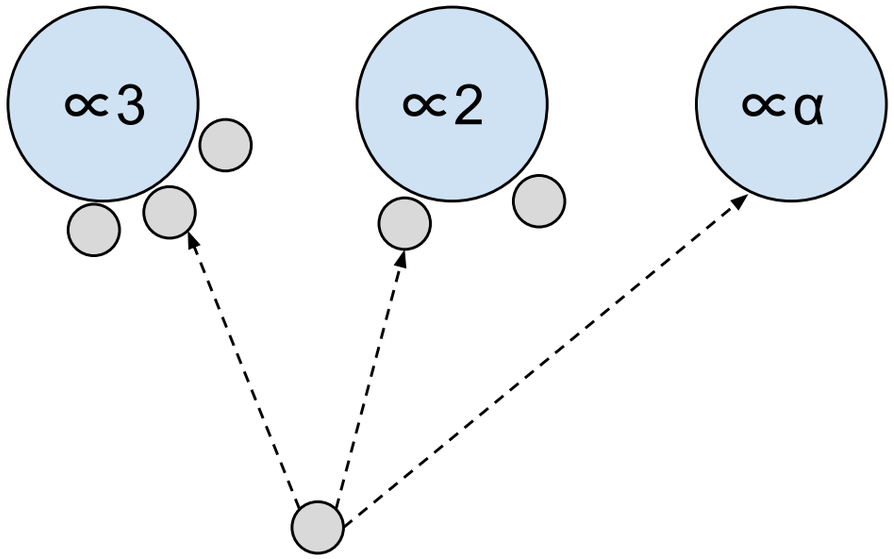
distance-dependent Chinese Restaurant Process

Сергей Бартунов
ВЦ РАН

Chinese Restaurant Process

Априорное распределение над принадлежностью кластарам

$$p(\mathbf{z}) = \prod_i p(z_i|\mathbf{z}_{-i})$$



Посетители заходят в ресторан один за одним

- Первый садится за первый стол $z_1 = 1$
- Каждый следующий выбирает стол следующим образом:

$$p(z_i = k|z_{\setminus i}) \propto \begin{cases} n_k, & n_k > 0 \\ \alpha, & n_k = 0 \end{cases}$$

Свойства CRP

- Число кластеров не фиксировано априорно
- Ожидаемое число кластеров растет логарифмически от числа посетителей
- Меняя параметр α, можно влиять на “разрешающую способность” процедуры кластеризации
- Предполагает, что порядок, в котором поступают данные, не имеет значения

Exchangeability (симметрическая зависимость)

- В явном или неявном виде предполагается во множестве алгоритмов

$$p(x_1, x_2, \ldots, x_N) = p(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(N)})$$

- Теорема де Финетти

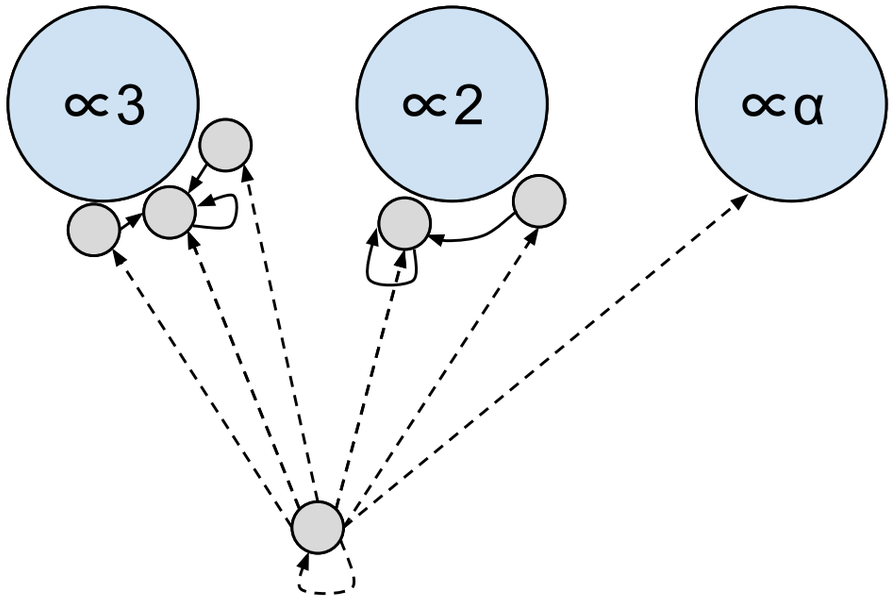
$$p(x_1, x_2, \ldots, x_N) = \int p(\theta) \prod_i^N p(x_i|\theta) d\theta$$

- Но структура может иметь значение

Эквивалентная формулировка CRP

Переформулируем CRP: вместо **z** теперь **c**

$$p(\mathbf{c}) = \prod_i p(c_i)$$



Посетители заходят в ресторан **одновременно**

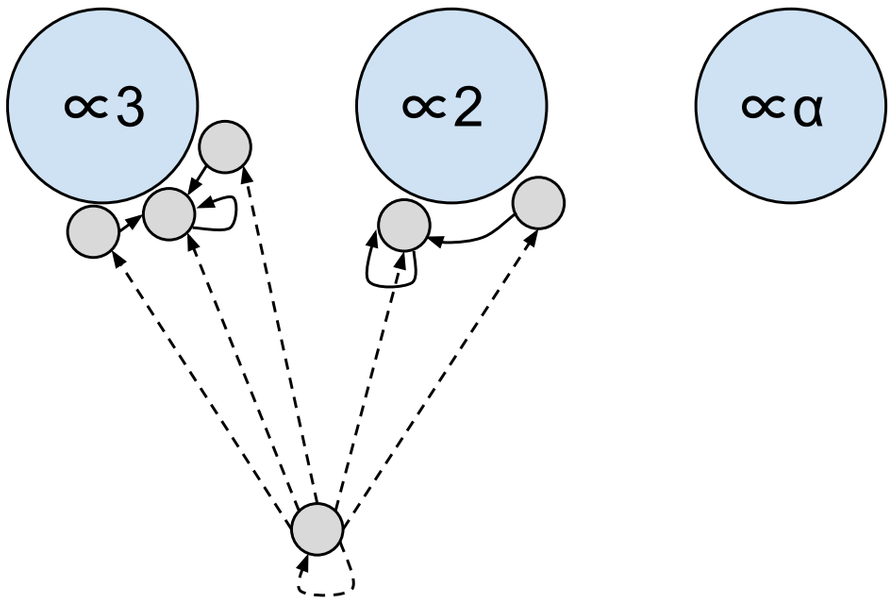
- Каждый берет за руку ровно одного посетителя с вероятностью $\propto 1$
- Либо берет за руку сам себя с вер-ю $\propto \alpha$

