лингвистических обоснований
- LDA не решает проблему неединственности разложения
- LDA — слабый и неинтересный регуляризатор

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.

Далее в этом разделе

- Прикладные проекты
  - текущие проекты, в которых есть данные, команда и примерное понимание путей решения
  - каждый проект слишком большой, чтобы делать его в одиночку, но разбивается на более простые подзадачи
- Задания по спецкурсу
  - практические задачи из проектов
  - синтетические задачи на исследование алгоритмов
  - задания по подготовке обзоров с докладом на семинаре
- Открытые проблемы
  - наука здесь!
  - перспективные направления исследований для дальнейшей работы в научной группе
  - творческие исследовательские задачи без чётких постановок и заранее известных правильных ответов

1. Новостной мониторинг для медиапланирования

- **Дано:**
  поток новостей СМИ ($\sim 100\,\text{K}/\text{день}$) и социальных медиа.
- **Найти:**
  1) иерархическая тематическая модель,
  2) спектр тем и спектр мнений по заданному тексту,
  3) фильтр потока по заданным спектрам тем и мнений,
  4) оценки разнообразия тем и мнений в потоке.
- **Критерий:**
  1) интерпретируемость и различность тем,
  2) интерпретируемость разделения тем на подтемы,
  3) асессорские оценки качества (около 10 критериев):
     — точность распознавания новых тем,
     — точность распознавания слов общей лексики и др.

2. Новостной мониторинг для поиска проблемных компаний

- **Дано:**
  1) поток новостных сообщений СМИ,
  2) семантические ядра тем по компаниям,
  3) семантические ядра тем по проблемным ситуациям,
  4) подвыборка известных случаев проблемных ситуаций.
- **Найти:**
  1) сообщения о проблемных ситуациях по компаниям,
  2) все темы по каждой компании,
  3) новые типы проблемных ситуаций.
- **Критерий:**
  1) интерпретируемость всех тем,
  2) точность и полнота поиска по известным случаям.

3. Сценарный анализ записей разговоров контакт-центра

- **Дано:**
  1) коллекция текстов разговоров,
  2) семантические ядра (обучающие тексты) тем,
  3) сегментная разметка подвыборки разговоров.
- **Найти:**
  1) граф сценариев разговоров,
  2) вероятность успешного исхода в любой точке разговора,
  3) онлайновые подсказки оператору,
  4) автоматические оценки качества работы операторов,
  5) рекомендации операторам.
- **Критерий:**
  1) точность выделения тем в разговорах,
  2) точность сегментации на размеченной подвыборке.

4. Агрегатор русскоязычного научно-популярного контента

- **Дано:**
  1) коллекции статей научно-популярных порталов,
  2) коллекция Википедии на русском языке.
- **Найти:**
  1) общая тематическая иерархия,
  2) контекстные рекомендации по заданному тексту,
  3) тематический разведочный поиск по заданному тексту,
  4) интерактивная графическая «карта знаний».
- **Критерий:**
  1) полнота и точность поиска,
  2) интерпретируемость и различность тем,
  3) интерпретируемость разделения тем на подтемы,
  4) точность асессорского поиска документа по иерархии,
  5) экспертные оценки «интересности» рекомендаций.

5. Тематический разведочный поиск по коллективному блогу

- **Дано:**
  1) коллекция Habrahabr.ru или TechCrunch.com,
  2) выборка тематических запросов (длинные тексты),
  3) асессорские оценки релевантности документов запросам.
- **Найти:**
  1) тематическая модель для разведочного поиска,
  2) признаки сходства тематических векторов,
  3) функции ранжирования документов по запросу.
- **Критерий:**
  1) точность и полнота поиска,
  2) качество ранжирования (MAP или NDCG).

Янина А.О., Воронцов К.В. Мультимодальные тематические модели для разведочного поиска в коллективном блоге. JMLDA. 2016.

6. Кросс-язычный разведочный поиск по патентной базе

- **Дано:**
  1) коллекция патентов США на английском языке,
  2) коллекция их машинных переводов на русский язык,
  3) коллекция двуязычных статей Википедии,
  4) выборка русскоязычных запросов (длинные тексты),
  5) асессорские оценки релевантности документов запросам.
- **Найти:**
  1) тематическая иерархия научно-технической информации,
  2) признаки сходства тематических векторов,
  3) функции ранжирования документов по запросу.
- **Критерий:**
  1) точность и полнота кросс-язычного поиска,
  2) качество ранжирования (MAP или NDCG).

7. Построение продуктовой иерархии по текстам госзакупок

- **Дано:**
  1) описания объектов закупок (~ 200 млн.),
  2) общероссийский классификатор продукции ОКПД.
- **Найти:**
  иерархический тематический классификатор продуктов.
- **Критерий:**
  1) интерпретируемость и различность тем,
  2) интерпретируемость разделения тем на подтемы,
  3) согласованность верхних уровней с ОКПД,
  4) точность асессорского поиска товара по иерархии.

