CTMP - Current Scientific Discussion Example

Collaborative Topic Model for Poisson distributed ratings
Hoa M. Le, Son Ta Cong, Quyen Pham The, Ngo Van Linh, Khoat Than
International Journal of Approximate Reasoning 95 (2018) 62–76

Introduction:

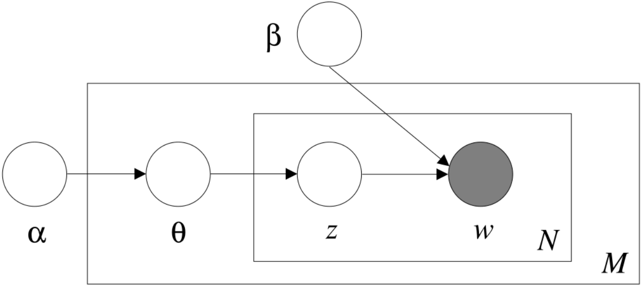
Latent Dirichlet Allocation
David M. Blei, Andrew Y. Ng and Michael I. Jordan
Journal of Machine Learning Research 3 (2003) 993-1022

Topic modeling: Formally, we define the following terms:

- **word** (slovo) - an item from a vocabulary $\{1, \ldots, V\}$, vektor s právě jednou jedničkou
- **document** - a sequence of N words $\mathbf{w} = (w_1, \ldots, w_N)$
- A **corpus** is a collection of M documents denoted by $D = \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M$.

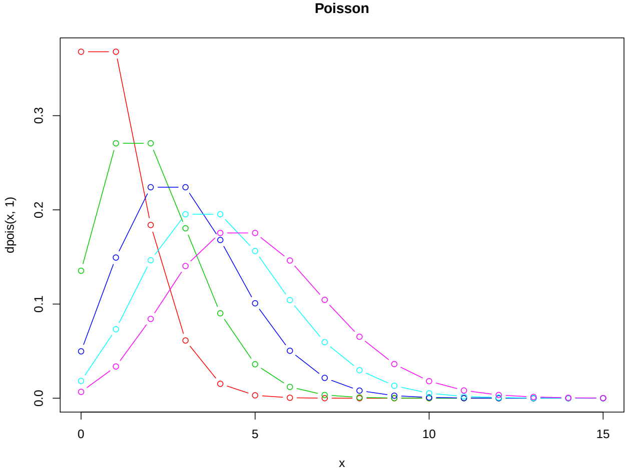
LDA Latend Dirichlet Allocation

- LDA assumes the following generative process for each document **w** in a corpus D :
  - Choose $N \approx Poisson(\xi)$.
  - Choose $\theta \approx Dirichlet(\alpha)$.
  - For each of the N words w_n :
    - Choose a topic $z_n \approx Multinomial(\theta)$ ($Categorical(\theta)$).
    - Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n



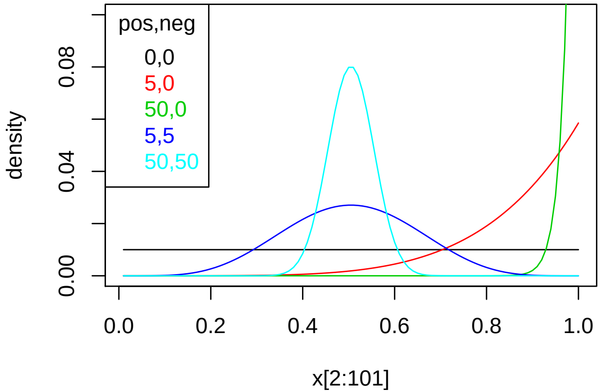
Document length - Poisson distribution

- ξ- rate; $p(N) = \frac{\xi^N e^{-\xi}}{N!}$
- $E(p(N)) = \xi$



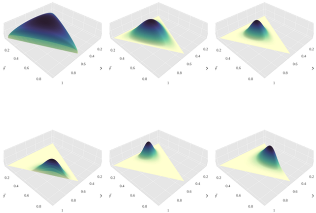
Beta distribution - Positive and negative exampleees

