Многокритериальная регуляризация вероятностных тематических моделей коллекций текстовых документов

Воронцов Константин Вячеславович

ВЦ РАН • МФТИ

Семинар «Стохастический анализ в задачах» НМУ–МФТИ
19 апреля 2014

Содержание

Задачи матричного разложения

Дано: матрица $Z = \|z_{ij}\|_{n\times m}$.

Найти: матрицы $X = \|x_{it}\|_{n\times k}$ и $Y = \|y_{tj}\|_{k\times m}$ такие, что

$$\|Z - XY\|_D = \sum_{i=1}^{n}\sum_{j=1}^{m} D\Big(z_{ij}, \sum_t x_{it}y_{tj}\Big) \to \min_{X,Y}$$

Типы задач:

- различные нормы D: Фробениуса, KL-дивергенция,...
- неотрицательные матричные разложения: $x_{it} \geqslant 0,\ \ y_{tj} \geqslant 0$
- стохастические матричные разложения: $\sum_i x_{it} = 1,\ \ \sum_t y_{tj} = 1$
- разреженные данные: известны только $z_{ij},\ \ (i,j) \in \Omega$:

$$\sum_{(i,j)\in\Omega} D\Big(z_{ij}, \sum_t x_{it}y_{tj}\Big) \to \min_{X,Y}$$

Примеры прикладных задач матричного разложения

1. **Анализ данных жидкостной хроматографии**

$$z(t,\lambda) = \sum_i X_i(t)\, Y_i(\lambda)$$

дано: $z(t,\lambda)$ — выход сканирующего УФ-детектора;
найти: $X_i(t)$ — хроматограмма i-го вещества, t — время
$Y_i(\lambda)$ — спектр i-го вещества, λ — длина волны.

2. **Анализ данных ДНК-микрочипов**

$$I(p,k) = \sum_g a_{pg}\, C_{gk}$$

дано: $I(p,k)$ — интенсивность свечения p-й пробы на k-м чипе;
найти: a_{pg} — коэффициент сродства p-й пробы g-му гену,
C_{gk} — концентрация g-го гена на k-м чипе.

Примеры прикладных задач матричного разложения

3. **Коллаборативная фильтрация**

$$R_{iu} = \sum_t p_i(t) q_u(t)$$

дано: R_{iu} — рейтинги товаров i, поставленные пользователем u;
найти: $p_i(t)$ — профиль интересов товара i;
$q_u(t)$ — профиль интересов пользователя u.

4. **Тематическое моделирование текстовых коллекций**

$$p(w|d) = \sum_t p(w|t) p(t|d)$$

дано: $p(w|d)$ — частоты слов w в документах d;
найти: $p(w|t)$ — распределения слов w в темах t,
$p(t|d)$ — распределения тем t в документах d.

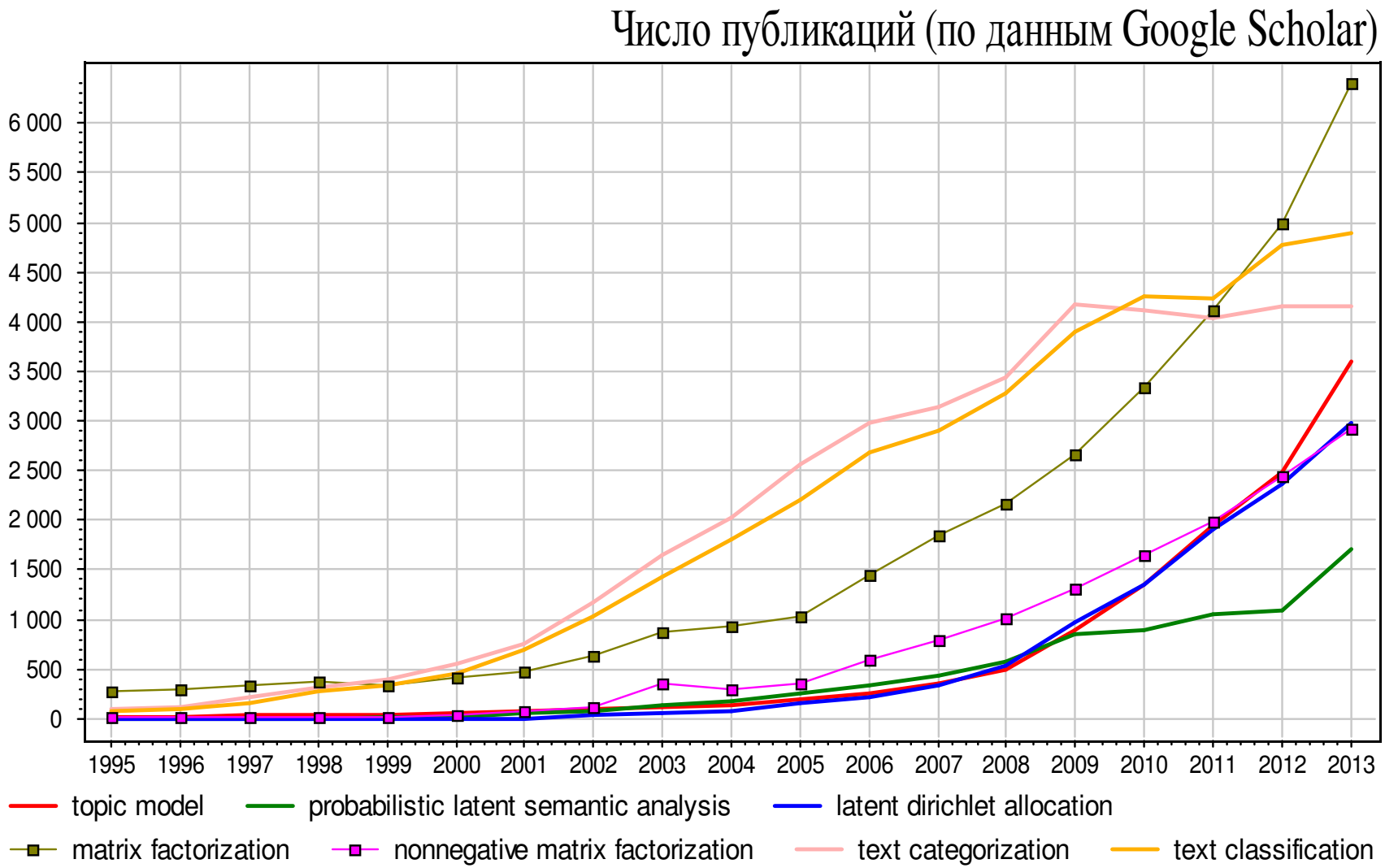
Цели вероятностного тематического моделирования (BTM)

- Тематический поиск документов $p(t|d)$ и объектов $p(t|x)$ по тексту любой длины или по любому объекту
- Категоризация, классификация, аннотирование, суммаризация, сегментация текстовых документов

Приложения:

- Поиск научной информации
- Выявление трендов и фронта исследований
- Поиск специалистов (expert search), рецензентов, проектов
- Анализ и агрегирование новостных потоков
- Рубрикация документов, изображений, видео, музыки
- Рекомендующие системы, коллаборативная фильтрация
- Аннотация генома и другие задачи биоинформатики
- Анализ дискретизированных биомедицинских сигналов

Матричные разложения и тематическое моделирование



Проблемы матричных разложений

- Неединственность решения

$$XY = (XS)(S^{-1}Y) = X'Y'$$

- Неустойчивость решения — следствие неединственности
- Сходимость к локальным экстремумам
- Сходимость к нестационарным точкам

Способы решения:

- Проблемно-ориентированные регуляризации
(будут рассмотрены для тематического моделирования)
- Подбор начальных приближений
- Модификации итерационных алгоритмов
(в следующем докладе Евгения Рябенко)

Задача вероятностного тематического моделирования (BTM)

Базовые предположения:

- каждое слово в документе связано с некоторой темой $t \in T$
- $D \times W \times T$ — дискретное вероятностное пространство
- выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$ случайная, независимая
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- гипотеза условной независимости: $p(w|d,t) = p(w|t)$

Вероятностная модель порождения документов:

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

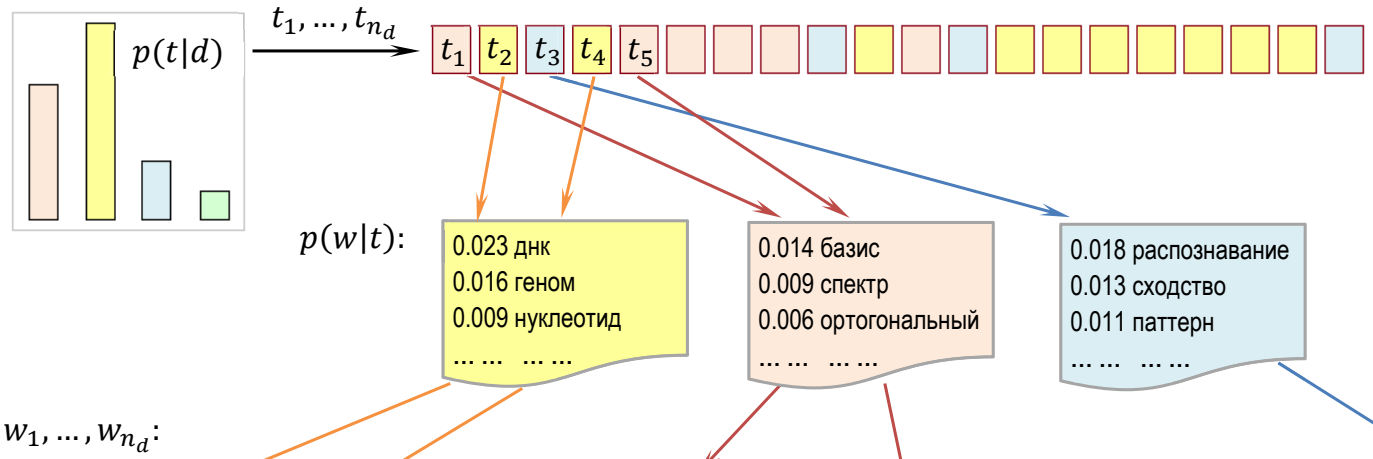
- $\phi_{wt} \equiv p(w|t)$ — распределение терминов в темах $t \in T$;
- $\theta_{td} \equiv p(t|d)$ — распределение тем в документах $d \in D$.

Прямая задача: по ϕ_{wt}, θ_{td} сгенерировать документ d.

Обратная задача: по $\hat{p}(w|d) \equiv \frac{n_{dw}}{n_d}$ найти параметры ϕ_{wt}, θ_{td}.

Прямая задача BTM: порождение коллекции

Порождение слов документа d из $p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Обратная задача BTM: стохастическое матричное разложение

Дано:

W — словарь, множество терминов (слов, словосочетаний),

D — множество (коллекция, корпус) текстовых документов,

n_{dw} — сколько раз термин $w \in W$ встретился в документе $d \in D$.

$\hat{p}(w|d) \equiv \frac{n_{dw}}{n_d}$ — эмпирические частоты терминов в документах

Найти: матричное разложение

$$\underset{W\times D}{F} \approx \underset{W\times T}{\Phi} \cdot \underset{T\times D}{\Theta}$$

$F = \left\|\hat{p}(w|d)\right\|_{W\times D}$ — известная матрица частот,

$\Phi = \left\|\phi_{wt}\right\|_{W\times T}$ — искомая матрица терминов тем $\phi_{wt} = p(w|t)$,

$\Theta = \left\|\theta_{td}\right\|_{T\times D}$ — искомая матрица тем документов $\theta_{td} = p(t|d)$.

Принцип максимума правдоподобия

Задача: максимизировать логарифм правдоподобия

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

при ограничениях неотрицательности и нормировки

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

Это взвешенная сумма дивергенций Кульбака–Лейблера между

$\hat{p}(w|d)=\frac{n_{dw}}{n_d}$ — эмпирическими частотами и

$p(w|d)=\sum_t \phi_{wt}\theta_{td}$ — тематическими моделями документов:

$$\mathrm{KL}(\hat{p}\|p)=\sum_{d\in D} n_d\sum_{w\in d}\hat{p}(w|d)\ln\frac{\hat{p}(w|d)}{p(w|d)}\to\min_{\Phi,\Theta},$$

Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Теорема

Если Φ, Θ — решение задачи максимизации правдоподобия, то оно удовлетворяет системе уравнений

$$\begin{cases} \text{E-шаг:} \quad p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum\limits_{s \in T} \phi_{ws}\theta_{sd}}; \quad \text{— вспомогательные переменные} \\ \text{M-шаг:} \quad \phi_{wt} = \dfrac{n_{wt}}{n_t}; \quad n_{wt} = \sum\limits_{d \in D} n_{dw} p_{tdw}; \quad n_t = \sum\limits_{w \in W} n_{wt}; \\ \qquad\quad \theta_{td} = \dfrac{n_{td}}{n_d}; \quad n_{td} = \sum\limits_{w \in d} n_{dw} p_{tdw}; \quad n_d = \sum\limits_{t \in T} n_{td}; \end{cases}$$

EM-алгоритм — чередование E- и M-шага до сходимости.
Это решение системы уравнений методом простых итераций.

✓ *Идея на будущее: можно использовать и другие методы!*

Вероятностная интерпретация шагов EM-алгоритма

E-шаг — это формула Байеса:

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

$n_{dwt} = n_{dw}p(t|d,w)$ — оценка числа троек (d,w,t) в коллекции

M-шаг — это частотные оценки условных вероятностей:

$$\phi_{wt} = \frac{n_{wt}}{n_t} \equiv \frac{\sum\limits_{d\in D} n_{dwt}}{\sum\limits_{d\in D}\sum\limits_{w\in d} n_{dwt}}, \qquad \theta_{td} = \frac{n_{td}}{n_d} \equiv \frac{\sum\limits_{w\in d} n_{dwt}}{\sum\limits_{w\in W}\sum\limits_{t\in T} n_{dwt}},$$

Краткая запись через знак пропорциональности $\propto$:

$$p(t|d,w) \propto \phi_{wt}\theta_{td}; \qquad \phi_{wt} \propto n_{wt}; \qquad \theta_{td} \propto n_{td};$$

Рациональный EM-алгоритм

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 документов $d \in D$ и всех слов $w \in d$
    $p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$ для всех $t \in T$;
    $n_{wt}, n_{td}, n_t, n_d \mathrel{+}= n_{dw} p_{tdw}$ для всех $t \in T$;
  $\phi_{wt} := n_{wt}/n_t$ для всех $w \in W$, $t \in T$;
  $\theta_{td} := n_{td}/n_d$ для всех $d \in D$, $t \in T$;

Эвристики

Разреживание распределений $p(t|d,w)$

- пропорциональное распределение без разреживания
- сэмплирование Гиббса: $t \sim p(t|d,w)$ для каждой позиции w_i
- сэмплирование: $t \sim p(t|d,w)$ для каждого слова (d,w)
- максимизация (оптимальный байесовский классификатор): $t = \arg\max_t p(t|d,w)$ для каждого слова (d,w)

Чередование сэмплирования и максимизации приводит к лучшему локальному максимуму правдоподобия [Д. Елшин]

Частое обновление параметров ϕ_{wt}, θ_{td}:

- после каждого прохода всей коллекции
- после каждого документа
- после каждого слова (самая быстрая сходимость)

Онлайновый EM-алгоритм (для больших коллекций)

инициализировать ϕ_{wt} для всех $w \in W,\ t \in T$;
$n_{wt} := 0,\ n_t := 0$ для всех $w \in W,\ t \in T$;
для всех пачек документов $D_j,\ j = 1, \ldots, J$
  $\tilde{n}_{wt} := 0,\ \tilde{n}_t := 0$ для всех $w \in W,\ t \in T$;
  для всех документов d из пачки D_j
    инициализировать θ_{td} для всех $t \in T$;
    повторять
      $p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$ для всех $w \in d,\ t \in T$;
      $\theta_{td} := \dfrac{1}{n_d} \sum\limits_{w \in d} n_{dw} p_{tdw}$ для всех $t \in T$;
    пока θ_d не сойдётся;
    $\tilde{n}_{wt}, \tilde{n}_t \mathrel{+}= n_{dw} p_{tdw}$ для всех $w \in d,\ t \in T$;
  $n_{wt} := \rho_j n_{wt} + \tilde{n}_{wt};\ \ n_t := \rho_j n_t + \tilde{n}_t$ для всех $w \in W,\ t \in T$;
  $\phi_{wt} := n_{wt} / n_t$ для всех $w \in W,\ t \in T$;

От PLSA к LDA (Latent Dirichlet Allocation)

Недостатки PLSA:

- Плохо оцениваются $p(w|t)$ и $p(t|d)$ редких слов и тем
- Переобучение при малых $\frac{|D|}{|T|}$

Способы устранения этих недостатков:

- Увеличить объём коллекции
- Наложить «естественные» ограничения на Φ и Θ

David Blei, Andrew Ng, Michael Jordan. Latent Dirichlet allocation Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.

