Вероятностные тематические модели
Лекция 6. Байесовский вывод для тематического моделирования

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 • 21 марта 2019

Содержание

Напоминание. Задача тематического моделирования

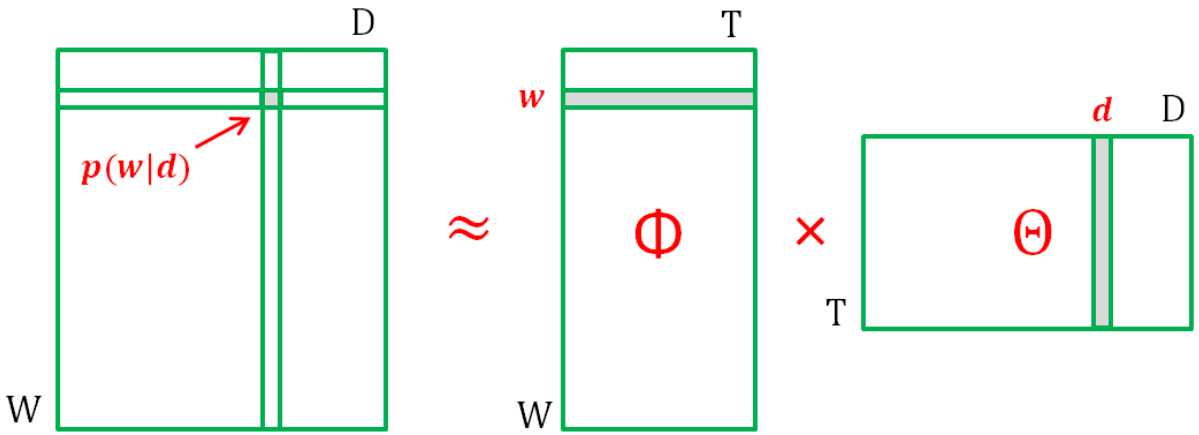
Дано: коллекция текстовых документов

- 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. Тематическая модель LDA

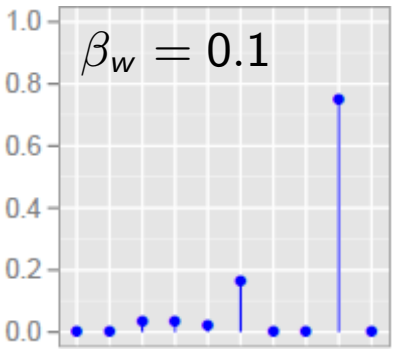
Распределения Дирихле с гиперпараметрами $\alpha \in \mathbb{R}^T$, $\beta \in \mathbb{R}^W$:

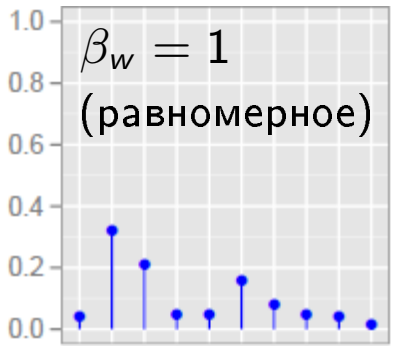
$$p(\Phi, \Theta|\alpha, \beta) = \prod_{t\in T} \mathrm{Dir}(\phi_t|\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d|\alpha)$$

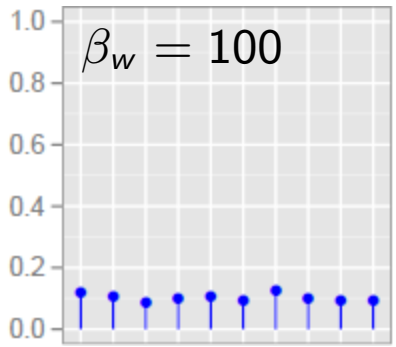
$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \phi_{wt} > 0; \quad \beta_0 = \sum_w \beta_w,\ \beta_t > 0;$$

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

Пример. Распределение $\mathrm{Dir}(\phi|\beta)$ при $|W| = 10$, $\phi, \beta \in \mathbb{R}^{10}$:





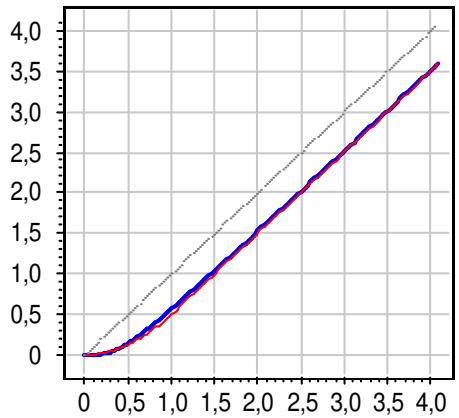


Свойства распределения Дирихле

- Математическое ожидание:
  $\mathrm{E}\theta_t = \int \theta_t \mathrm{Dir}(\theta|\alpha)\, d\theta = \frac{\alpha_t}{\alpha_0} = \underset{t}{\mathrm{norm}}(\alpha_t)$
- Мода: $\hat{\theta}_t = \frac{\alpha_t - 1}{\alpha_0 - T} = \underset{t}{\mathrm{norm}}(\alpha_t - 1)$
- Дисперсия: $\mathrm{D}\theta_t = \frac{\alpha_t(\alpha_0 - \alpha_t)}{\alpha_0^2(\alpha_0+1)}$
- Математическое ожидание логарифма:
  $\mathrm{E}\ln\theta_t = \int \ln\theta_t \mathrm{Dir}(\theta|\alpha)\, d\theta = \psi(\alpha_t) - \psi(\alpha_0)$,
  где $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ — дигамма-функция.