- Beta distribution



Document topic ratios θ - Dirichlet distribution

- Generalized Beta distribution
- Parameters $K \geq 2$ number of categories (integer), $\alpha_1, \ldots, \alpha_K$ concentration parameters, where $\alpha_i > 0$
- $\theta_1, \ldots, \theta_K$ where $\theta_i \in (0, 1)$ and $\sum_{i=1}^{K} \theta_i = 1$
- PDF $\frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}$ where $\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$



(clockwise, starting from the upper left corner): (1.3, 1.3, 1.3), (3,3,3), (7,7,7), (2,6,11), (14, 9, 5), (6,2,6)

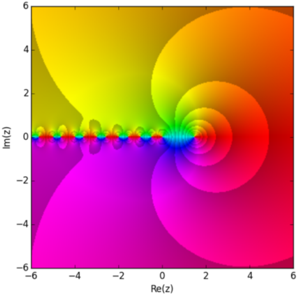
Expectation $ln(\theta)$ - Digamma function

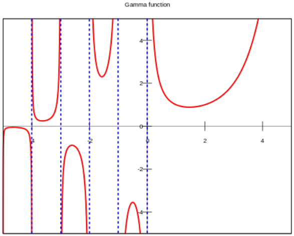
$$E[\ln \theta_i] = \psi(\alpha_i) - \psi(\sum_k \alpha_k)$$

$$\psi(x) = \frac{d}{dx} ln(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)}$$

$$\Gamma(n) = (n-1)!$$

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$$





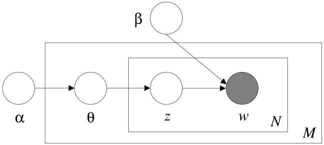
Topic z_n and word w_n probabilities

Uff. Konečně diskrétní.

- z_n one topic; Multinomial with probabilities $\theta, \sum_i \theta_i = 1$
  - categorical (r=1): select z_n according probabilities θ.
  - $p(x) = \theta_1^{[x=1]} \cdots \theta_k^{[x=k]}$
  - binomial: $k = 2$, number of successes in r trials.
- Multinomial - r samples, histogram $x_1, \ldots, x_k$:
  - $f(x_1, \ldots, x_k | r, \theta_1, \ldots, \theta_k) = \frac{r!}{x_1! \cdots x_k!} {\theta_1}^{x_1} \cdots {\theta_k}^{x_k}$
  - $f(x_1, \ldots, x_k | \theta_1, \ldots, \theta_k) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^{k} \theta_i^{x_i}$

Word probability

- discrete conditional $\beta_{ij} = p(w_j = 1 | z_i = 1)$



Document, Corpus probability

- Join probability for a single document

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

- document 'marginal' probability

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) d\theta$$

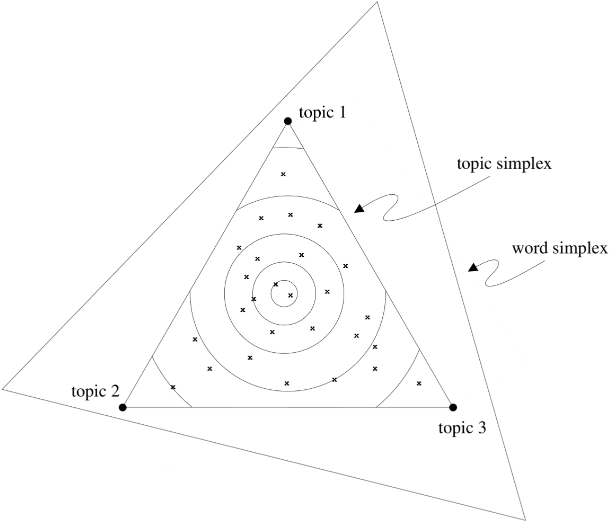
- corpus 'marginal' probability

$$p(\mathbf{D}|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta) p(w_{dn}|z_{dn}, \beta) d\theta_d$$

We search α, β maximizing $p(\mathbf{D}|\alpha, \beta)$.