Yi Wang. Distributed Gibbs Sampling of Latent Dirichlet Allocation: The Gritty Details. 2011.

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.

Латентное размещение Дирихле, LDA

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

Гипотеза об априорных распределениях Дирихле:

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

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

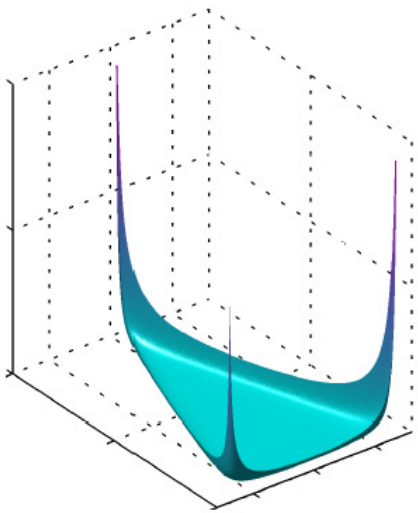
- $\phi_t = (\phi_{wt})_{w\in W} \in \mathbb{R}^{|W|}$ — случайные векторы из распределения Дирихле с параметром $\beta \in \mathbb{R}^{|W|}$:

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

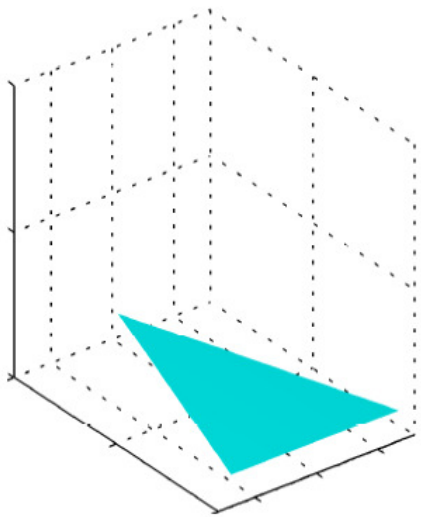
Почему именно распределение Дирихле?

– Является сопряжённым к мультиномиальному распределению
– Может порождать сглаженные или разреженные векторы
– Неплохо описывает кластерные структуры на симплексе

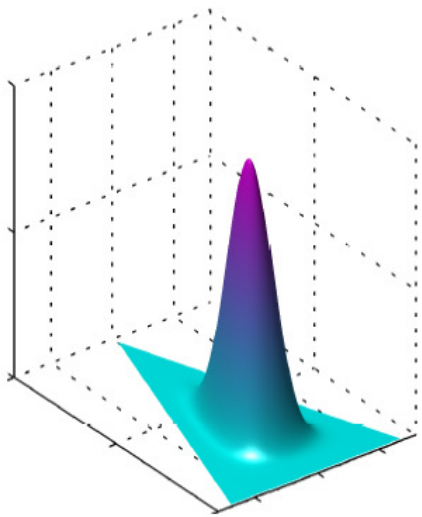
Пример. $\mathrm{Dir}(\theta|\alpha)$ при $|T| = 3$, $\theta, \alpha \in \mathbb{R}^3$:



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



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



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

Байесовская оценка параметров $\theta_{td} \equiv p(t|d)$

Пусть темы слов в документах $d \in D$ выбираются из θ_d:

$$X_d = \{t_1, \ldots, t_{n_d}\} \sim \theta_d.$$

Тогда вероятность встретить каждую из тем t ровно n_{td} раз подчиняется мультиномиальному распределению:

$$p(X_d|\theta_d) = \text{Mult}(n_{1d}, \ldots, n_{Td}|\theta_d) = \frac{n_d!}{\prod_t n_{td}!} \prod_t \theta_{td}^{n_{td}}.$$

Если предположить, что $\theta_d \sim \text{Dir}(\alpha)$, то по формуле Байеса апостериорное распределение также из $\text{Dir}(\alpha')$, $\alpha'_t = \alpha_t + n_{td}$:

$$p(\theta_d|X_d) = \frac{p(X_d|\theta_d)\,\text{Dir}(\theta_d|\alpha)}{\int p(X_d|\theta)\,\text{Dir}(\theta|\alpha)\,d\theta} \propto \prod_t \theta_{td}^{n_{td}} \theta_{td}^{\alpha_t - 1} = \text{Dir}(\theta_d; \alpha').$$

Распределение Дирихле — *сопряжённое* к мультиномиальному, что упрощает байесовское оценивание параметров ϕ_{wt} и θ_{td}.

Байесовская оценка параметров $\theta_{td} \equiv p(t|d)$

Оценка θ_{td} при априорном распределении:

$$\mathrm{E}p(t|d,\alpha) = \int \theta_{td}\mathrm{Dir}(\theta_d|\alpha)\,d\theta_d = \frac{\alpha_t}{\alpha_0}.$$

Пусть известна выборка тем $X_d = \{t_1,\ldots,t_{n_d}\} \sim \theta_d$.
Оценка θ_{td} при апостериорном распределении:

$$\mathrm{E}p(t|d,X_d,\alpha) = \int \theta_{td}\mathrm{Dir}(\theta_d|\alpha')\,d\theta_d = \frac{n_{td}+\alpha_t}{\sum\limits_{t'} n_{t'd}+\alpha_{t'}} = \frac{n_{td}+\alpha_t}{n_d+\alpha_0},$$

n_{td} — сколько раз слово документа d было отнесено к теме t,
n_d — длина документа в словах.

Замечание. Эта оценка переходит в МП-оценку при $\alpha_t \equiv 0$, хотя при $\alpha_t = 0$ распределение Дирихле не определено.

Байесовская оценка параметров $\phi_{wt} \equiv p(w|t)$

Оценка ϕ_{wt} при априорном распределении:

$$\mathrm{E}p(w|t,\beta) = \int \phi_{wt} \mathrm{Dir}(\phi_t|\beta)\, d\phi_t = \frac{\beta_w}{\beta_0}.$$

Коллекция порождается двумя распределениями $p(t|d)$, $p(w|t)$.
Часть коллекции, порождённая темой t:

$$X_t = \big\{(d,w,t)\colon d \in D,\ w \sim \phi_t\big\}.$$

Апостериорное распределение для ϕ_t по формуле Байеса:

$$p(\phi_t|X_t,\beta) = \frac{p(X_t|\phi_t)\,\mathrm{Dir}(\phi_t|\beta)}{\int p(X_t|\phi)\,\mathrm{Dir}(\phi|\beta)\,d\phi} = \mathrm{Dir}(\phi_t|\beta'), \quad \beta'_w = \beta_w + n_{wt}.$$

Оценка ϕ_{wt} через апостериорное распределение:

$$\mathrm{E}p(w|t,X_d,\beta) = \int \phi_{wt} \mathrm{Dir}(\phi_t|\beta')\, d\phi_t = \frac{n_{wt}+\beta_w}{n_t+\beta_0}.$$

Главное отличие LDA от PLSA

Оценки условных вероятностей $\phi_{wt} \equiv p(w|t)$, $\theta_{td} \equiv p(t|d)$:

- в PLSA — несмещённые оценки максимума правдоподобия:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad \theta_{td} = \frac{n_{td}}{n_d}$$

- в LDA — сглаженные байесовские оценки:

$$\phi_{wt} = \frac{n_{wt} + \beta_w}{n_t + \beta_0}, \qquad \theta_{td} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}.$$

Различия между LDA и PLSA практически исчезают
— на больших данных
— при игнорировании редких слов (робастные алгоритмы)

Воронцов К.В., Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных, 2013. — Т. 1, № 6. — С. 657–686.

Недостатки LDA

1. распределение Дирихле удобно математически, но не имеет лингвистических обоснований
2. сглаживание вместо разреживания
3. байесовский вывод требует интегрирования по пространству параметров модели, которое только в базовом варианте LDA элементарно
4. построение композитных и многофункциональных моделей становится громоздкой математической задачей
5. на больших данных почти нет различий между LDA и PLSA
6. переобучение PLSA связано только с редкими словами, и это отнюдь не главный недостаток PLSA (плохо поняли суть проблемы и боремся не с тем врагом)

Неустойчивость! Эксперимент на модельных данных

Модельные коллекции порождаются заданными матрицами Φ_0 и Θ_0 при $|D| = 500$, $|W| = 1000$, $|T| = 30$, $n_d \in [100, 600]$.