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

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



Напоминание. Общие обозначения и общий EM-алгоритм

$X = (d_i, w_i)_{i=1}^n$ — 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: зная X, найти апостериорное распределение $p(\Omega|X, \gamma)$

Максимизация неполного правдоподобия:

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

EM-алгоритм для решения данной задачи максимизации:

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

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

Две основные идеи Variational Bayesian inference

Идея 1.
Раньше $\Omega = (\Phi, \Theta)$ были неслучайными параметрами.
Теперь столбцы Φ, Θ порождаются распределениями Дирихле:

$$p(\Phi, \Theta|\alpha, \beta) = \prod_t \mathrm{Dir}(\phi_t|\beta) \prod_d \mathrm{Dir}(\theta_d|\alpha),$$

$\Omega = (\alpha, \beta)$ — параметры вероятностной модели вместо (Φ, Θ),
(Z, Φ, Θ) — скрытые переменные вместо Z.

Идея 2.
Если $q(Z) = p(Z|X, \Omega)$ не вычислимо, то ищем $q(Z)$ в виде факторизации по группам переменных Z_j, $j \in J$:

$$q(Z) = \prod_{j \in J} q_j(Z_j)$$

Основная теорема вариационного байесовского вывода

Теорема. Решение задачи $\mathrm{KL}\big(q(Z) \,\big\|\, p(Z|X,\Omega)\big) \to \min_q$
в семействе факторизованных распределений $q(Z) = \prod_j q_j(Z_j)$
удовлетворяет системе уравнений

$$\ln q_j(Z_j) = \mathrm{E}_{q\setminus j} \ln p(X, Z|\Omega) + \mathrm{const},$$

где $\mathrm{E}_{q\setminus j}$ — матожидание по всем переменным кроме Z_j,
const — нормировочный множитель распределения q_j.

Для решения этой системы используют метод простой итерации.

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

$$\sum_{Z_j} q_j(Z_j) \underbrace{\sum_{Z\setminus Z_j} \prod_{i\neq j} q_i(Z_i) \ln p(X, Z|\Omega)}_{\mathrm{E}_{q\setminus j} \ln p(X,Z|\Omega)} - \sum_{Z_j} q_j(Z_j) \ln q_j(Z_j) \;\to\; \min_q$$

Доказательство

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

$$\sum_Z q(Z) \ln \frac{p(Z|X,\Omega)}{q(Z)} \to \max_q \Leftrightarrow \sum_Z q(Z) \ln \frac{p(X,Z|\Omega)}{q(Z)} - \sum_Z q(Z) \ln p(X|\Omega) \to \max_q$$

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

$$\sum_{Z_j} q_j(Z_j) \underbrace{\sum_{Z \setminus Z_j} \prod_{i \neq j} q_i(Z_i) \ln p(X,Z|\Omega)}_{\mathsf{E}_{q \setminus j} \ln p(X,Z|\Omega)} - \sum_{Z_j} q_j(Z_j) \underbrace{\sum_{Z \setminus Z_j} \prod_{i \neq j} q_i(Z_i) \sum_{k \in J} \ln q_k(Z_k)}_{\ln q_j(Z_j) + \text{const}} \to \max_{q_j}$$

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

$$\underbrace{\sum_{Z \setminus Z_j} \prod_{i \neq j} q_i(Z_i) \sum_{k \neq j} \ln q_k(Z_k)}_{\text{не зависит от } q_j} + \underbrace{\sum_{Z \setminus Z_j} \prod_{i \neq j} q_i(Z_i)}_{1} \ln q_j(Z_j)$$

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

$$\ln q_j(Z_j) = \mathsf{E}_{q \setminus j} \ln p(X,Z|\Omega) + \text{const}.$$

■

Снова максимизация неполного правдоподобия

Теперь (Z, Φ, Θ) — это скрытые переменные,
$\Omega = (\alpha, \beta)$ — это параметры вероятностной модели.