Word and topic sipmlex

The topic simplex for three topics embedded in the word simplex for



three words

Inference

- First, we need θ and z for a given document ('Estimation of hidden variables').
- We want:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}$$

$$\begin{aligned} p(\mathbf{w}|\alpha, \beta) &= \int p(\theta|\alpha) \prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) d\theta \\ &= \frac{\Gamma(\sum_i \alpha_i + 1)}{\prod_i \Gamma(\alpha_i + 1)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta \end{aligned}$$

Calculation intractable, we use approximation.

Variational Inference

- we remove edges - coupling between θ, z, w
- we consider set of distributions parametrized by γ, ϕ_n:

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n)$$

- finding a thight lower bound on log-likehood corresponds to minimizing KL-divergence D:

$$(\gamma^*, \phi^*) = argmin_{(\gamma,\phi)} D(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta))$$

- by setting derivatives zero we get:

$$\phi_{ni} \leftarrow \beta^n_{iw} exp E_q[log(\theta_i)|\gamma] = \beta^n_{iw} exp(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)$$

$$\gamma_i \leftarrow \alpha_i + \sum_{n=1}^{N} \phi_{ni}$$

- complexity roughly $O(N^2 k)$.

LDA Hidden Variables Estimation

initialize $\phi_{ni}^0 := 1/k$ for all i and n
initialize $\gamma_i := \alpha_i + N/k$ for all i
repeat
  for $n = 1$ **to** N
    for $i = 1$ **to** k
      $\phi_{ni}^{t+1} := \beta_{iw_n} \exp(\Psi(\gamma_i^t))$
    normalize ϕ_n^{t+1} to sum to 1.
  $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$
until convergence

Parameter Estimation (Learning)

$$LL(\alpha, \beta; D) = \sum_{d=1}^{M} \log \; p(w_d|\alpha, \beta)$$

(variational) EM-algorithm

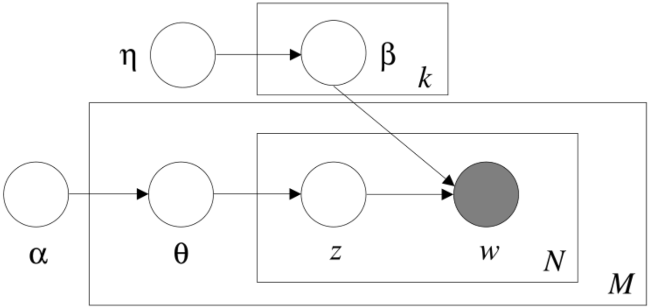
- E: find $\{(\gamma_d^*, \phi_d^*|d \in D)\}$
- M:
  - $\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni}^* w_{dn}^j$
  - α iteratively (Newton-Raphson method)

Vyhlazování

- Přidáme 'prior' β a nový variational parametr λ a aproximujeme pomocí:

$$q(\beta_{1:k}, z_{1:M}, \theta_{1:M}|\lambda, \phi, \gamma) = \prod_{i=1}^{k} Dirichlet(\beta_i|\lambda_i) \prod_{d=1}^{M} q_d(\theta_d, z_d|\phi_d, \gamma_d)$$

- We get the update: $\lambda_{ij} \propto \eta + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^j_{dn}$



Dirichlet-multinomial

$\alpha_0 = \sum \alpha_k$

$$P(x|\alpha) = \frac{n!\Gamma(\alpha_0)}{\Gamma(n+\alpha_0)} \prod_{k=1}^{K} \frac{\Gamma(x_k+\alpha_k)}{x_k!\Gamma(\alpha_k)} = \frac{nBeta(\alpha_0, n)}{\prod_{k:x_k>0} Beta(\alpha_k, x_k)}$$