Отклонение восстановленных распределений $p(i|j)$ от исходных модельных распределений $p_0(i|j)$ измеряются средним расстоянием Хеллингера:

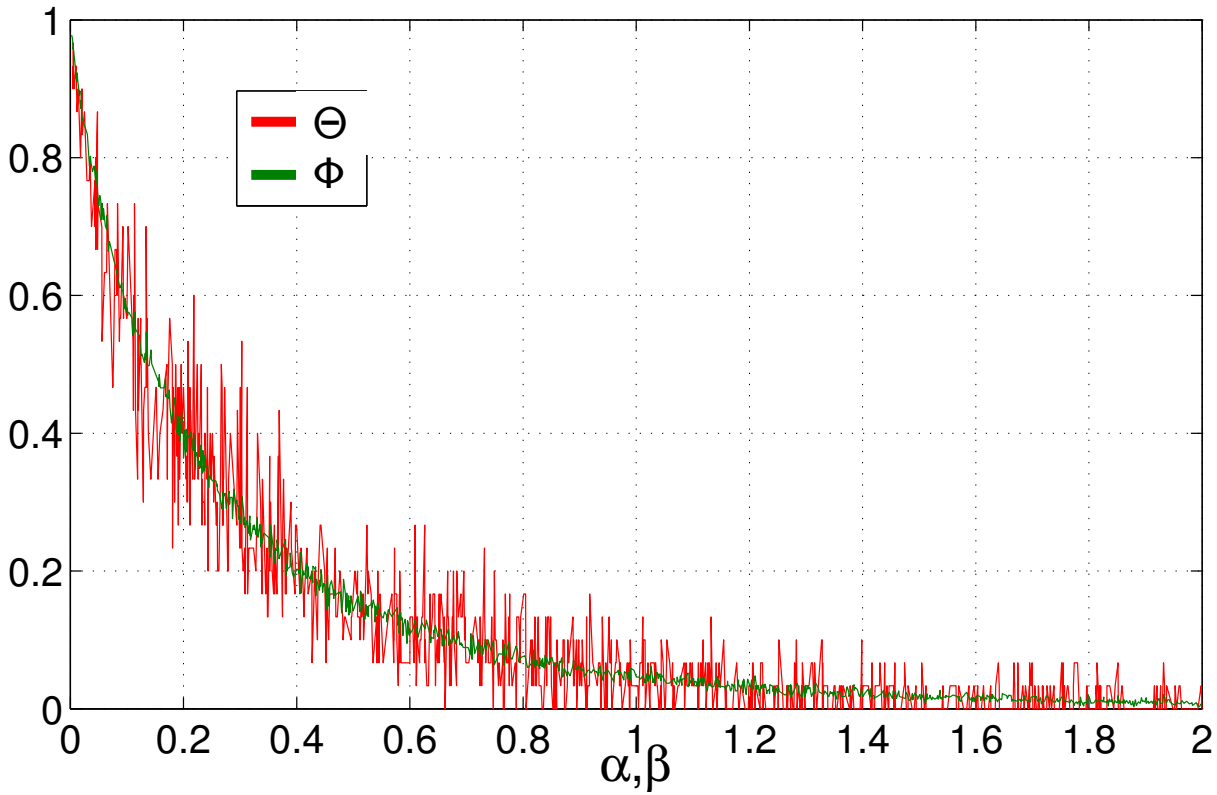
$$H(p, p_0) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \left(\sqrt{p(i|j)} - \sqrt{p_0(i|j)} \right)^2},$$

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

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

Генерация модельных данных различной степени разреженности

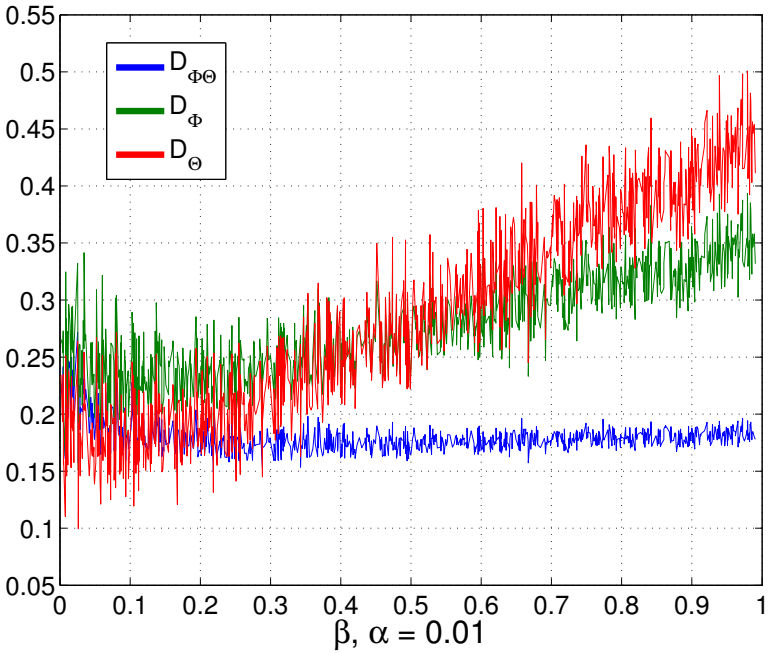
Зависимость разреженности (доли почти нулевых элементов) распределений $\theta_d^0 \sim \mathrm{Dir}(\alpha)$ и $\phi_t^0 \sim \mathrm{Dir}(\beta)$ от параметров α и β симметричного распределения Дирихле:



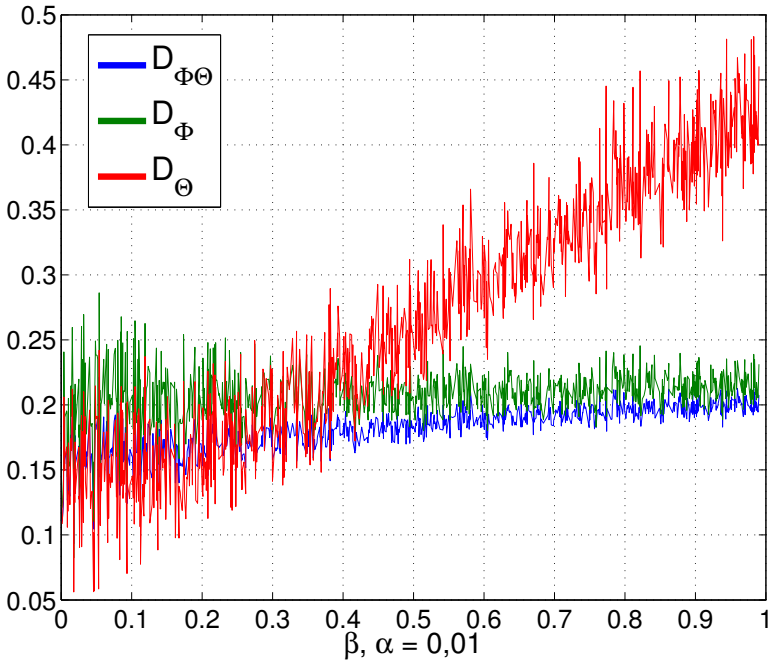
Эксперимент: неустойчивость восстановления Φ, Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Φ_0