Задача максимизации неполного правдоподобия:

$$p(X|\alpha,\beta) = \sum_Z \int_\Phi \int_\Theta p(X, Z, \Phi, \Theta|\alpha,\beta) d\Phi d\Theta \;\to\; \max_{\alpha,\beta}$$

EM-алгоритм для решения данной задачи имеет вид:

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

$$\text{M-шаг:}\quad \sum_Z \int_\Phi \int_\Theta q(Z,\Phi,\Theta) \ln p(X,Z,\Phi,\Theta|\alpha,\beta) d\Phi d\Theta \;\to\; \max_{\alpha,\beta}$$

Теперь всё происходит на E-шаге (аналог старых E и M шагов)
Новый M-шаг часто опускают, фиксируя α, β без оптимизации

Вариационная аппроксимация для тематической модели LDA

Ищем $q(Z, \Phi, \Theta)$ в семействе факторизованных распределений (если переменные независимы, решение будет точным):

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

$J = \{1, \ldots, n\} \sqcup T \sqcup D$ — индексы всех скрытых переменных.

Основная теорема вариационного байесовского вывода

Решение задачи E-шага удовлетворяет системе уравнений

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

где $\mathsf{E}_{q \setminus j}$ — матожидание по всем переменным кроме j-й,
const — нормировочный множитель распределения q_j.

Для решения этой системы используют метод простой итерации.

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

Применяя *основную теорему*, переводим независящие от скрытых переменных множители в const:

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

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

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

Теперь надо брать матожидания $\mathsf{E}_{q\setminus j}$ от этой величины по всем распределениям $q_t(\phi_t)$, $q_d(\theta_d)$, $q_i(t_i)$, кроме j-го.

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

Распределения скрытых переменных $q_t(\phi_t)$, $q_d(\theta_d)$, $q_i(t_i)$

Распределение скрытой переменной $\phi_t \in \mathbb{R}^W$:

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

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

Распределения скрытых переменных $q_t(\phi_t)$, $q_d(\theta_d)$, $q_i(t_i)$

Распределение скрытой переменной $\theta_d \in \mathbb{R}^T$:

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

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

Распределения скрытых переменных $q_t(\phi_t)$, $q_d(\theta_d)$, $q_i(t_i)$

Распределение скрытой переменной $t_i \in T$:

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

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

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

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

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

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

Собираем всё воедино

В итерационном процессе чередуются два шага:

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

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

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

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

Точечные оценки параметров по матожиданию или моде:

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

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

Промежуточный итог

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

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

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

Основные идеи Gibbs Sampling

$\Omega = (\alpha, \beta)$ — параметры вероятностной модели.
Поочерёдное выполнение двух шагов (аналог EM-алгоритма):

$$\text{E-шаг:} \quad Z \sim p(Z|X, \Omega)$$
$$\text{M-шаг:} \quad (\Phi, \Theta) \sim p(\Phi, \Theta|X, Z, \Omega)$$

Основная теорема о сходимости сэмплирования Гиббса

Процесс сэмплирования одномерных случайных величин

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

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

Распределение Дирихле — сопряжённое к мультиномиальному

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

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

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

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

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

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

$$\propto \prod_{t\in T} \mathrm{Dir}(\phi_t|\tilde\beta_t) \prod_{d\in D} \mathrm{Dir}(\theta_d|\tilde\alpha_d), \quad \tilde\beta_{wt} = n_{wt}+\beta_w, \quad \tilde\alpha_{td} = n_{td}+\alpha_t.$$

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

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

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

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

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

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

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

Итак, мы только что получили распределение

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

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

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

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

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

Для сэмплирования Гиббса нужно одномерное распределение

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

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

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

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

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

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

Собираем всё воедино

Выделены отличия от вариационного алгоритма

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

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

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

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

Параметры Φ, Θ сэмплируются или оцениваются точечно:

$$(\phi_{wt})_w \sim \text{Dir}\left(\phi \,\middle|\, (n_{wt} + \beta_w)_w\right) \qquad (\theta_{td})_t \sim \text{Dir}\left(\theta \,\middle|\, (n_{td} + \alpha_t)_t\right)$$

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

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

Алгоритм сэмплирования Гиббса

Вход: коллекция D, число тем $|T|$, параметры α, β;
Выход: распределения Φ и Θ;

$n_{wt}, n_{td}, n_t, n_d := 0$ для всех $d \in D$, $w \in W$, $t \in T$;
для всех итераций $k := 1, \ldots, k_{\max}$
  для всех документов $d \in D$ и терминов $w = w_1, \ldots, w_{n_d} \in d$
    если $k \geqslant 2$ **то** $t := t_{dw}$; $--n_{wt}$; $--n_{td}$; $--n_t$; $--n_d$;
    $p(t|d,w) = \underset{t \in T}{\text{norm}}\left(\frac{n_{wt}+\beta_w}{n_t+\beta_0} \cdot \frac{n_{td}+\alpha_t}{n_d+\alpha_0}\right)$ для всех $t \in T$;
    сэмплировать одну тему t из распределения $p(t|d,w)$;
    $t_{dw} := t$; $++n_{wt}$; $++n_{td}$; $++n_t$; $++n_d$;

$\phi_{wt} := n_{wt}/n_t$ для всех $w \in W$, $t \in T$;
$\theta_{td} := n_{td}/n_d$ для всех $d \in D$, $t \in T$;