LDA: Z topics, n_v^k number of word v in topic k

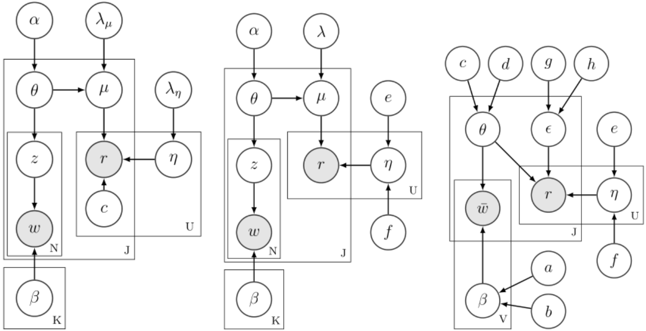
$$P(W|\alpha, Z) = \prod_{k=1}^{K} DirMult(W_k|Z, \alpha) = \prod_{k=1}^{K} \frac{\Gamma(\sum_v \alpha_v)}{\Gamma(\sum_v n_v^k + \alpha_v)} \prod_{v=1}^{V} \frac{\Gamma(n_v^k + \alpha_v)}{\Gamma(\alpha_v)}$$

Urn model

- vracím 2 od barvy: Dirichlet-multinomial
- vracím 1 od barvy: multinomial
- nevracím: multivariate hypergeometric distribution

CTMP

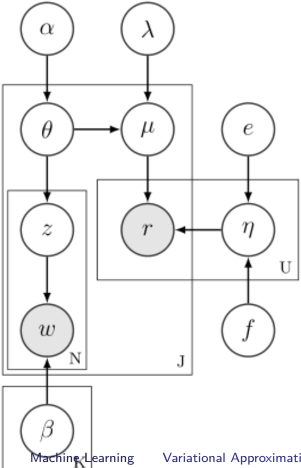
Collaborative Topic Model for Poisson distributed ratings
Hoa M. Le, Son Ta Cong, Quyen Pham The, Ngo Van Linh, Khoat Than
International Journal of Approximate Reasoning 95 (2018) 62–76



(a) CTR (b) CTMP (c) CTPF

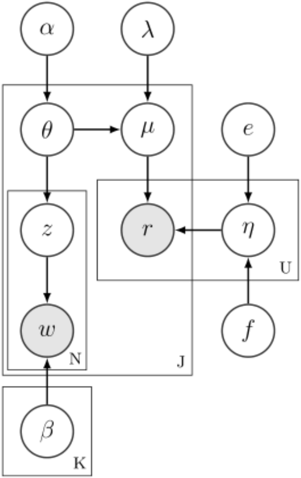
CTR- gaussian R, CTPM - poisson R.

- For each user u, draw η_u where $\eta_{uk} \propto Gamma(e, f)$
- For each item j:
  - (a) Draw topic proportion $\theta_j \propto Dirichlet(\alpha)$
  - For the n-th word of item j:
    - Draw topic index $z_{jn} \propto Categorical(\theta)_j$
    - Draw word $w_{jn} \propto Categorical(\beta_{z_{jn}})$
    - Draw latent factor $\mu_j \propto N(\theta_j, \lambda^{-1} I_K)$
  - For each user-item pair (u, j) , draw $r_{uj} \propto Poisson(\eta_u^T \mu_j)$



Predictive Score

- For a given document j and a user u we predict a score s_{ju}
- $s_{ju} \approx \mu_j \cdot \frac{rte_u}{shp_u}$.
- $\frac{rte_{uk}}{shp_{uk}} = E_{q(\eta_{uk}|shp_{uk}, rte_{uk})}[\eta_{uk}] \approx E[\eta_{uk}|D, \mu_{jk}]$