PLSA



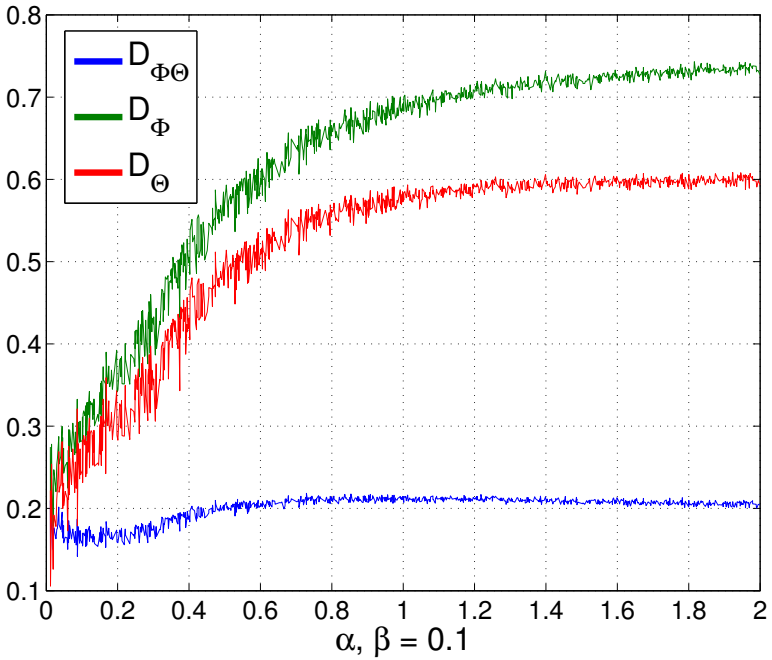
LDA



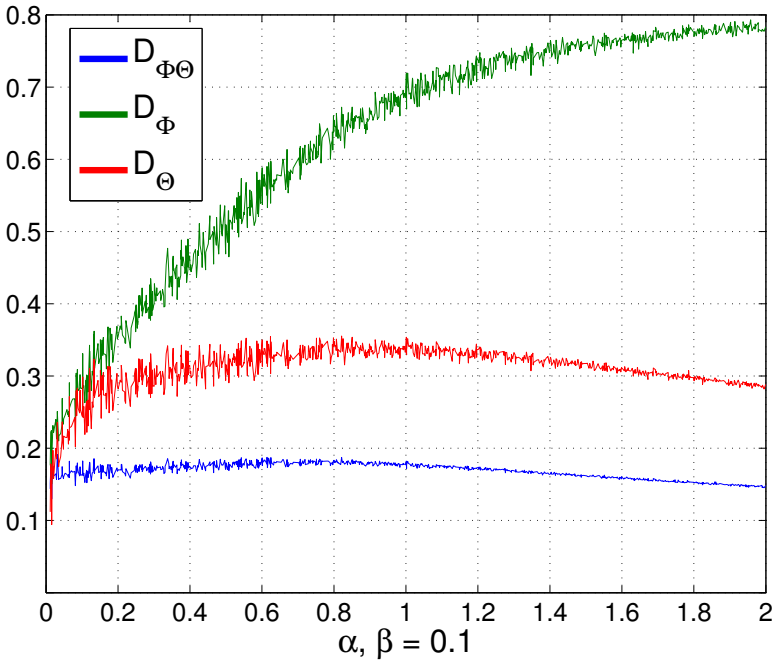
Эксперимент: неустойчивость восстановления Φ, Θ

Зависимость точности восстановления матриц Φ, Θ и ΦΘ от разреженности матрицы Θ_0

PLSA



LDA



Выводы

1. Произведение $\Phi\Theta$ восстанавливается устойчиво, точность восстановления не зависит от разреженности исходных модельных данных Φ_0, Θ_0
2. Матрицы Φ, Θ восстанавливаются неустойчиво, результат зависит от случайной инициализации
3. Методы PLSA и LDA одинаково неустойчивы (сглаживание не спасает от неединственности)
4. Устойчивое восстановление матриц Φ, Θ происходит только при сильной разреженности (более 80% нулей)

Реализация экспериментов:
Виталий Глушаченков. Магистерская диссертация. МФТИ, 2013.
Михаил Колупаев. Курсовая работа. ВШЭ, 2013.

Аддитивная регуляризация тематической модели

Пусть, наряду с правдоподобием, требуется максимизировать ещё n критериев $R_i(\Phi, \Theta)$, $i = 1, \ldots, n$ — *регуляризаторов*.

Метод многокритериальной оптимизации — *скаляризация*.

Задача: максимизировать регуляризованное правдоподобие

$$\underbrace{\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}}_{\text{log-likelihood } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

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

где $\tau_i > 0$ — *коэффициенты регуляризации*.

EM-алгоритм с регуляризацией M-шага

Теорема

Если Φ, Θ — решение задачи максимизации регуляризованного правдоподобия, то оно удовлетворяет системе уравнений

$$\begin{cases} \text{E-шаг:} \quad p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T}\phi_{ws}\theta_{sd}}; \\ \text{M-шаг:} \quad \phi_{wt} \propto \left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\right)_+; \quad n_{wt} = \sum\limits_{d\in D} n_{dw}p_{tdw}; \\ \qquad\quad \theta_{td} \propto \left(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\right)_+; \quad n_{td} = \sum\limits_{w\in d} n_{dw}p_{tdw}; \end{cases}$$

При $R(\Phi, \Theta) = 0$ это формулы EM-алгоритма для PLSA.

Напоминания. 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, $j = 1, \ldots, k$:

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

Доказательство Теоремы о регуляризации M-шага

1. Условия KKT для ϕ_{wt}:

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

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. Учтём ограничение $\phi_{wt} \geqslant 0$ и предположение $\lambda_t > 0$:

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

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

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

5. Подставим λ_t из (4) в (3), получим требуемое. ■

EM-алгоритм с регуляризацией E-шага

Теорема

Если регуляризатор зависит от Φ, Θ через $p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$,

$$R(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} r(p_{1dw}, \ldots, p_{Tdw}),$$

то решение задачи максимизации регуляризованного правдоподобия удовлетворяет системе уравнений

$$\begin{cases} \text{E-шаг:} \quad \tilde{p}_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}} \left(1 + \dfrac{\partial r}{\partial p_{tdw}} - \sum_{s \in T} p_{sdw} \dfrac{\partial r}{\partial p_{sdw}}\right); \\ \text{M-шаг:} \quad \phi_{wt} \propto \left(\sum_{d \in D} n_{dw} \tilde{p}_{tdw}\right)_+; \quad \theta_{td} \propto \left(\sum_{w \in d} n_{dw} \tilde{p}_{tdw}\right)_+; \end{cases}$$

Обобщения BTM и учёт дополнительной информации

- Разреживание $p(w|t)$, $p(t|d)$, $p(t|d,w)$
- Разделение слов на тематические и функциональные
- Оптимизация числа тем T
- Построение иерархии тем
- Выделение словосочетаний, n-граммные модели
- Учёт ключевых слов по темам (частичное обучение)
- Учёт тезаурусов или онтологий предметных областей
- Учёт структуры документа: предложения, разделы
- Учёт метаданных: дата/время, авторы, источник, и т.д.
- Учёт цитат и/или гиперссылок: исходящие, входящие
- Учёт рубрикации или иной классификации документов
- Учёт связей с изображениями, именами, пользователями
- Мультиязычные тематические модели

Напоминания. Дивергенция Кульбака–Лейблера

Функция расстояния между распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$:

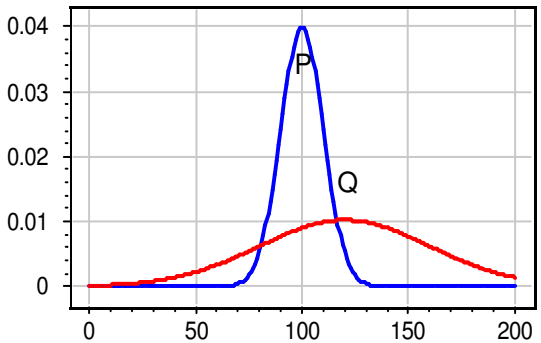
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

1. $\mathrm{KL}(P\|Q) \geqslant 0$; $\quad \mathrm{KL}(P\|Q) = 0 \;\Leftrightarrow\; P = Q$;

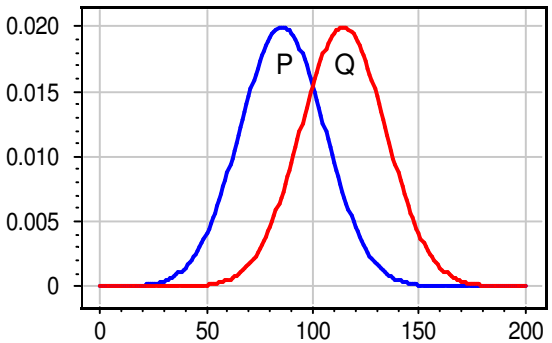
2. Минимизация KL эквивалентна максимизации правдоподобия:

$$\mathrm{KL}(P\|Q(\alpha)) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i(\alpha)} \to \min_\alpha \quad \Longleftrightarrow \quad \sum_{i=1}^n p_i \ln q_i(\alpha) \to \max_\alpha.$$

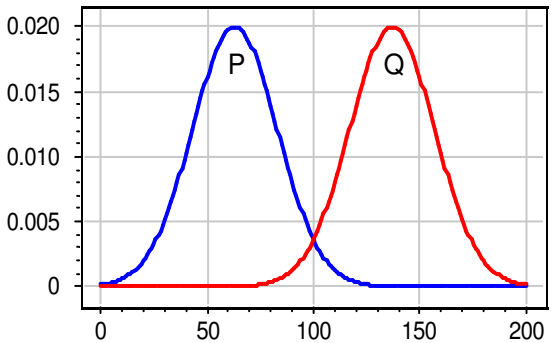
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