$$p(c_i = j) \propto \begin{cases} 1, & i > j \\ \alpha, & i = j \end{cases}$$

$$z_i(c) = \begin{cases} z_{c_i}(c), & c_i \neq i \\ i, & c_i = i \end{cases}$$

(sequential) distance-dependent CRP

Переформулируем CRP: вместо **z** теперь **c**

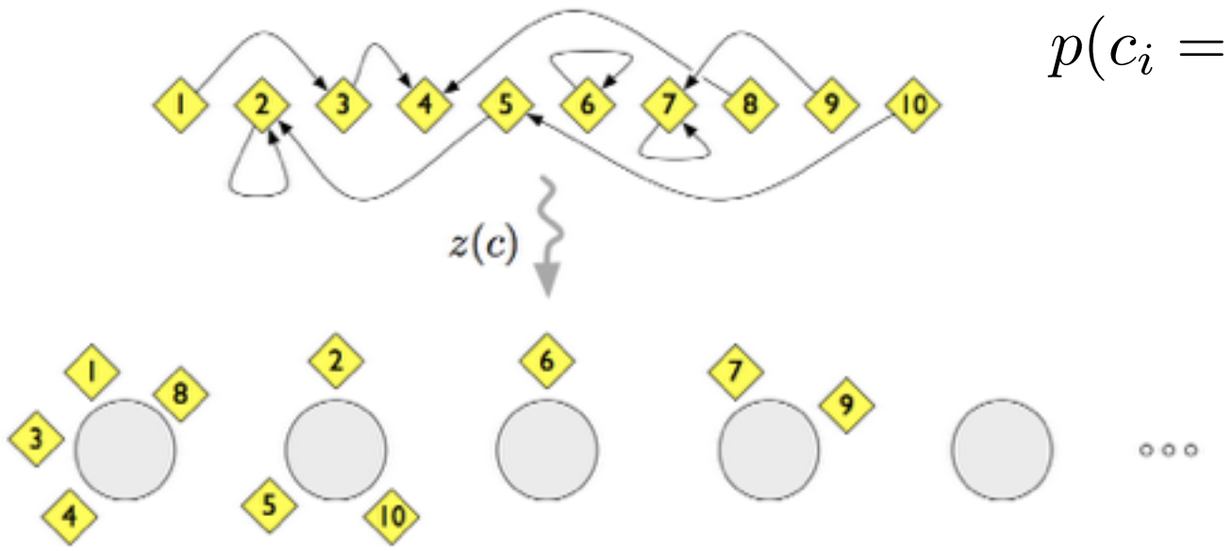


$$p(\mathbf{c}) = \prod_i p(c_i)$$

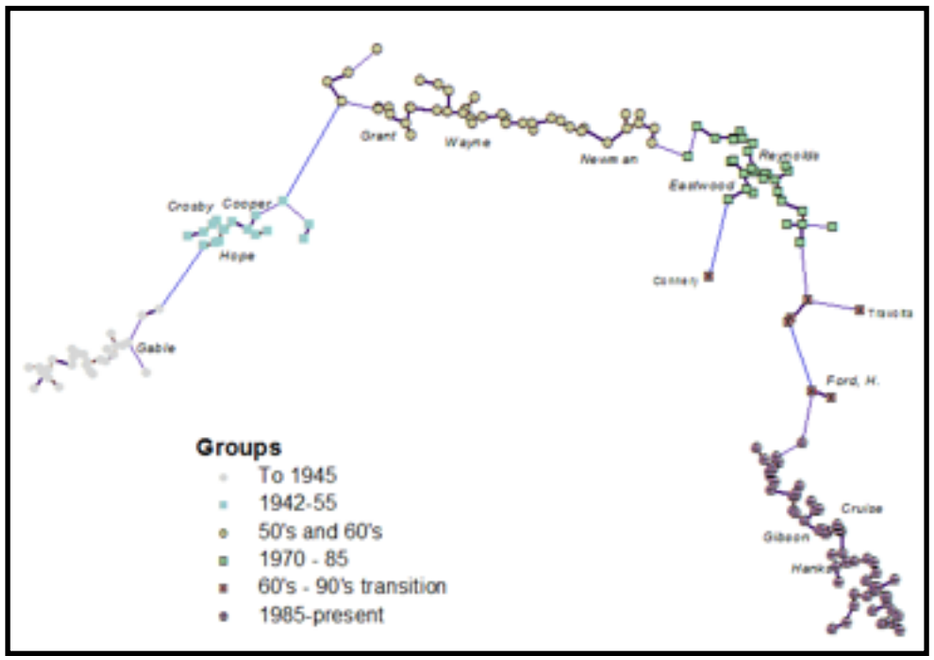
Посетители заходят в ресторан **одновременно**

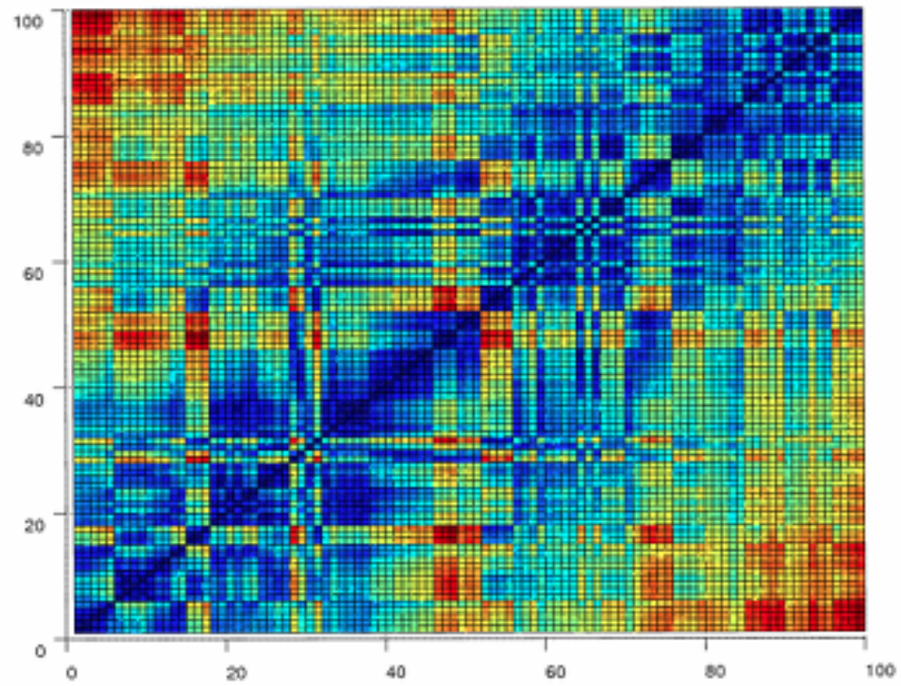
- Каждый берет за руку ровно одного посетителя с вероятностью $\propto f(d_{ij})$
- Либо берет за руку сам себя с вер-ю $\propto \alpha$

$$p(c_i = j | D, f, \alpha) \propto \begin{cases} f(d_{ij}) & i > j \\ \alpha & i = j \end{cases}$$