- To be honest, intermediate variable y, which $r_{uj} = \sum_{k=1}^{K} y_{ujk}$ and $y_{ujk} \approx Poisson(\eta_{uk}\mu_{jk})$

Log likelihoood to be optimized

- $$L = \log P(\theta, \mu, D|\alpha, \beta, \lambda, e, f) = \sum_{j=1}^{J} \log P\left(\theta_j, \mu_j, w_j|\alpha, \beta\right) + \sum_{u=1}^{U}\sum_{j=1}^{J} \log P(r_{uj}|\mu_j, e, f)$$
$$= \sum_{j=1}^{J} \log P(\theta_j, w_j|\alpha, \beta) + \sum_{j=1}^{J} \log P(\mu_j|\theta_j, \lambda)$$
$$+ \sum_{u=1}^{U}\sum_{j=1}^{J} \log \int \sum_{y_{uj}} P(r_{uj}, y_{uj}, \eta_u|\mu_j, e, f) d\eta_u$$

- last term approximated by

$$q(\eta_u, y_{uj}) = q(y_{uj}|r_{uj}, \phi_{uj}) \prod_{k=1}^{K} q(\eta_{uk}|shp_{uk}, rte_{uk})$$

- with distributions $q(y_{uj}|r_{uj}, \phi_{uj}) = Mult(y_{uj}|r_{uj}, \phi_{uj})$, $q(\eta_{uk}|shp_{uk}, rte_{uk}) = Gamma(\eta_{uk}|shp_{uk}, rte_{uk})$

Function to be maximized

$$
\begin{aligned}
l(\theta, \mu, \phi, \text{shp}, \text{rte}, \beta) = & \sum_{j}^{J}((\alpha - 1)\sum_{k}^{K} \log \theta_{jk} + \sum_{v}^{V} c_j^v \log \sum_{k}^{K} \theta_{jk}\beta_{kv}) - \sum_{j}^{J} \frac{\lambda}{2} \|\theta_j - \mu_j\|_2^2 \\
& + \sum_{u}^{U}\sum_{j}^{J}\sum_{k}^{K} r_{uj}\phi_{ujk} \log(\mu_{jk}) - \sum_{u}^{U}\sum_{j}^{J}\sum_{k}^{K} r_{uj}\phi_{ujk} \log(\phi_{ujk}) \\
& + \sum_{u}^{U}\sum_{k}^{K}(\text{rte}_{uk} - f - \sum_{j}^{J} \mu_{jk}) \frac{\text{shp}_{uk}}{\text{rte}_{uk}} \\
& + \sum_{u}^{U}\sum_{k}^{K}(\sum_{j}^{J} r_{uj}\phi_{ujk} + e - \text{shp}_{uk})(\Psi(\text{shp}_{uk}) - \log(\text{rte}_{uk})) \\
& - \sum_{u}^{U}\sum_{k}^{K} \text{shp}_{uk} \log(\text{rte}_{uk}) + \sum_{u}^{U}\sum_{k}^{K} \log(\Gamma(\text{shp}_{uk})) + Constant.
\end{aligned}
$$

Algorithm 1 Learning CTMP by coordinate ascent.

Input: Observed data w, r and hyperparameters α, λ, e, f
Output: Estimates θ, μ, ϕ_{uj}, shp_{uk}, rte_{uk} and β
  init Initialize θ, β by their respective estimates from LDA [8]
  repeat
    for $j = 1 : J$ **do**
      Update θ_j by Algorithm 2
      Update μ_j as in Equation (8)
    end for
    for $u = 1 : U, k = 1 : K$ **do**
      Update variational parameters as in Table 2
      $\phi_{ujk} \propto \exp\left\{\log \mu_{jk} + \psi\left(\text{shp}_{uk}\right) - \log\left(\text{rte}_{uk}\right)\right\} \; \forall j \text{ if } r_{uj} > 0$
      $\text{shp}_{uk} \leftarrow e + \sum_j r_{uj}\phi_{ujk}$
      $\text{rte}_{uk} \leftarrow f + \sum_j \mu_{jk}$
    end for
    $\beta_{kv} \propto \sum_j c_j^v \theta_{jk} \; \forall k, v$
  until convergence