$\mathrm{KL}(P\|Q) = 0.442$
$\mathrm{KL}(Q\|P) = 2.966$



$\mathrm{KL}(P\|Q) = 0.444$
$\mathrm{KL}(Q\|P) = 0.444$



$\mathrm{KL}(P\|Q) = 2.969$
$\mathrm{KL}(Q\|P) = 2.969$

Регуляризатор №1: Сглаживание (совпадает с LDA)

Гипотеза сглаженности:
распределения ϕ_{wt} близки к заданным распределениям β_w
распределения θ_{td} близки к заданным распределениям α_t

$$\sum_{t\in T} \mathrm{KL}_w(\beta_w \| \phi_{wt}) \to \min_{\Phi}; \qquad \sum_{d\in D} \mathrm{KL}_t(\alpha_t \| \theta_{td}) \to \min_{\Theta}.$$

Максимизируем сумму этих регуляризаторов:

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

Подставляем, получаем формулы M-шага LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0\beta_w, \qquad \theta_{td} \propto n_{td} + \alpha_0\alpha_t.$$

Этого вы не найдёте в *D.Blei, A.Ng, M.Jordan.* Latent Dirichlet allocation // Journal of Machine Learning Research, 2003. — Vol. 3. — Pp. 993–1022.

Регуляризатор №2: Частичное обучение (обобщение LDA)

Гипотеза: имеется дополнительная информация от экспертов:
1) списки тем $T_d \subset T$ для некоторых документов $d \in D_0$,
2) списки терминов $W_t \subset W$ для некоторых тем $t \in T_0$.

ϕ^0_{wt} — распределение, равномерное на W_t
θ^0_{td} — распределение, равномерное на T_d

Максимизируем сумму этих регуляризаторов:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in T_0}\sum_{w\in W_t} \phi^0_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d\in D_0}\sum_{t\in T_d} \theta^0_{td} \ln \theta_{td} \to \max$$

Подставляем, получаем обобщение LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0 \phi^0_{wt} \qquad \theta_{td} \propto n_{td} + \alpha_0 \theta^0_{td}$$

Nigam K., McCallum A., Thrun S., Mitchell T. Text classification from labeled and unlabeled documents using EM // Machine Learning, 2000, no. 2–3.

Регуляризатор №2: Частичное обучение (второе обобщение LDA)

Гипотеза: вместо логарифма можно взять любую монотонно возрастающую функцию μ, в том числе $\mu(z) = z$
(максимизируется сумма ковариаций $\text{cov}(\phi_t^0, \phi_t)$, $\text{cov}(\theta_d^0, \theta_d)$)

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W_t} \phi_{wt} + \alpha_0 \sum_{d \in D_0} \sum_{t \in T_d} \theta_{td} \to \max.$$

Подставляем, получаем ещё одно обобщение LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0 \phi_{wt}, \; w \in W_t \qquad \theta_{td} \propto n_{td} + \alpha_0 \theta_{td}, \; t \in T_d.$$

Преимущество ковариационного регуляризатора:
Если θ_{td}^0 равномерно на T_d, то ковариация не накладывает ограничений на распределение θ_{td} между темами из T_d.

Регуляризатор №3: Разреживание (третье обобщение LDA)

Гипотеза разреженности: среди ϕ_{wt}, θ_{td} много нулей.

Чем сильнее разрежено распределение, тем ниже его энтропия.
Максимальной энтропией обладает равномерное распределение.

Максимизируем дивергенцию между распределениями β_w, α_t (равномерными?) и искомыми распределениями ϕ_{wt}, θ_{td}:

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

Подставляем, получаем «анти-LDA»:

$$\phi_{wt} \propto \left(n_{wt} - \beta_0\beta_w\right)_+, \qquad \theta_{td} \propto \left(n_{td} - \alpha_0\alpha_t\right)_+.$$

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions.

Регуляризатор №4: Декорреляция

Гипотеза некоррелированности тем:
чем различнее темы, тем лучше они интерпретируются.

Минимизируем ковариации между вектор-столбцами ϕ_t:

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

Подставляем, получаем ещё один вариант разреживания — постепенное контрастирование строк матрицы Φ:

$$\phi_{wt} \propto \left(n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \right)_+ .$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp. 224–228.

Регуляризатор №5: Удаление незначимых тем

Гипотеза: если тема собрала мало слов, то она не нужна.

Разреживаем распределение $p(t) = \sum_d p(d)\theta_{td}$, максимизируя KL-дивергенцию между $p(t)$ и равномерным распределением:

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

Подставляем, получаем:

$$\theta_{td} \propto \left(n_{td} - \tau \frac{n_d}{n_t} \theta_{td} \right)_+.$$

Эффект:
строки матрицы Θ могут целиком обнуляться для тем t, собравших мало слов по коллекции, $n_t = \sum_d \sum_w n_{dwt}$.

Регуляризатор №6: Разреживание распределений $p(t|d,w)$

Гипотеза разреженности распределений $p(t|d,w)$:
в документе слово может относиться только к одной теме.

Максимизируем KL-дивергенцию между $\hat{p}(t) = \frac{1}{|T|}$ и $p(t|d,w)$:

$$R(\Phi,\Theta) = -\tau \sum_{d\in D} \sum_{w\in d} n_{dw} \frac{1}{|T|} \sum_{t\in T} \ln p_{tdw}.$$

Подставляем, получаем формулы модифицированного E-шага:

$$p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{s\in T} \phi_{ws}\theta_{sd}}(1+\tau) - \frac{\tau}{|T|}.$$

Эффект:
если $p(t|w) < \frac{1}{|T|}$, то ϕ_{wt} уменьшается;
если $p(t|d) < \frac{1}{|T|}$, то θ_{td} уменьшается.

Регуляризатор №7: Максимизация когерентности тем

Гипотеза: тема лучше интерпретируется, если она содержит *когерентные* (часто встречающиеся рядом) слова $u, w \in W$.

Пусть C_{uw} — оценка когерентности, например $\hat{p}(w|u) = \frac{N_{uw}}{N_u}$.
Согласуем ϕ_{wt} с оценками $\hat{p}(w|t)$ по когерентным словам,

$$\hat{p}(w|t) = \sum_u p(w|u)p(u|t) = \frac{1}{n_t}\sum_u C_{uw} n_{ut};$$

$$R(\Phi,\Theta) = \tau \sum_{t\in T} n_t \sum_{w\in W} \hat{p}(w|t) \ln \phi_{wt} \to \max.$$

Подставляем, получаем ещё один вариант сглаживания:

$$\phi_{wt} \propto n_{wt} + \tau \sum_{u \in W\setminus w} C_{uw} n_{ut}.$$

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Empirical Methods in Natural Language Processing, EMNLP-2011. — Pp. 262–272.

Регуляризатор №8: Связи между документами

Гипотеза: чем больше n_{dc} — число ссылок из d на c, тем более близки тематики документов d и c.

Минимизируем ковариации между вектор-столбцами связанных документов θ_d, θ_c:

$$R(\Phi, \Theta) = \tau \sum_{d,c \in D} n_{dc} \text{cov}(\theta_d, \theta_c) \to \max,$$

Подставляем, получаем ещё один вариант сглаживания:

$$\theta_{td} \propto n_{td} + \tau \theta_{td} \sum_{c \in D} n_{dc} \theta_{tc}.$$

Dietz L., Bickel S., Scheffer T. Unsupervised prediction of citation influences // ICML 2007. — Pp. 233–240.

Регуляризатор №9: Классификация документов

Пусть C — множество классов документов
(категории, авторы, ссылки, годы, пользователи,...)

Гипотеза:
классификация документа d объясняется его темами:

$$p(c|d) = \sum_{t \in T} p(c|t)p(t|d) = \sum_{t \in T} \psi_{ct}\theta_{td}.$$

Минимизируем дивергенцию между моделью $p(c|d)$
и «эмпирической частотой» классов в документах m_{dc}:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct}\theta_{td} \to \max.$$

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 1–2.

Регуляризатор №9: Классификация документов

EM-алгоритм дополняется оцениванием параметров ψ_{ct}.