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

Промежуточный итог

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

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

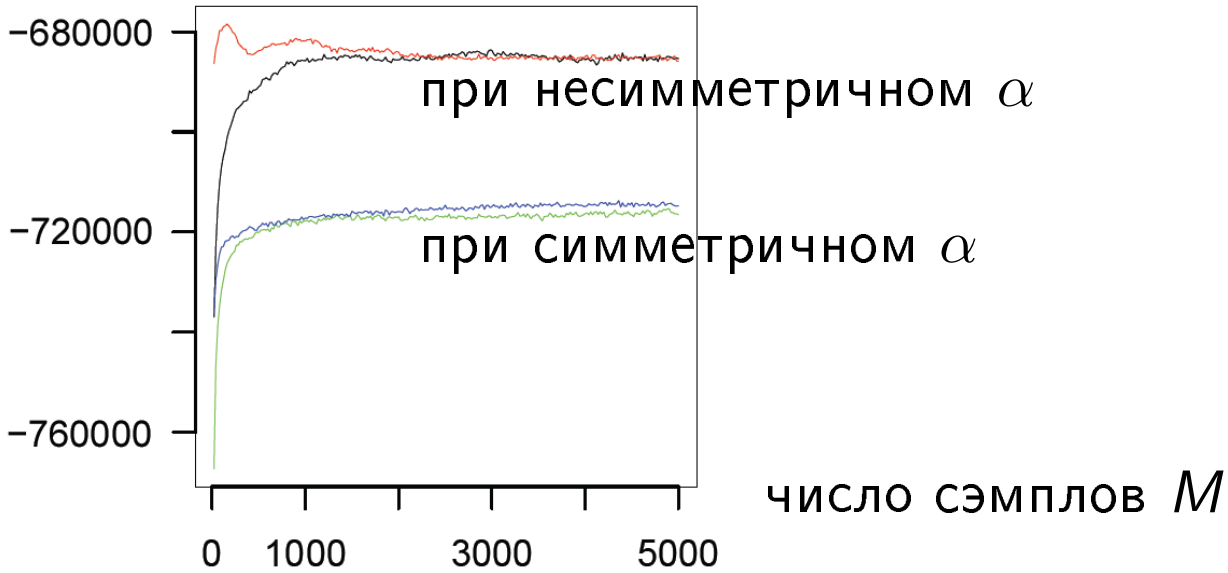
Проблема выбора гиперпараметров α и β

Стандартная рекомендация [2004]: $\alpha_t = 50/|T|$, $\beta_w = 0.01$.

Выводы по результатам более тонкого исследования [2009]:

- $p(t|d) \sim \mathrm{Dir}(\theta; \alpha)$, оптимизировать $\alpha = (\alpha_1, \ldots, \alpha_T)$.
- $p(w|t) \sim \mathrm{Dir}(\phi; \beta)$, взять симметричное $\beta_1 = \cdots = \beta_T \ll 1$.

правдоподобие



H.Wallach, D.Mimno, A.McCallum. Rethinking LDA: why priors matter. NIPS, 2009.

Оптимизация гиперпараметра α

Обоснованность (evidence) модели на коллекции D:

$$P(D|\alpha) = \prod_{d \in D} \frac{\Gamma(\alpha_0)}{\Gamma(n_d + \alpha_0)} \prod_{t \in T} \frac{\Gamma(n_{td} + \alpha_t)}{\Gamma(\alpha_t)} \to \max_{\alpha}$$

Метод неподвижной точки [Minka, 2003] — итерационный процесс, встраиваемый между проходами по всей коллекции:

$$\alpha_t := \alpha_t \frac{\sum_d \psi(n_{td} + \alpha_t) - \psi(\alpha_t)}{\sum_d \psi(n_d + \alpha_0) - \psi(\alpha_0)},$$