$$\mu_{jk} = \frac{-\sum_u \frac{\text{shp}_{uk}}{\text{rte}_{uk}} + \lambda\theta_{jk} + \sqrt{\Delta}}{2\lambda}$$

$$\text{where } \Delta = \left(-\sum_u \frac{\text{shp}_{uk}}{\text{rte}_{uk}} + \lambda\theta_{jk}\right)^2 + 4\lambda \sum_k r_{uj}\phi_{ujk}$$

Algorithm 2 Topic proportion θ_j estimation by OPE.

Input: $\lambda,\ \beta,\ \mu_j, \alpha, w_j = \{c_j^v\}_{v=1}^V$

Output: θ_j that maximizes $g\left(\theta_j\right)$

init Initialize $\theta_j^{(1)}$ arbitrarily in $\overline{\Delta}_K = \left\{x \in \mathbb{R}^K : \sum_k x_k = 1,\ x_k \geq \epsilon > 0\right\}$

for $t = 1, ..., \infty$ **do**

Draw $g^{(t)}$ uniformly from $\left\{-\frac{\lambda}{2} \left\|\theta_j - \mu_j\right\|_2^2;\ (\alpha - 1) \sum_k \log \theta_{jk} + \sum_v c_j^v \log\left(\sum_k \theta_{jk}\beta_{kv}\right)\right\}$

$G\left(\theta_j\right) \leftarrow \frac{1}{t} \sum_{h=1}^t g^{(h)}$

$e^{(t)} \leftarrow \arg\max_{e_k \in \overline{\Delta}_K} \left\langle \nabla G\left(\theta_j^{(t)}\right), e_k \right\rangle$

$\theta_j^{(t+1)} \leftarrow ae^{(t)} + (1 - a)\,\theta_j^{(t)}$ where $a = \frac{2}{t+2}$

end for

Experiments

Table 3
Statistics of the experimented datasets. *Sparsity* indicates proportion of the entries that do not have any positive ratings in each rating matrix *R*.

	#Users	#Items		#Ratings	Sparsity (%)	Vocab. (#words)	#Words per item	Corpus size (#words)
		TOTAL	cold-item					
CiteULike	5,551	16,980	3,396	204,986	99.78	8,000	66	1,120,680
Soha	99,211	49,234	26,725	12,929,892	99.74	45,435	134	6,597,356
Lazada	42,308	46,200	470	879,071	99.96	42,030	26	1,201,200
Muachung	13,034	12,615	8,541	210,612	99.87	4,100	25	315,375
Movielens 10M	69,876	10,681	293	9,520,883	98.72	21,218	32	341,792
Movielens 20M	138,248	27,278	1,853	19,761,133	99.47	33,736	30	818,340

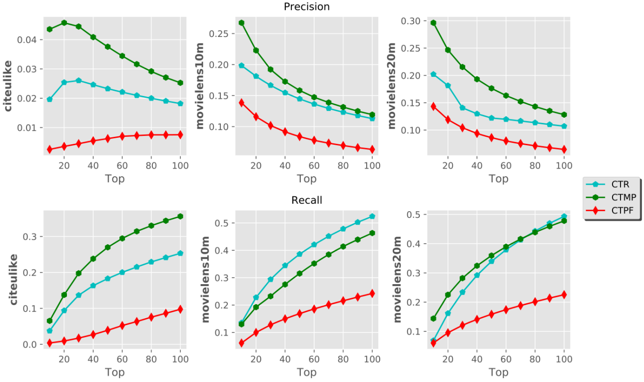


Fig. 5. Average precision and recall in top-10 to top-100 recommendation on out-of-matrix.

Sparse/general topics