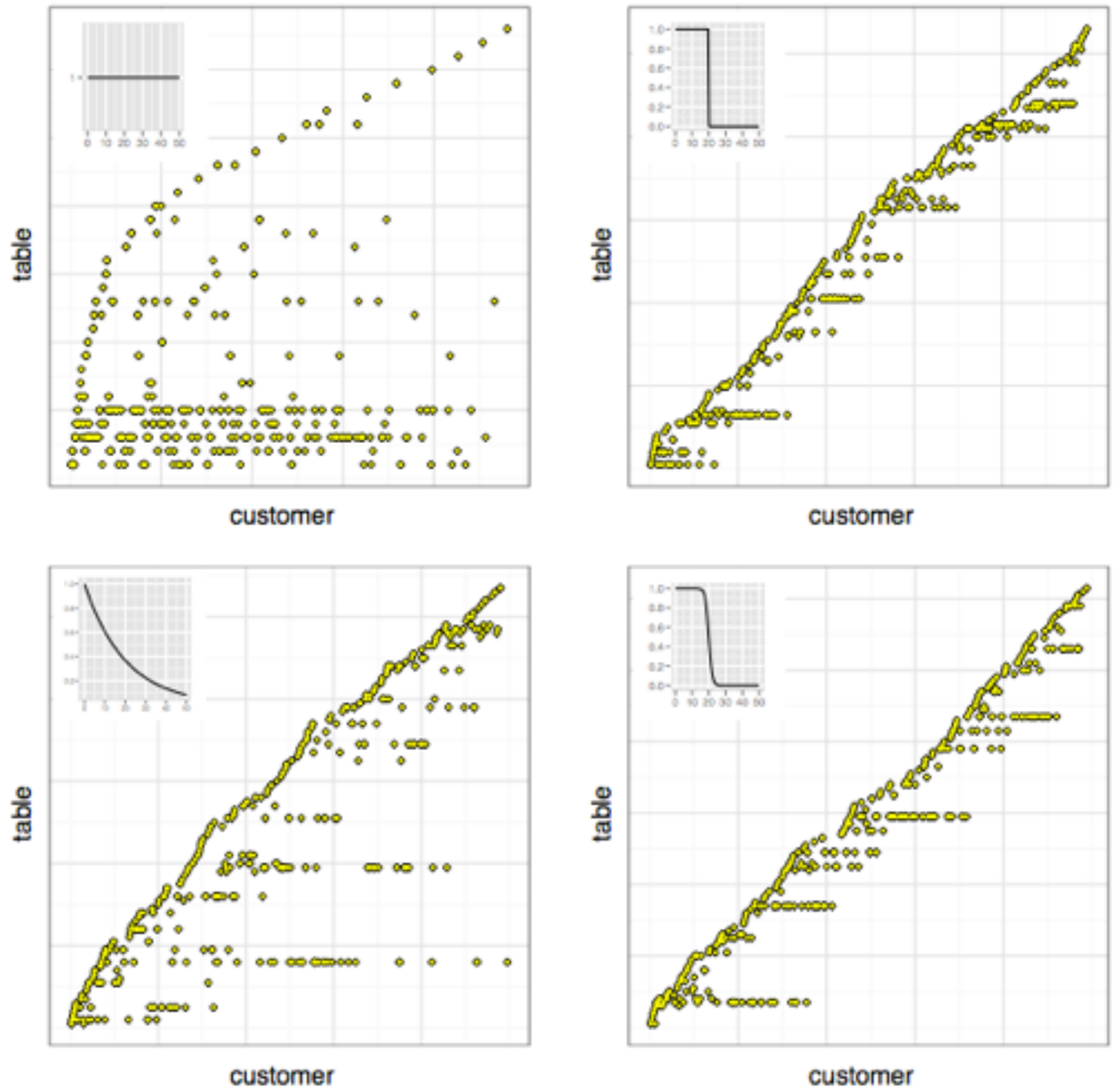


Матрица расстояний





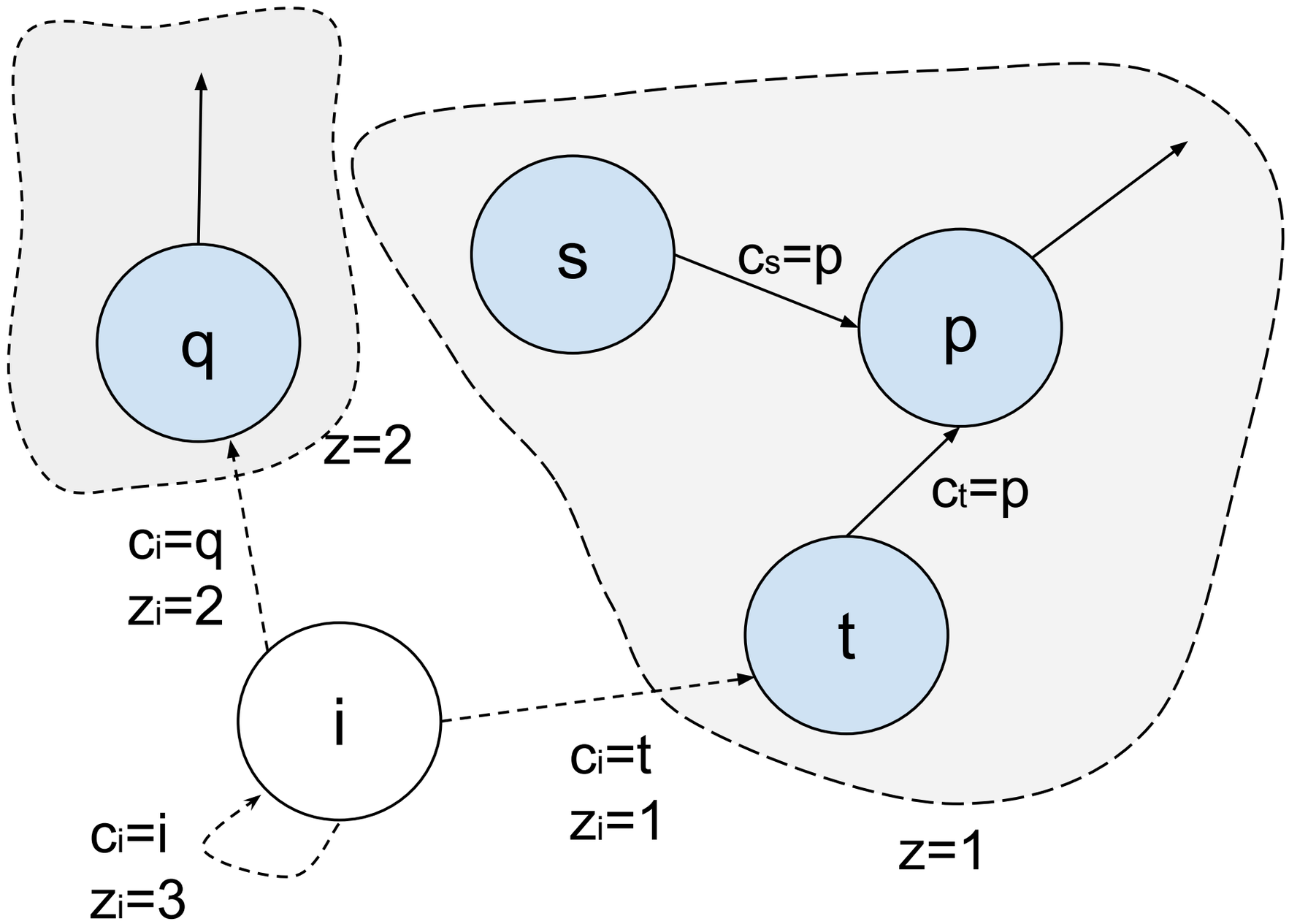
Функция угасания



Как от хватаний за руку перейти к кластерам?

- (seq)ddCRP формулируется не в терминах столов, а в терминах хватаний за руку
- Хватания за руку происходят одновременно и независимо
- Но кластеризация (номер стола) каждого объекта вообще-то зависит от того, кого он взял за руку

Как от хватаний за руку перейти к кластерам?



Смесь распределений с (seq)ddCRP

1. Для каждого объекта $c_i \sim p(c_i|f, D, \alpha)$
2. Детерминировано определяем столы $\mathbf{z} = z(\mathbf{c})$
3. Для каждого стола
    1. Порождаем параметр $\theta_k \sim G_0(\theta)$
    2. Порождаем объекты $x_i \sim p(x|\theta_k), \quad i \in z_k(\mathbf{c})$

Отличия ddCRP от CRP и других процессов

- Модель, основанная на случайной мере:

$$p(\mathbf{x}, \mathbf{z}, \mathbf{\Theta}) = \prod_{k=1}^{\infty} p(\theta_k) \prod_i p(z_i) p(x_i | z_i, \mathbf{\Theta})$$

- Свойство marginal invariance:

$$\int p(\mathbf{x}, \mathbf{z}, \mathbf{\Theta}) dx_j dz_j = \prod_{k=1}^{\infty} p(\theta_k) \prod_{i \neq j} p(z_i) p(x_i | z_i, \mathbf{\Theta}) \int p(x_j, z_j | \mathbf{\Theta}) dx_j dz_j =$$
$$= p(\mathbf{x}_{\setminus j}, \mathbf{z}_{\setminus j}, \mathbf{\Theta})$$