E-шаг. По формуле Байеса:

$$p(t|d,w) = \frac{\phi_{wt}\theta_{td}}{\sum_{s\in T}\phi_{ws}\theta_{sd}} \qquad p(t|d,c) = \frac{\psi_{ct}\theta_{td}}{\sum_{s\in T}\psi_{cs}\theta_{sd}}$$

M-шаг. Максимизация регуляризованного правдоподобия:

$$\phi_{wt} \propto n_{wt} \qquad n_{wt} = \sum_{d\in D} n_{dw} p(t|d,w)$$

$$\theta_{td} \propto n_{td} + \tau m_{td} \qquad n_{td} = \sum_{w\in W} n_{dw} p(t|d,w) \qquad m_{td} = \sum_{c\in C} m_{dc} p(t|d,c)$$

$$\psi_{ct} \propto m_{ct} \qquad m_{ct} = \sum_{d\in D} m_{dc} p(t|d,c)$$

Регуляризатор №10: Категоризация документов

Снова регуляризатор для классификации:

$$R(\Psi,\Theta) = \tau \sum_{d\in D}\sum_{c\in C} m_{dc} \ln \sum_{t\in T} \psi_{ct}\theta_{td} \to \max$$

Недостаток: для «эмпирической частоты классов» приходится необоснованно брать равномерное распределение:

$$m_{dc} = n_d \tfrac{1}{|C_d|}[c \in C_d]$$

Ковариационный регуляризатор:

$$R(\Psi,\Theta) = \tau \sum_{d\in D}\sum_{c\in C} m_{dc} \sum_{t\in T} \psi_{ct}\theta_{td} \to \max$$

приводит к естественному аналитическому решению

$$\psi_{ct} = \big[c = c^*(t)\big], \quad c^*(t) = \arg\max_{c\in C} \sum_{d\in D} m_{dc}\theta_{td}$$

Эффект: Каждая категория c распадается на свои темы.

Регуляризатор №11: Динамическая тематическая модель

Y — моменты времени (например, годы публикаций),
$y(d)$ — метка времени документа d,
$D_y \subset D$ — все документы, относящиеся к моменту $y \in Y$.

Гипотеза 1: распределение $p(t|y) = \sum_{d \in D_y} \theta_{td} p(d)$ разрежено:

$$R_1(\Theta) = -\tau_1 \sum_{y \in Y} \sum_{t \in T} \ln p(t|y) \to \max.$$

Эффект — разреживание тем t с малым $p(t|y(d))$:

$$\theta_{td} \propto \left(n_{td} - \tau_1 \frac{\theta_{td} p(d)}{p(t|y(d))} \right)_+.$$

Гипотеза 2: $p(t|y)$ меняются плавно, с редкими скачками:

$$R_2(\Theta) = -\tau_2 \sum_{y \in Y} \sum_{t \in T} \left| p(t|y) - p(t|y-1) \right| \to \max.$$

Этапы решения прикладных задач с помощью APTM

1. понимание требований
2. формализация требований с помощью регуляризаторов (при необходимости реализация новых регуляризаторов)
3. определение критериев качества модели (при необходимости реализация новых критериев)
4. подбор траектории регуляризации
5. эксперименты

APTM позволяет сосредоточиться на формализации требований, а алгоритм строить автоматически:

Подбор траектории регуляризации

Пусть задана линейная комбинация регуляризаторов:

$$R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)$$

Задача: выбрать вектор коэффициентов $\tau = (\tau_i)_{i=1}^{n}$

Ближайший аналог: «Regularization Path» в задачах регрессии с L_1- и L_2-регуляризацией (Elastic Net)

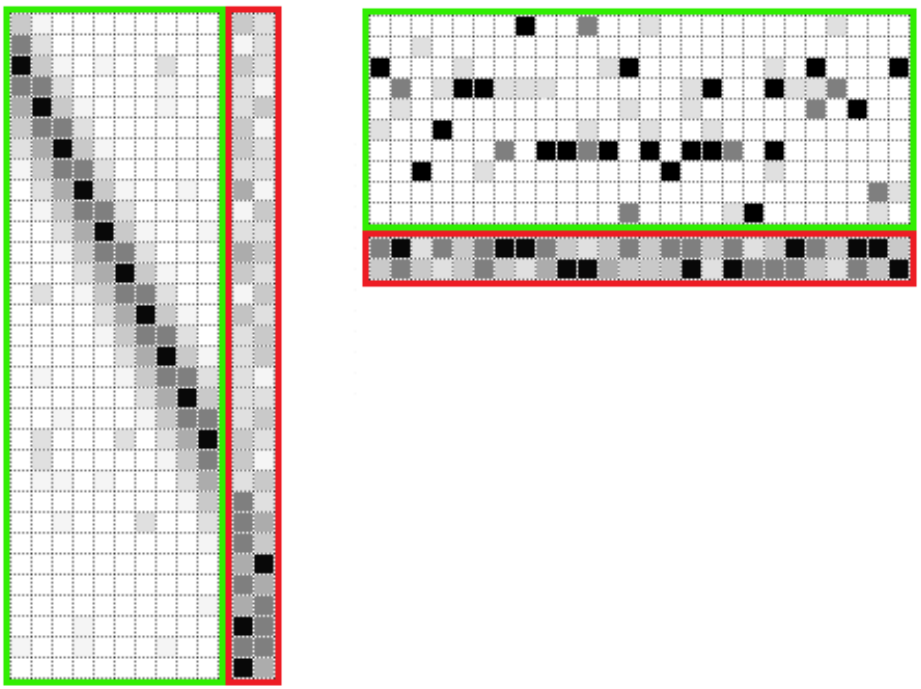
Общие соображения о выборе траектории регуляризации:

- мониторить много критериев качества в ходе итераций,
- усиливать регуляризаторы постепенно,
- сначала достичь сходимости нерегуляризованного PLSA,
- чередовать регуляризаторы, если они мешают друг другу,
- ослаблять регуляризаторы, когда их цель уже достигнута.

Гипотеза о структуре интерпретируемых тем

1. Предметные темы разреженные, существенно различные, имеют **ядро**, состоящее из терминов предметной области.
2. Фоновые темы плотные, содержат слова общей лексики.

$\Phi_{W \times T}$ $\Theta_{T \times D}$



Комбинирование разреживания, сглаживания и декорреляции

Задача: улучшить интерпретируемость, не ухудшив перплексию

Набор регуляризаторов:

№1 сглаживание фоновых тем — столбцов Φ, строк Θ

№3 разреживание предметных тем — столбцов Φ, строк Θ

№4 декоррелирование предметных тем — столбцов Φ

№5 удаление незначимых тем — строк Θ

Данные: NIPS (Neural Information Processing System)

- $|D| = 1566$ статей конференции NIPS на английском языке;
- суммарной длины $n \approx 2.3 \cdot 10^6$,
- словарь $|W| \approx 1.3 \cdot 10^4$.
- контрольная коллекция: $|D'| = 174$.