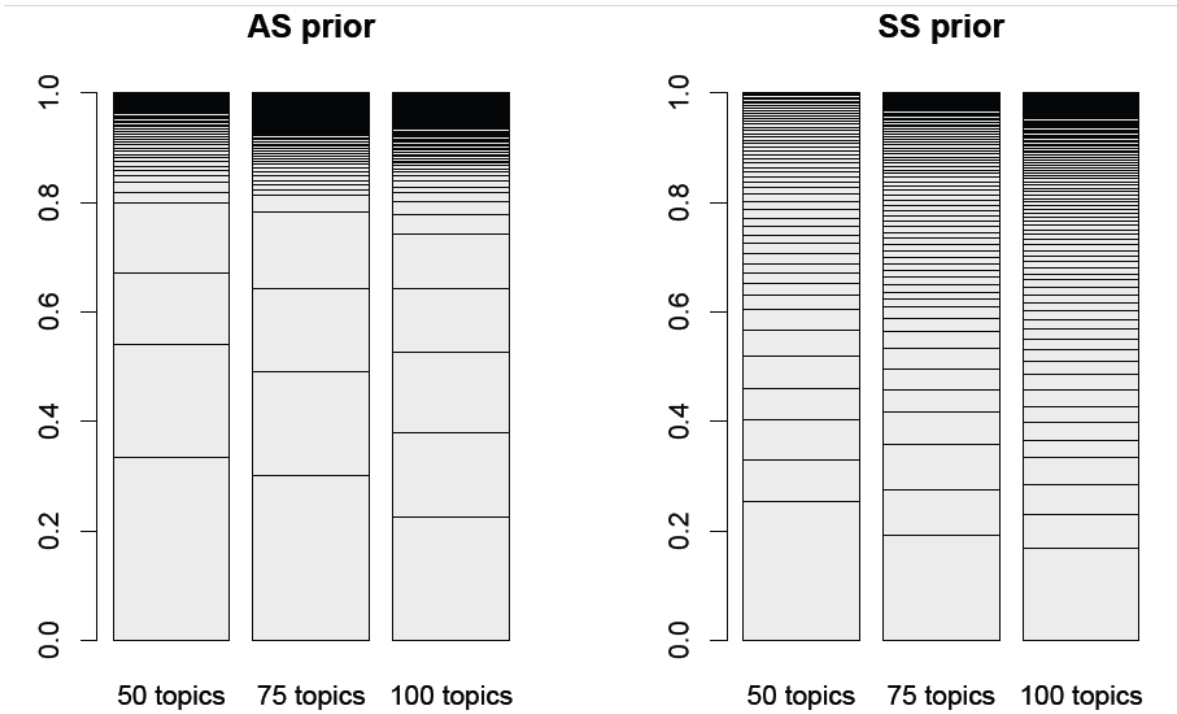
где $\psi(z) = \big(\ln \Gamma(z)\big)' = \Gamma'(z)/\Gamma(z)$ — дигамма-функция.

Thomas Minka. Estimating a Dirichlet distribution. 2003.

Hanna Wallach. Structured Topic Models for Language. PhD thesis, University of Cambridge, 2008.

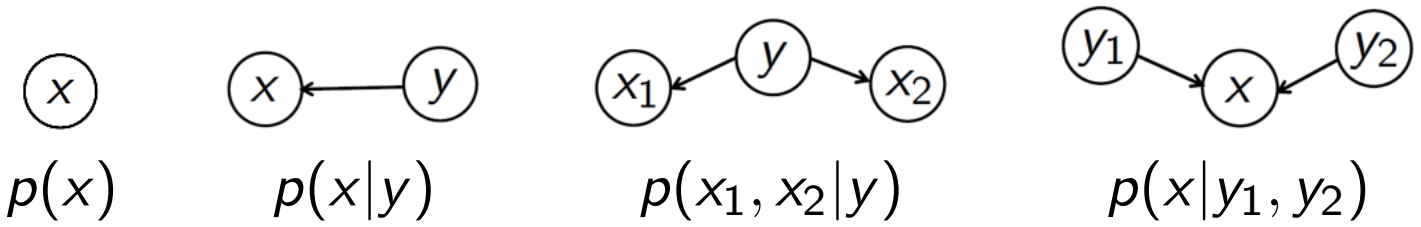
Преимущество оптимизации гиперпараметра α

- Правдоподобие существенно выше.
- Сходимость быстрее.
- Меньшая чувствительность к избыточному $|T|$.
- Более естественная несбалансированность тем.



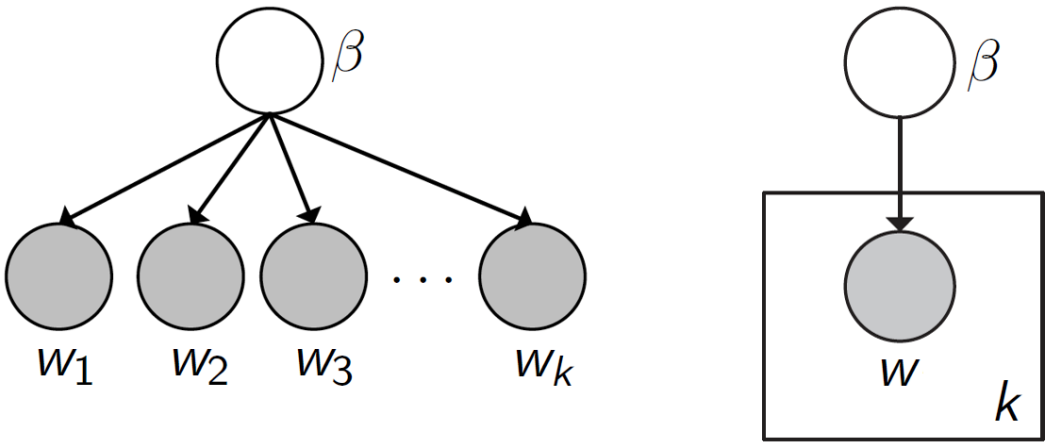
Графическая нотация «plate notation»

Графическое представление условных зависимостей



$p(x)$ $p(x|y)$ $p(x_1, x_2|y)$ $p(x|y_1, y_2)$

Графическое представление выборки $w_1, \ldots, w_k$, порождаемой распределением $\beta_w = p(w)$