- ddCRP этим свойством не обладает:

$$\int p(\mathbf{x}, \mathbf{c}, \mathbf{\Theta}) dx_j dc_j = \prod_{i \neq j} p(c_i) \int p(c_j) \prod_k p(\theta_k) p(\mathbf{x}_{z_k(\mathbf{c})} | \theta_k) dx_j dc_j$$

Вариацонный вывод для (seq)ddCRP

Исходное распределение: $p(\mathbf{x}, \mathbf{c}, \mathbf{\Theta}) = \prod_i^N p(c_i) \prod_k^{|z(\mathbf{c})|} \left[G_0(\theta_k) \prod_{i \in z_k(\mathbf{c})} p(x_i|\theta_k) \right]$

Преобразуем его в $p(\mathbf{x}, \mathbf{c}, \mathbf{\Theta}) = \left[\prod_i^N \prod_j^N p(c_i = j)^{c_{ij}} \right] \prod_j^N \left[G_0(\theta_j) \prod_i^N p(x_i|\theta_j)^{z_{ij}(\mathbf{c})} \right]$

Обозначим:

$z_{ij}(\mathbf{c}) = 1 \Leftrightarrow r_{ij}(\mathbf{c}) = 1 \wedge c_j = j$ (столы индексируются объектом с наименьшим номером)

$r_{ij}(\mathbf{c}) = 1$ - существует путь из i в j

$c_i \in \mathbb{R}^N, \quad c_i = j \Leftrightarrow c_{ij} = 1$

Будем искать апостериорное распределение $q(\mathbf{c}, \mathbf{\Theta}) = \prod_i^N q(c_i) \prod_j^N q(\theta_i) \approx p(\mathbf{c}, \mathbf{\Theta}|\mathbf{x})$

$\min_q D_{KL}(q||p(\mathbf{c}, \mathbf{\Theta}|\mathbf{x})) \Leftrightarrow \max_q \mathcal{L} \leq \log p(\mathbf{x})$

Покоординатная оптимизация q(.)