Чиркова Н.А., Воронцов К.В. Аддитивная регуляризация мультимодальных иерархических тематических моделей JMLDA. 2016.

8. Выявление структуры отрасли по транзакционным данным

- **Дано:**
  1) база банковских транзакций между компаниями,
  2) коды ОКВЭД для компаний.
- **Найти:**
  1) латентные темы — виды экономической деятельности,
  2) их соответствие ОКВЭДам,
  3) граф товарно-денежных потоков отрасли,
  4) типовые бизнес-схемы компаний — лидеров отрасли.
- **Критерий:**
  1) точность описания транзакционных данных,
  2) интерпретируемость графа отрасли.

9. Диагностика заболеваний по электрокардиограмме

- **Дано:**
  1) электрокардиограммы, закодированные в символьные последовательности методом В.М.Успенского,
  2) диагнозы по 32 заболеваниям для каждой ЭКГ.
- **Найти:**
  1) диагностические эталоны каждого заболевания,
  2) решающее правило по каждому заболеванию.
- **Критерий:**
  1) чувствительность и специфичность диагностики,
  2) качество ранжирования диагнозов.

Тематическая модель классификации с частичным обучением

- **Дано:**
  коллекция текстов, частично классифицированных
- **Цель:**
  исследовать зависимость качества тематической классификации от числа классифицированных документов
- **Этапы решения:**
  1) разделить коллекцию на обучение и тест,
  2) построить классификатор, используя только классифицированные документы,
  3) улучшить качество классификации путём использования неразмеченных данных.

Переобучение тематической модели классификации

- **Дано:**
  коллекция текстов, классифицированных по темам
- **Цель:**
  исследовать зависимость переобучения от стратегии разреживания матрицы Θ
- **Этапы решения:**
  1) разделить коллекцию на обучение и тест
  2) построить тематические модели классификации при различных стратегиях разреживания
  3) подобрать оптимальный режим регуляризации

Фоновые темы в двухуровневой иерархии

- **Дано:**
  коллекция новостей, двухуровневая тематическая иерархия
- **Цель:**
  сравнить несколько стратегий регуляризации фоновых тем
- **Этапы решения:**
  1) фоновые темы переходят с родительского уровня на дочерний
  2) фоновые темы дочернего уровня не зависят от родительского
  3) для каждой родительской темы выделяется своя фоновая тема

Динамическая тематическая модель новостного потока

- **Дано:**
  коллекция новостей
- **Цель:**
  построить критерий, решающий, создавать ли новую тему или продолжать существующую
- **Этапы решения:**
  1) построить тематическую модель коллекции новостей
  2) построить алгоритм добавления блока новостей
  3) составить и разметить выборку пар «новость–продолжение»
  4) оптимизировать пороговое правило в критерии

Сравнение триплетов статей arXiv

- **Дано:**
  1) коллекция статей arXiv,
  2) набор триплетов статей
- **Цель:**
  показать, что тематические модели решают задачу семантической близости документов
- **Этапы решения:**
  1) восстановить соответствие документов коллекции статей arXiv и url статей,
  2) Обучить нейронную сеть Paragraph2vec и тематическую модель ARTM на коллекции статей arXiv,
  3) добиться для обеих моделей результатов, сопоставимых с результатами статьи

Конкурс Fakenews Challenge

- **Дано:**
  Коллекция англоязычных новостей и их заголовков, поделённая на 2 широких класса (relevant, not relevant) и 4 узких (discuss, agree, disagree, not relevant)
- **Цель:**
  1) определение релевантности заголовка статье,
  2) проверка адекватности тематического моделирования.
- **Этапы решения:**
  1) предобработка, разбиение коллекции для кросс-валидации,
  2) тематическое моделирование, кросс-валидация,
  3) сравнение с другими моделями (LR, SVM),
  4) добиться улучшения качества модели путём регуляризации.

Темы исследований, где есть открытые проблемы

- Устойчивость и полнота набора тем.
- Оптимизация параметров онлайнового алгоритма.
- Адаптивная оптимизация коэффициентов регуляризации.
- Эффективная инициализация матрицы Φ.
- Создание новых тем в потоке новостей.
- Автоматическое выделение терминов-словосочетаний.
- Тематическая сегментация.
- Тематические модели дистрибутивной семантики.
- Суммаризация тем.
- Автоматическое именование тем.
- Интеграция с анализом тональности и выявлением мнений.
- Интеграция с синтаксическими анализаторами.

Резюме

- Задача тематического моделирования некорректно поставлена, её решение не единственно и не устойчиво.
- Регуляризация — стандартный приём решения таких задач.
- Подход ARTM позволяет комбинировать регуляризаторы и строить тематические модели с требуемыми свойствами.
- Реализация — в проекте с открытым кодом BigARTM.
- Модель LDA — слишком слабый регуляризатор, не решает проблему неединственности и неустойчивости.
- Модель LDA лучше описывает вероятности редких слов, но для выявления тематики они как раз и не важны.
- Регуляризаторы и модальности — в следующих лекциях.