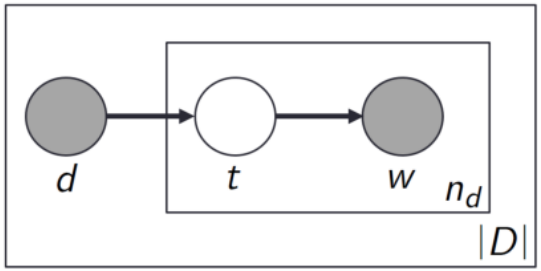


Графическая нотация для моделей PLSA и LDA

Модель PLSA:

каждый $d \in D$ порождает скрытые темы:
$t_i \sim p(t|d), \quad i = 1, \ldots, n_d$;

каждая тема t_i порождает слово:
$w_i \sim p(w|t_i), \quad i = 1, \ldots, n_d$.

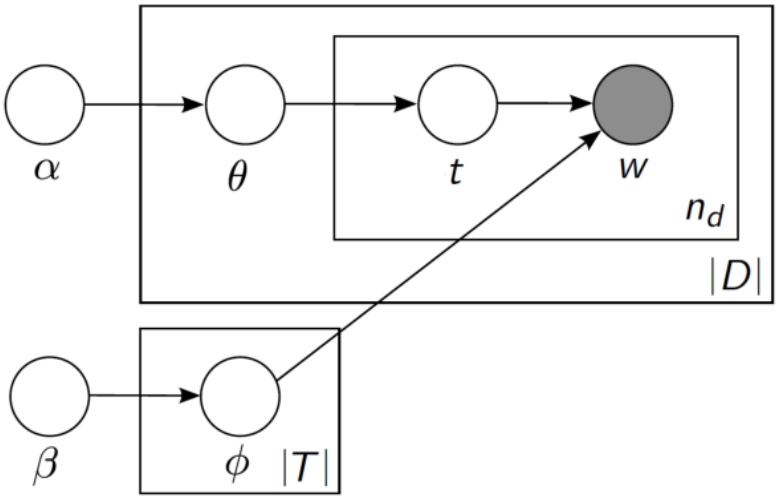


Модель LDA:

α порождает векторы документов:
$\theta_d \sim \mathrm{Dir}(\theta|\alpha), \quad d \in D$;

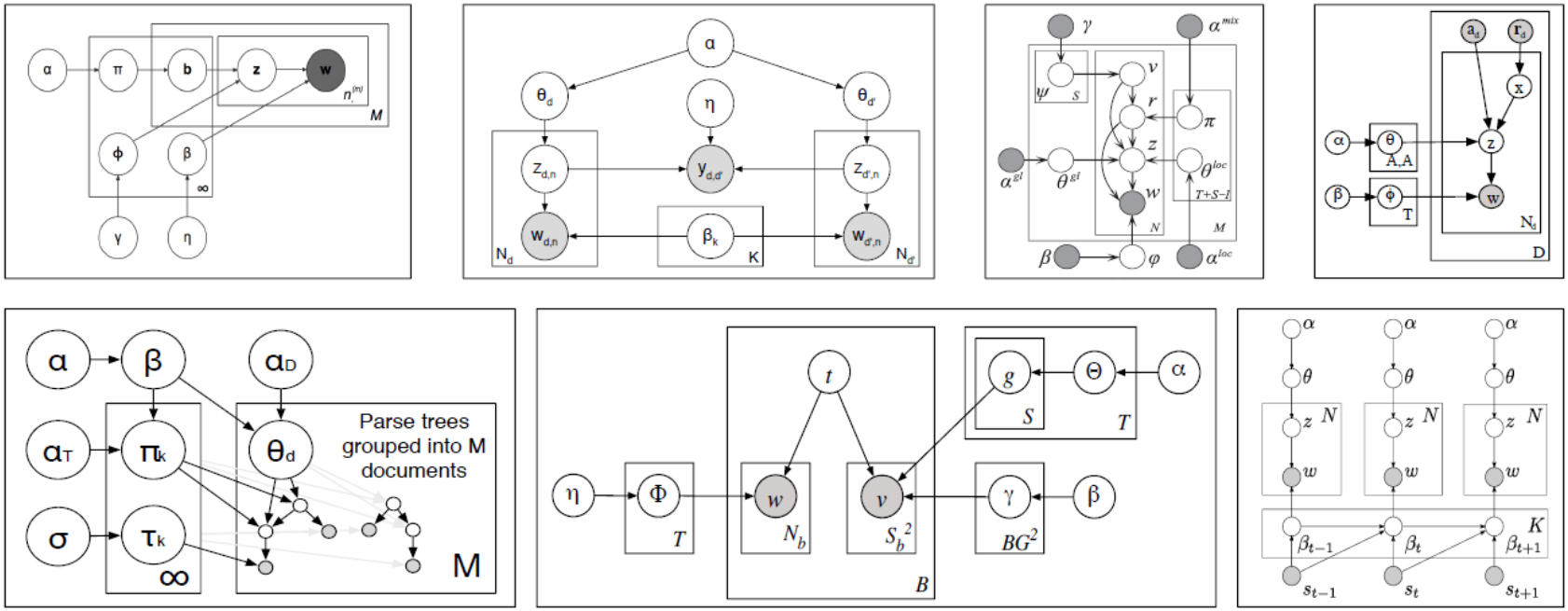
β порождает векторы тем:
$\phi_t \sim \mathrm{Dir}(\phi|\beta), \quad t \in T$;

далее как в PLSA.