Зафиксируем все распределения, будем оптимизировать $q(c_s)$

Согласно основной формуле вар. вывода: $\log q(c_s) \propto \mathbb{E}_{q(\mathbf{c}_{\setminus s})q(\mathbf{\Theta})} \log p(\mathbf{x}, \mathbf{c}, \mathbf{\Theta})$

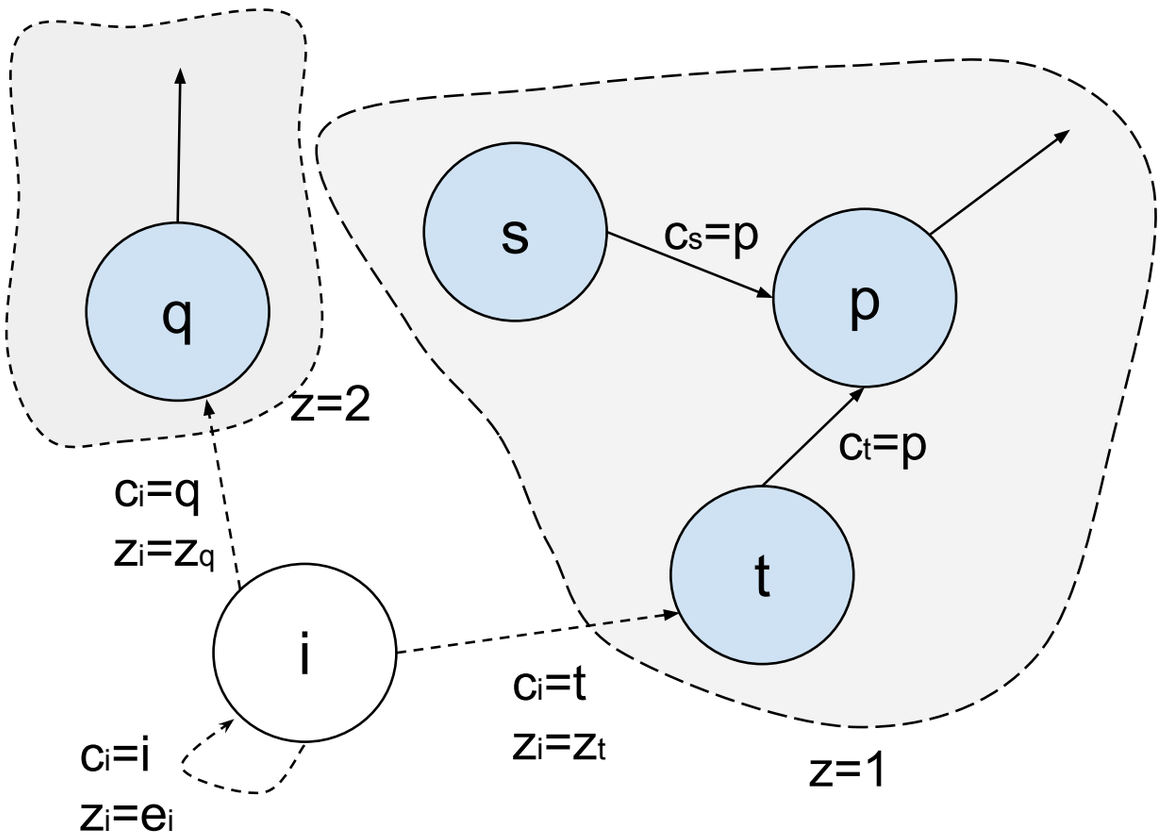
Итого, формула для пересчета:

$$\log q(c_s) \propto \log p(c_s) + \sum_{j<s} \left[\mathbb{E}_{q(\theta_j)} \log p(\theta_j|G_0) + \sum_{i>s} \mathbb{E}_{q(\mathbf{c}_{\setminus s})} z_{ij}(\mathbf{c}) \mathbb{E}_{q(\theta_j)} \log p(x_i|\theta_j)\right]$$

Пересчет распределений на параметры смеси:

$$\log q(\theta_j) \propto \log p(\theta_j|G_0) + \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{c})} z_{ij}(\mathbf{c}) \log p(x_i|\theta_j)$$

Расчет мат. ожиданий **z**(**c**)



1. $c_i = (c_{i1}, \ldots, c_{iN}) \to (q(c_i = 1), \ldots, q(c_i = N))$
   $z_i = (z_{i1}, z_{i2}, \ldots, z_{iN})$

2. $z_i = z_{c_i}$

3. $z_i = c_{ii} e_i + \sum_{j \neq i} c_{ij} z_j$

4. $\mathbf{z} = diag(\mathbf{c}) + (\underbrace{\mathbf{c} - diag(\mathbf{c})}_{=A})\mathbf{z}$

5. $\mathbf{z} = (I - A)^{-1} diag(\mathbf{c})$

$$\mathbf{r} = (I - A)^{-1} = L^{-1}$$

Градиентная процедура вывода

- $q(c_s)$ - дискретное распределение

- Будем оптимизировать $\mathcal{L}$ как функцию от N^2 аргументов с ограничениями

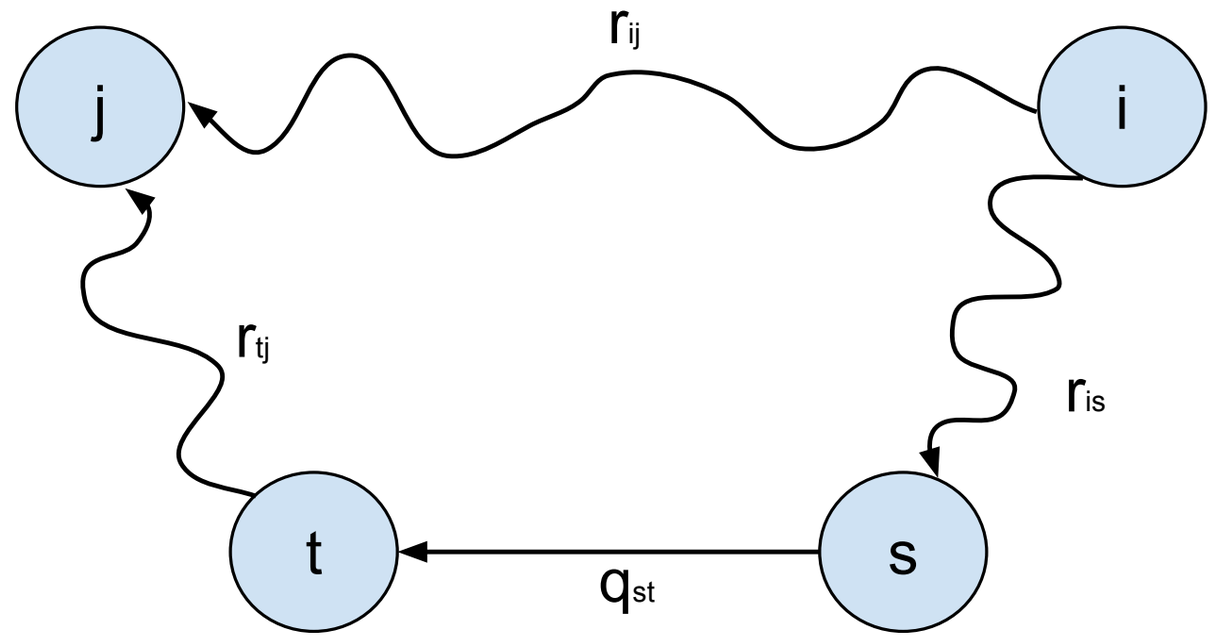
- Метод проекции градиента:

$$q(\mathbf{c})^{t+1} = P(q(\mathbf{c})^t + \alpha_t \nabla \mathcal{L})$$

- После шага по градиенту, каждое распределение проецируется на симплекс

Производная z(c)