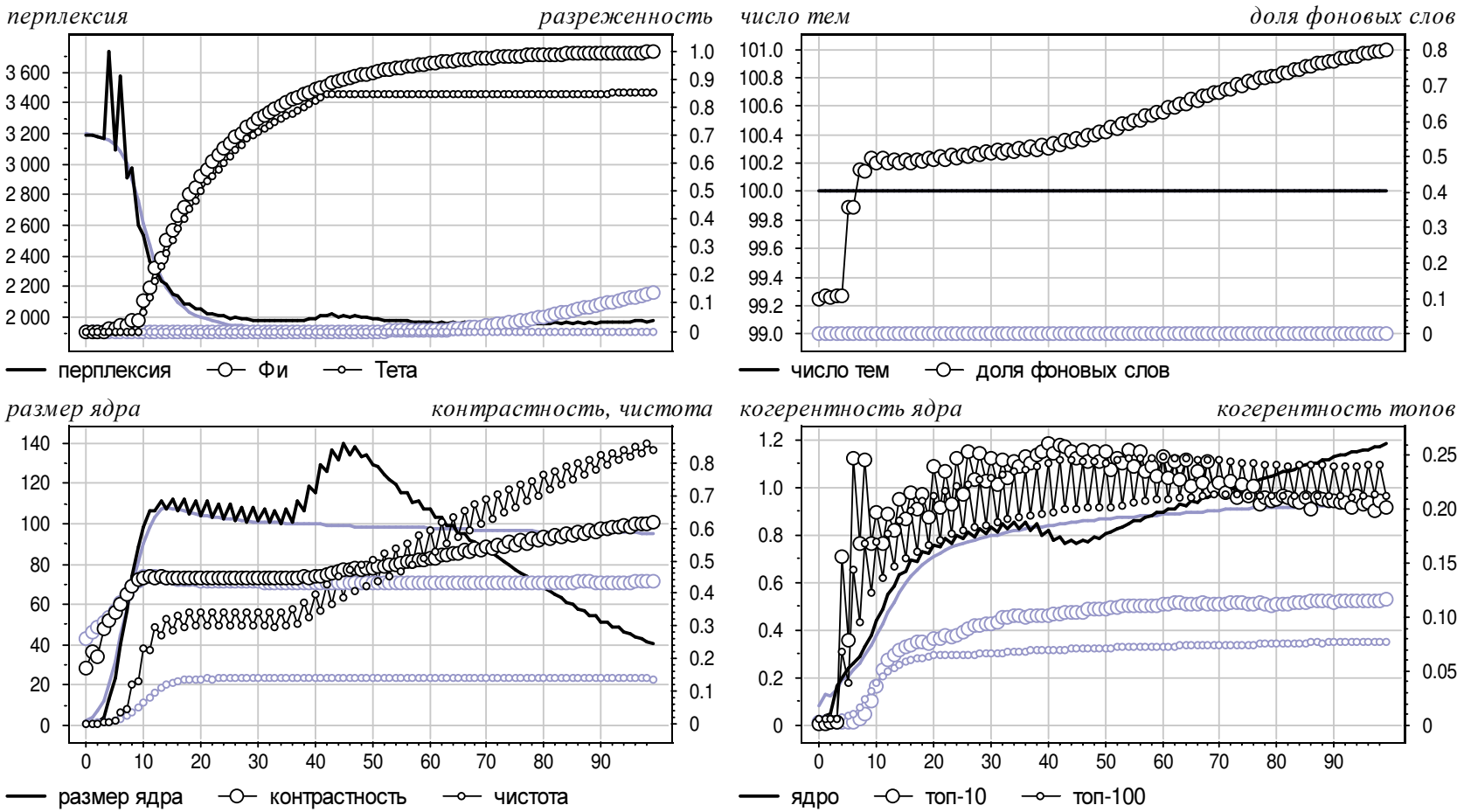
Критерии качества модели

- Перплексия контрольной коллекции: $\mathcal{P} = \exp(-\mathscr{L}/N)$
- Разреженность — доля нулевых элементов в Φ и Θ
- Характеристики интерпретируемости тем:
  - когерентность темы: [Newman, 2010]
  - размер ядра темы: $|W_t|$, ядро $W_t = \{w : p(t|w) > 0.25\}$
  - чистота темы: $\sum_{t \in W_t} p(w|t)$
  - контрастность темы: $\frac{1}{|W_t|} \sum_{t \in W_t} p(t|w)$
- Вырожденность тематической модели:
  - число тем: $|T|$
  - доля фоновых слов: $\frac{1}{n} \sum_{d \in D} \sum_{w \in d} \sum_{t \in B} p(t|d, w)$

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

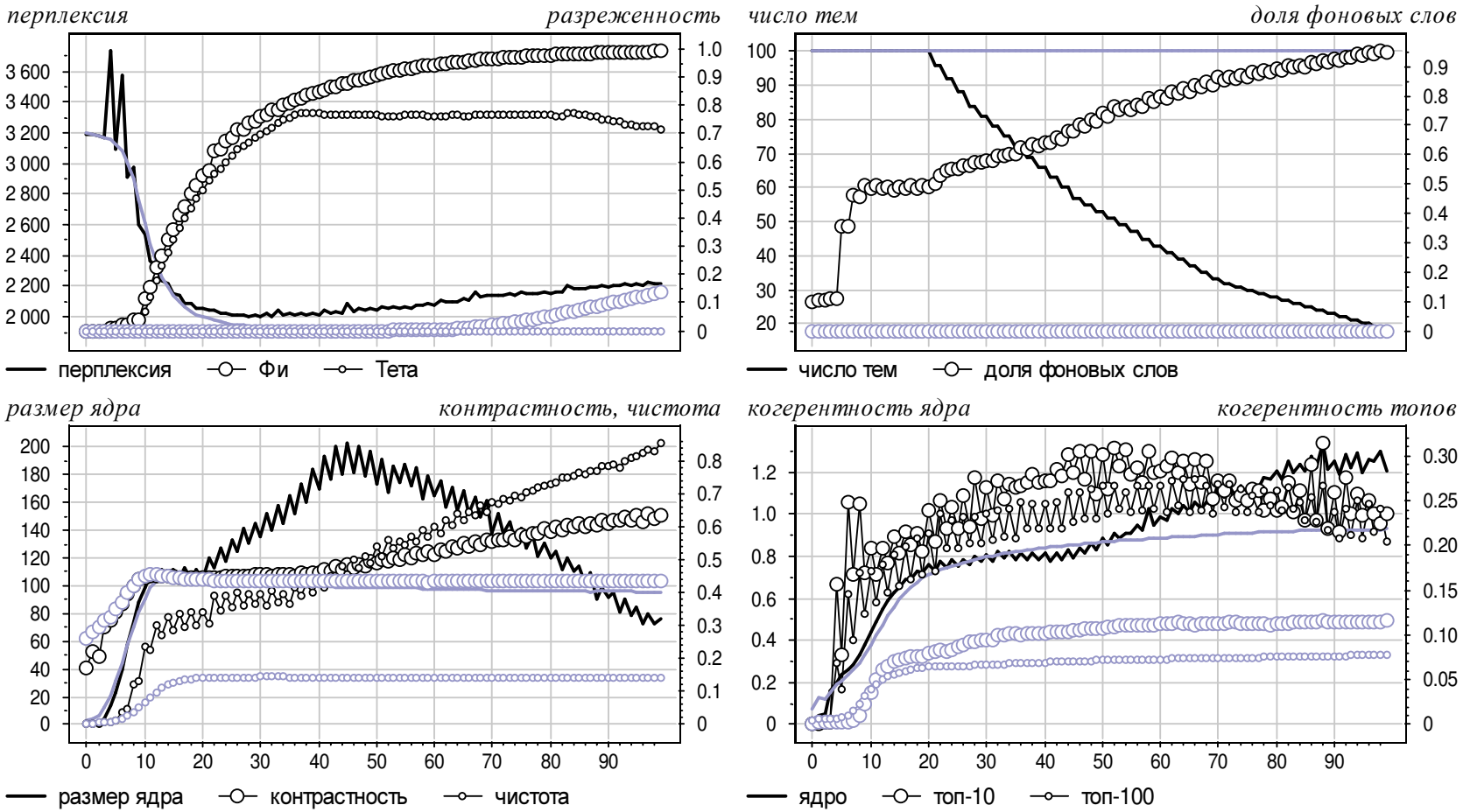
Комбинирование разреживания, сглаживания и декорреляции

Зависимости критериев качества от итераций EM-алгоритма (серый — PLSA, чёрный — ARTM)



Все те же, с удалением незначимых тем

Зависимости критериев качества от итераций EM-алгоритма (серый — PLSA, чёрный — ARTM)



Выводы

Показана возможность одновременного:

- усиления разреженности (до 98%)
- улучшения интерпретируемости (когерентности) тем
- повышения различности (чистоты и контрастности) тем
- при размере ядер тем 50–150 слов
- почти без потери перплексии (правдоподобия) модели

Подобраны траектории регуляризации:

- разреживание включать постепенно после 10-20 итераций
- сглаживание включать сразу
- декорреляцию включать сразу и как можно сильнее
- удаление незначимых тем включать постепенно,
- никогда не совмещая с декорреляцией на одной итерации

Направления ближайших исследований

- Лингвистическая регуляризация, отказ от «мешка слов»
  - учёт линейной структуры текста
  - учёт лингвистических ресурсов (тезаурусов, онтологий)
  - выделение терминов-словосочетаний
- Разработка BigARTM — библиотеки с открытым кодом
  - параллельные вычисления
  - распределённое хранение коллекции
  - любые сочетания регуляризаторов
- Улучшение сходимости
  - оптимизация начального приближения
  - методы глобальной оптимизации
  - нисходящие иерархические модели
- Визуализация результатов поиска научной информации

Воронцов Константин Вячеславович
voron@yandex-team.ru

Страницы на www.MachineLearning.ru:

- Участник:Vokov
- Вероятностные тематические модели (курс лекций, К. В. Воронцов)
- Тематическое моделирование

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН, Т. 456, №3, 2014.

Vorontsov K. V., Potapenko A. A., Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // AIST'14. Springer. 2014.