Графическая нотация: представление структуры модели

Большое структурное разнообразие тематических моделей:



David Blei. Probabilistic topic models // Communications of the ACM, 2012.

Обсуждение «Stop using Plate Notation»

Единственное достоинство и куча недостатков:

- + хорошо запоминающийся наглядный образ модели
- – множественность вариантов отображения одной модели
- – неполнота и неоднозначность интерпретации
- – не ясен переход от картинки к модели и алгоритму
- – во многих статьях этот переход скрыт или скомкан
- – иллюзия понятности и практическая бесполезность

Один из комментариев:

Every now and then the topic comes up as to why algorithms and procedures are explained in obtuse forms across the entirety of the paper it is described in, usually we just conclude that *it would look too simple if it were explained any other way*.

Rob Zinkov. Stop using Plate Notation. 2013-07-28.
`http://zinkov.com/posts/2013-07-28-stop-using-plates`

Язык псевдокода порождающего процесса (generative story)

Пример: вероятностная порождающая модель PLSA

Вход: $p(w|t)$ для всех $t \in T$, $p(t|d)$ для всех $d \in D$;
Выход: коллекция документов;

для всех документов $d \in D$
  для всех позиций слов $i = 1, \ldots, n_d$ в документе d
    выбрать тему t_i из $p(t|d)$;
    выбрать слово w_i из $p(w|t_i)$;

- $+$ легко понимать модель, описание недвусмысленно
- $-$ не ясен переход от модели к алгоритму
- $-$ во многих статьях этот переход скрыт или скомкан

Язык задач оптимизации: модель → критерий → алгоритм

1. Вероятностная модель порождения данных:

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

2. Постановка задачи оптимизации (ARTM):

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

3. Алгоритм решения — итерационный процесс:

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

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

Математический язык оптимизационных задач

Один недостаток и куча достоинств:

- $-$ плохо понятен специалистам гуманитарных профессий
- $+$ ясность и однозначность записи модели
- $+$ ясность критерия оптимальности
- $+$ ясность вариантов выбора методов оптимизации
- $+$ переход от модели к алгоритму — это теорема
- $+$ возможность комбинирования моделей
- $+$ унифицированная модульная реализация (BigARTM)

Байесовское обучение — доминирующий подход в TM

Модель LDA предопределила дальнейшее развитие тематических моделей на основе *байесовского вывода*:

$$\text{Posterior}(\Phi, \Theta|\text{data}) \propto \text{Prior}(\Phi, \Theta)\, \text{P}(\text{data}|\Phi, \Theta)$$

Prior и Posterior в модели LDA — распределения Дирихле.

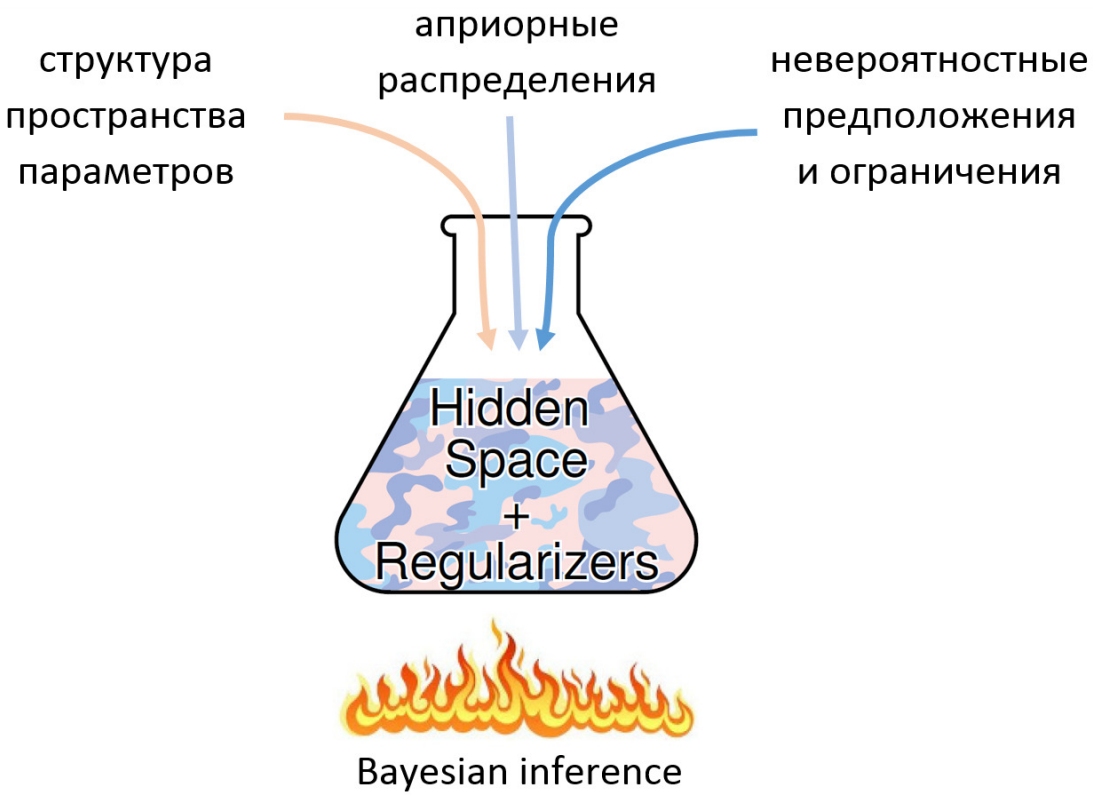
Проблемы:

- Нам нужны лишь значения Φ, Θ, а не их распределения
- Prior Дирихле имеет слабые лингвистические обоснования
- Задача сильно усложняется для несопряжённых Prior
- Байесовский вывод уникален для каждой модели
- Технически трудно обобщать и комбинировать модели

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

Алхимия байесовского вывода в тематическом моделировании

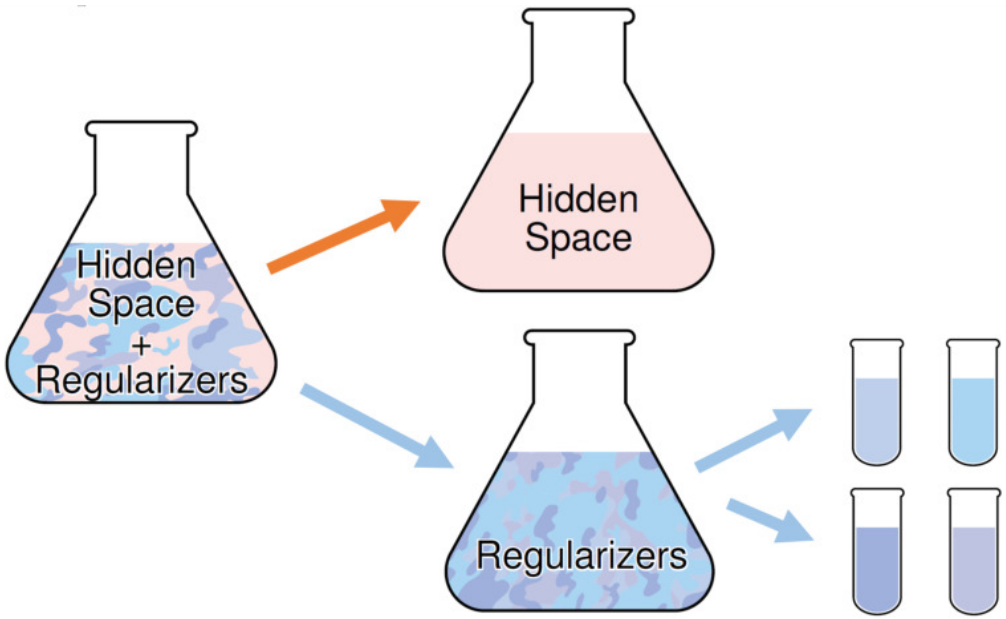
Вероятностная модель порождения данных объединяет в едином описании структуру пространства параметров, априорные распределения и дополнительные знания о задаче.



ARTM — новая алхимия на основе классической регуляризации

Простая *порождающая модель* описывает структуру пространства.
Регуляризаторы суммируются с весами, в любых сочетаниях,
и каждый описывает только одно дополнительное требование.

Декомпозиция — классический способ упрощения задачи



ARTM: модульный подход к синтезу требуемых моделей

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

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

-- *нестандартизуемые этапы, уникальная разработка для каждой задачи*

-- *стандартизуемые этапы*

Резюме

- Максимизация апостериорной вероятности (MAP) даёт точечные оценки (Φ, Θ) и полностью совместима с ARTM.
- *Байесовский вывод* оценивает $p(\Phi, \Theta|X)$ вместо (Φ, Θ). Такое усложнение задачи представляется избыточным.
- Итерационный процесс в байесовских методах VB и GS не сильно отличается от EM-алгоритма для MAP.
- Байесовский вывод уникален для каждой модели. Нет общего алгоритма. Нет модульной реализации.
- Нет оптимизационной постановки задачи для (Φ, Θ). Вместо регуляризаторов — априорные распределения.
- Распределение Dir упрощает байесовский вывод, но не имеет убедительных лингвистических обоснований.
- Другие распределения не являются сопряжёнными мультиномиальному, что усложняет байесовский вывод.