- $$\frac{\partial \mathbb{E} z_{ij}(\mathbf{c})}{\partial q_{st}} = q_{jj} \mathbb{E} r_{is}(\mathbf{c}) \mathbb{E} r_{tj}(\mathbf{c}), \quad s \neq j, t \neq j$$



- Подсчет градиента требует только одного обращения матрицы

Языковая модель на основе seqddCRP

1. Каждая лексема выбирает, кого взять за руку

$$c_i \sim p(c_i | f, D, \alpha)$$

2. Детерминировано объединяются столы

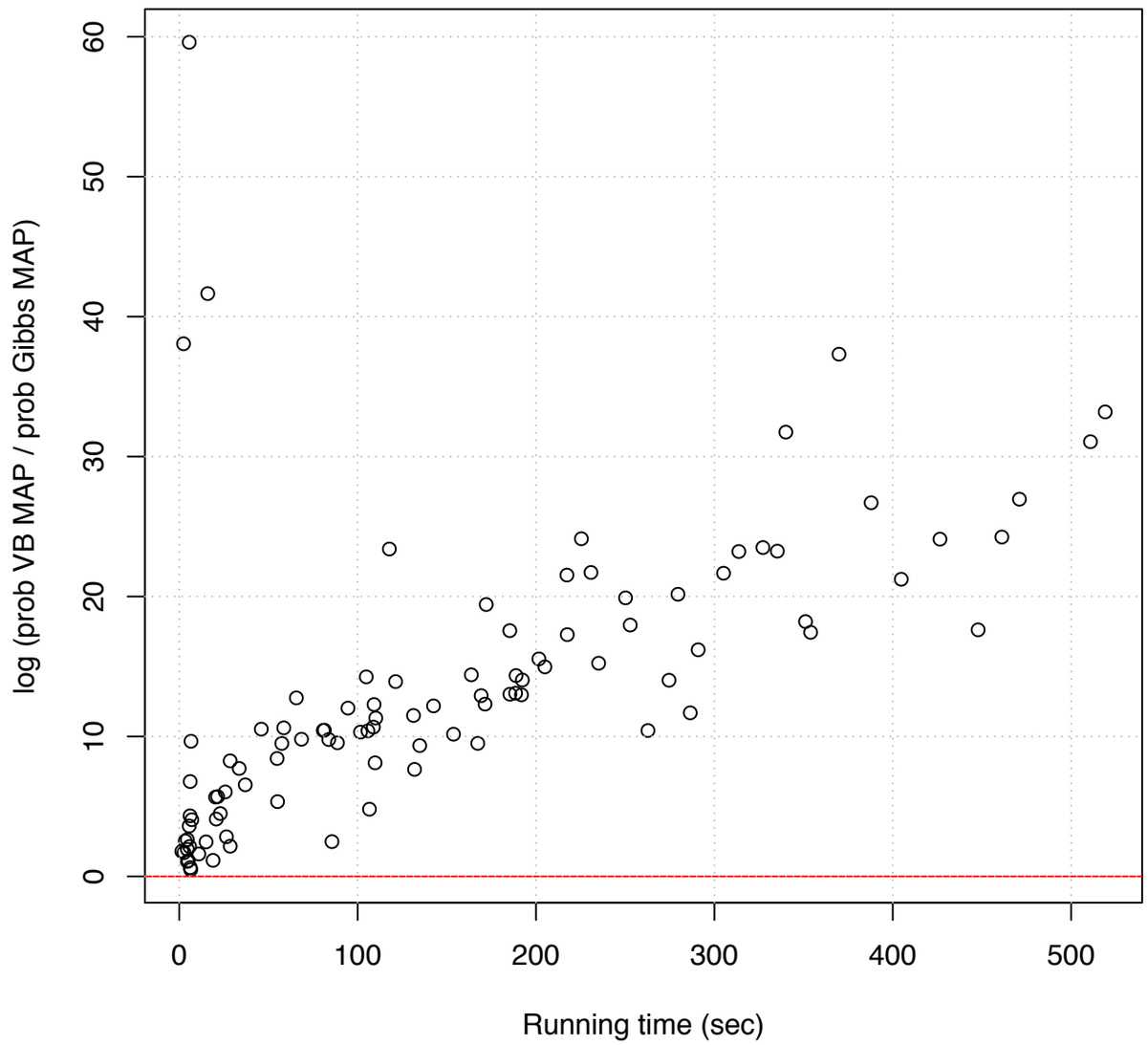
$$\mathbf{z} = z(\mathbf{c})$$

3. Для всех лексем за столом выбирается слово

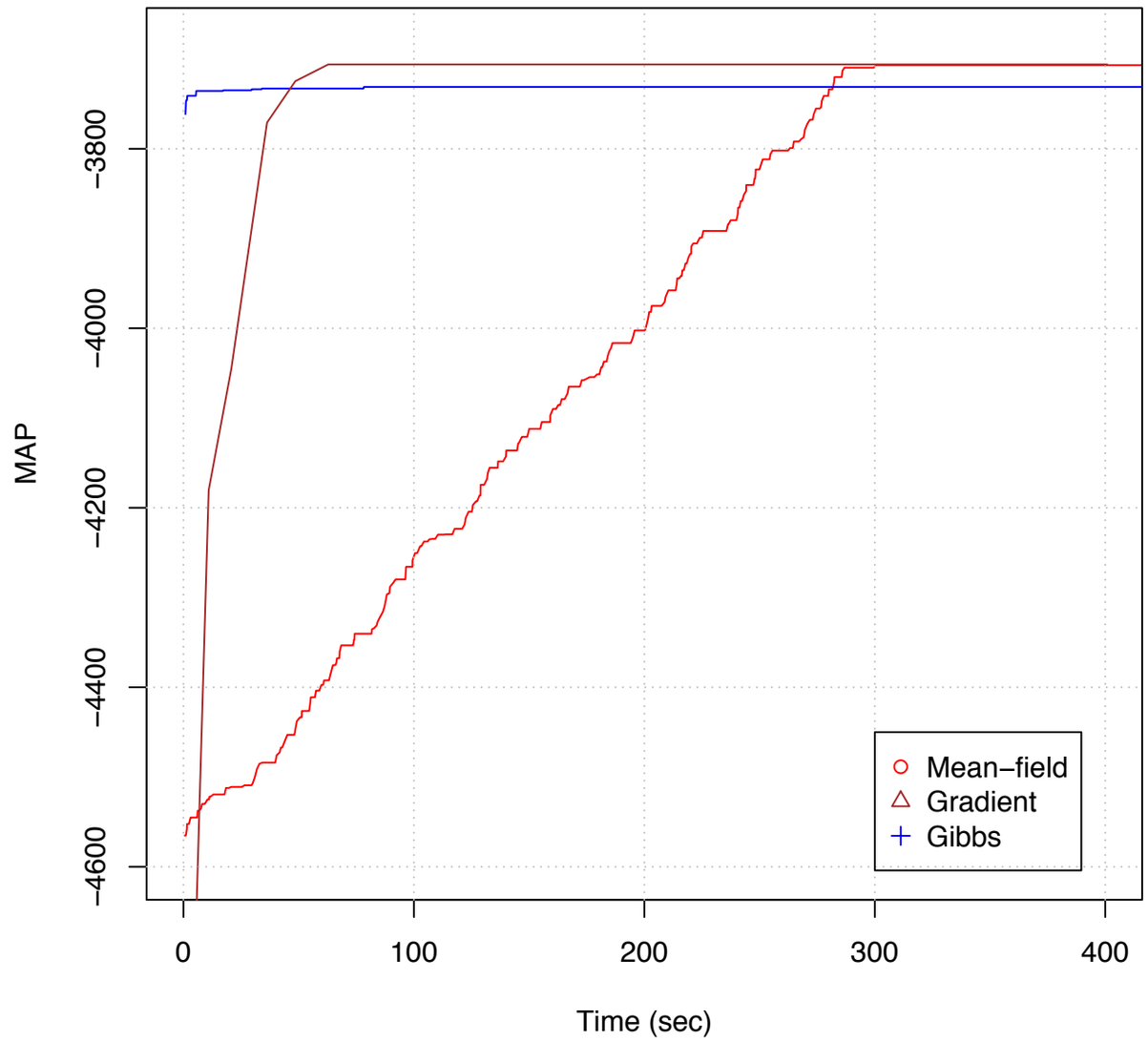
$$\mathbf{w}_{z_k(\mathbf{c})} \sim p(w)$$

Нет скрытых переменных

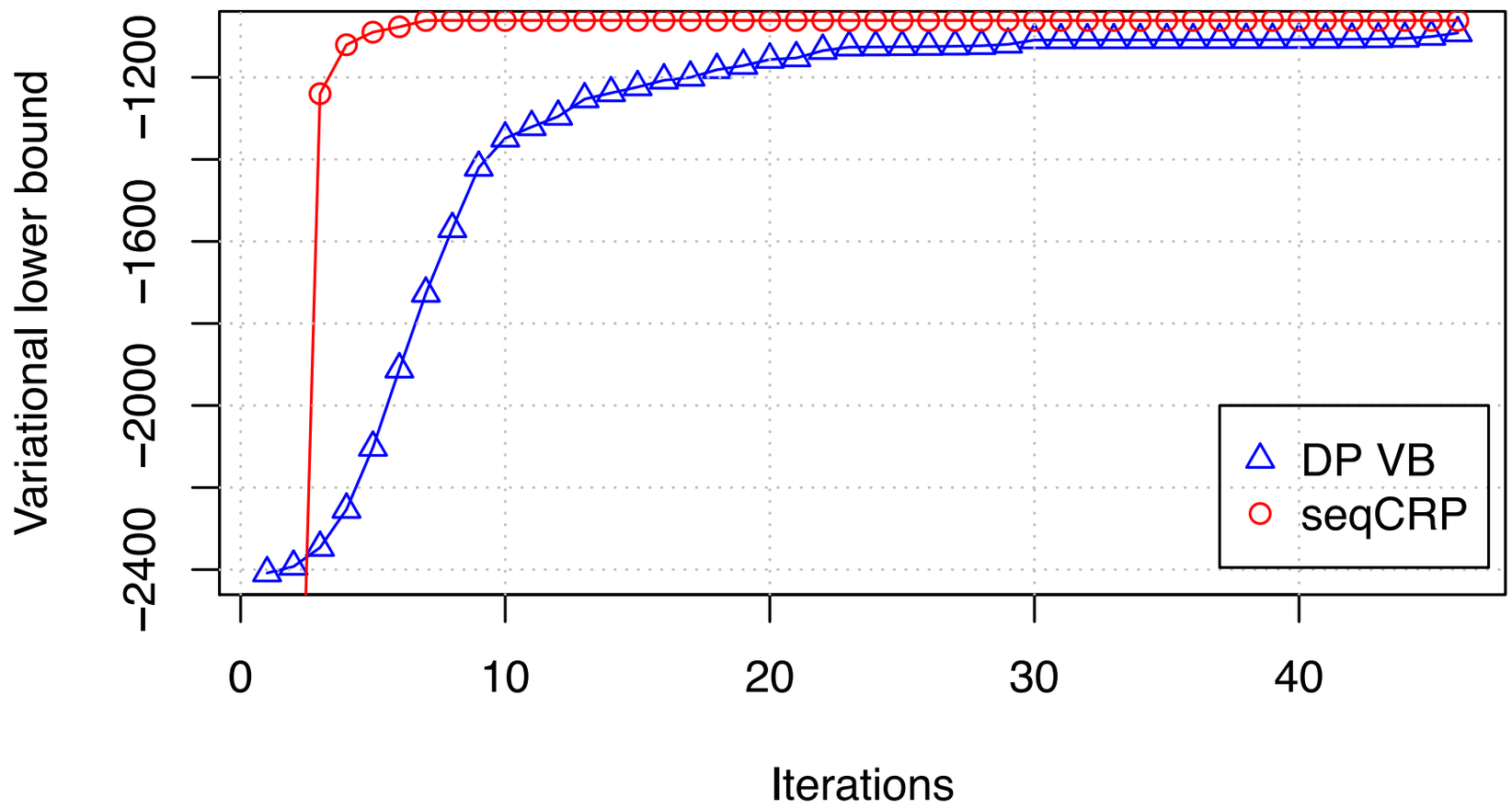
Гиббс-сэмплер против вариационного вывода



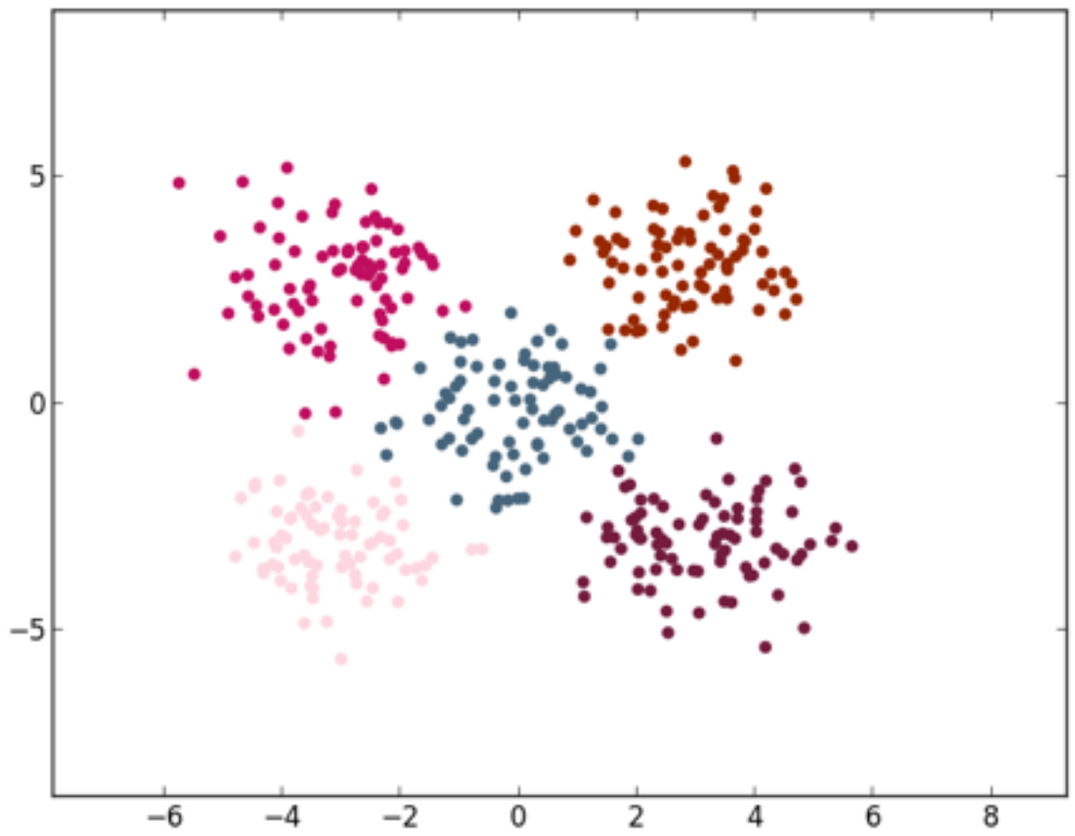
Сравнение скорости поиска моды



Вариационная нижняя оценка на задаче разделения CRP-смеси



seqddCRP против seqCRP



$$p(\mathbf{x}_{\text{test}}|\mathbf{x}, \hat{\Theta}) = \sum_{\mathbf{c}_{\text{test}}, \mathbf{c}} p(\mathbf{c}|\mathbf{x}, \hat{\Theta}) p(\mathbf{c}_{\text{test}}) p(\mathbf{x}_{\text{test}}|\mathbf{c}_{\text{test}}, \mathbf{c}, \hat{\Theta})$$

$$\hat{\Theta} = \mathbb{E}_q \Theta$$

$^{\text{algorithm}}/_{R}$	seqddCRP	seqCRP
1	-689.38	-691.96
2	-825.26	-836.58
3	-864.93	-888.10
4	-900.70	-905.75
5	-863.55	-965.76

Приложение #1

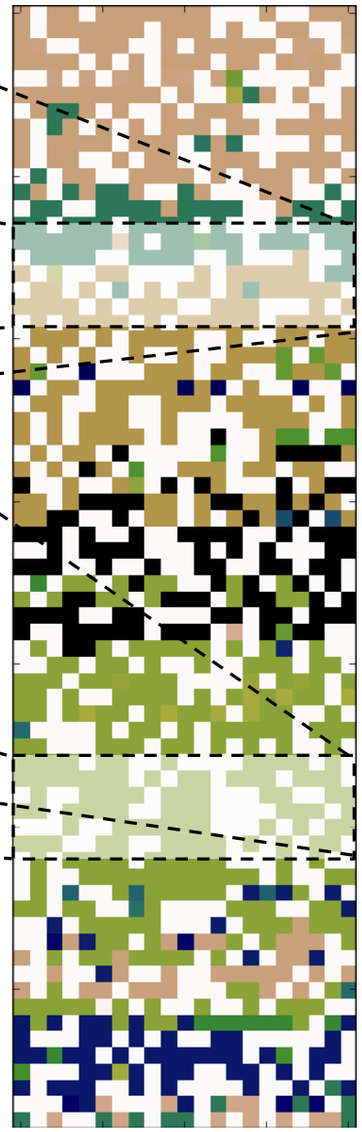
- Твиты очень короткие, поэтому методы типа LDA плохо работают
- С помощью seqddCRP можно вероятностно объединять твиты в документы
- Расстояния могут задавать время, расстояния в графе подписки, хештеги и многое другое

#ICML accepted papers are finally online. @vova could get one in, congrats!	We are going to the worst #belka **bar** in Moscow to **celebrate** this	First I need to read one named “**bayesian nonparametrics** for suicide **detection**”	Wow, they introduce new **stochastic process** for failed suicides	Funny picture of **drunk** and **happy** me. jpg	Can anyone on campus bring me **pills** from **headache**?	I’m ready for any **side effects** especially **amnesia**...
19:12 home	19:20 home	19:21 home	20:03 home	22:50 belka bar	9:04 cs lab	9:05 cs lab

Приложение #2

The natural chemical balance of grapes lets them ferment without the addition of sugars, acids, enzymes, water, or other nutrients. The term ”wine” can also refer to starch-fermented or fortified beverages having higher alcohol content, such as barley wine or sake.

Examples of recognized non-European locales include Napa Valley and Sonoma Valley in California; Willamette Valley in Oregon; Columbia Valley in Washington; Barossa Valley in South Australia and Hunter Valley in New South Wales; Luján de Cuyo in Argentina;



Заключение

- (seq)ddCRP это нов непараметрическое распределение, учитывающее зависимости в данных
- Можно представить, что (seq)ddCRP - это что-то вроде скрытой марковской модели с бесконечным числом состояний и немного другой моделью переходов
- Предложен вариацонный вывод для seqddCRP и, как следствие, альтернатива stick-breaking для процесса Дирихле