潜在トピック間の因果構造のモデリング

宮澤　脩一　　総合研究大学院大学　統計科学専攻　博士課程(5年一貫制)5年

【目的】

文書の潜在トピックを推定する確率的トピックモデルにおいて，文単位のメイントピックから単語単位のサブトピックが生成する2重の階層化を行い，トピック間の因果構造を推定する手法を確立する.

【関連研究】

Correlated Topic Model (CTM)

トピック間の相関構造を多変量対数正規分布の共分散行列Σとして推定するトピックモデル．提案モデルでは対象な共分散行列ではなく，非対称な重み付き隣接行列として方向性を持つトピック間の因果性を推定する.

Sparse Topic Model (sparseTM)

多くのトピックモデルでトピック-単語分布は，語彙サイズの次元を持つ非ゼロ成分のみから成る確率ベクトルであるが，spraseTMではベータ-ベルヌーイ過程で生成したスパースなベクトルに係数を掛け，ディリレクパラメータとして用いることで，多くの成分をゼロとするスパースなトピック単語分布を生成する．提案モデルにおいても，メイントピック→サブトピックの組み合わせによって異なるスパースな単語分布を扱えるようなモデル化を行う.

【モデル】

トピックの生成過程

文書 d の各文 s にメイントピック x があり，そのメイントピックに従って各単語のサブトピック z を生成する．メイントピックとサブトピックで同じ K 種のトピック集合を共有し，あるメイントピック j からは同じインデックスのサブトピック $k=j$ が最も生成しやすいとする．また，連続する文 $s, s+1$ で同じメイントピック j を扱いやすいという仮定を置き，文間のメイントピックの遷移を隠れマルコフモデルでモデル化する．これらの同じトピックインデックスを生成しやすいという性質を切断正規分布からの潜在変数の生成によってモデル化する.

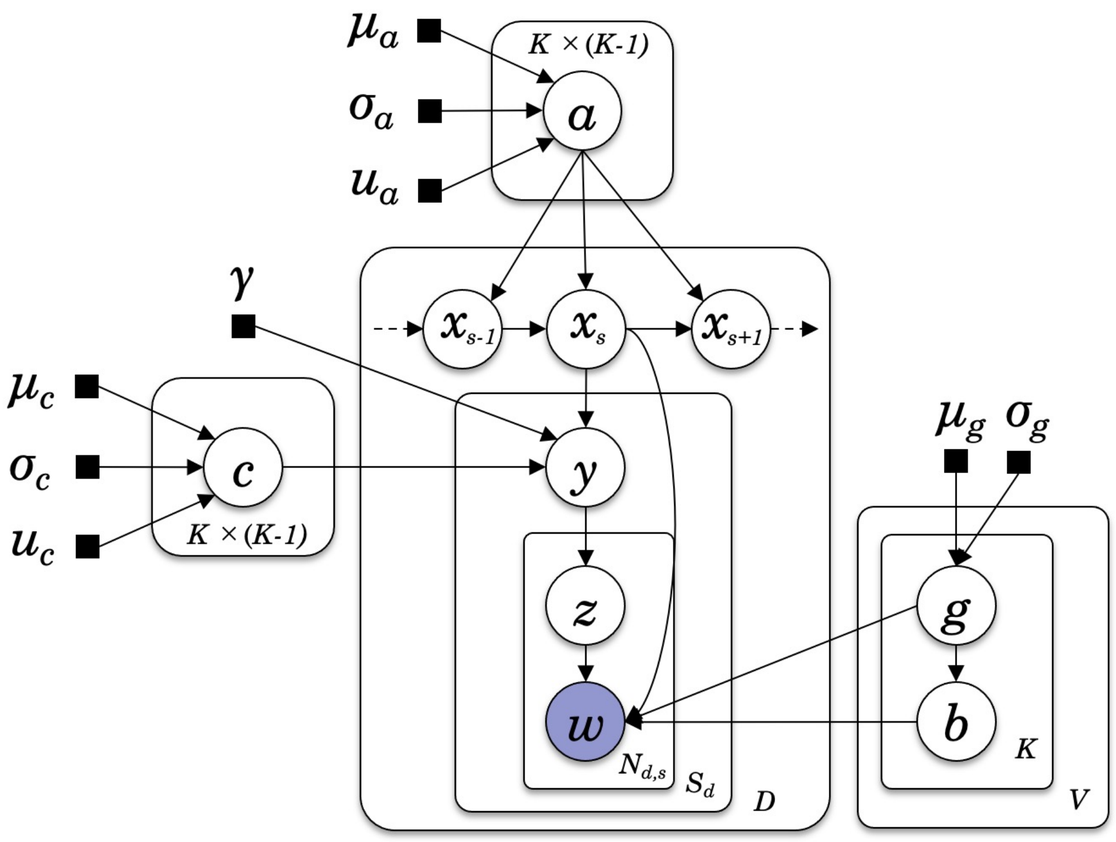
トピック単語分布の生成

各トピック k について，正規分布から生成する単語選好ベクトル $\boldsymbol{g}_k$ およびベルヌーイ分布から生成する単語フィルタ one-hot ベクトル $\boldsymbol{b}_k$ を割り当て，メイントピック j →サブトピック k の経路で生成する単語の分布は，メイントピック j の単語フィルタ $\boldsymbol{b}_j$ およびサブトピック kの単語選好$\boldsymbol{g}_k$を用いて以下のように定義する.

$$\boldsymbol{g}_k \in \mathbb{R}^V \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \qquad \boldsymbol{b}_k \in \{0,1\}^V \sim \mathrm{Bern}(\mathrm{sigmoid}(\boldsymbol{g}_k))$$

メイントピックとサブトピックが一致する/異なる場合

$$\boldsymbol{\phi}_{j \to j} = \frac{\exp(\boldsymbol{g}_j)}{\sum_{v=1}^{V} \exp(g_{jv})} \qquad \boldsymbol{\phi}_{j \to k} = \frac{\exp(\boldsymbol{g}_k) \circ \boldsymbol{b}_j}{\sum_{v=1}^{V} \exp(g_{kv}) b_{jv}}$$



提案モデルのグラフィカルモデル

文間のメイントピックの遷移行列 $\boldsymbol{A}$ の生成

$$a_{j,j} = u_a \; \forall j = 1, \cdots, K \qquad a_{j,j' \neq j} \sim \mathcal{TN}(0, \sigma_a^2, -\infty, a_{j,j})$$

文のメイントピック系列 $\boldsymbol{x}$ の生成

$$x_{d,1} \sim \mathrm{rand}(1, \cdots, K) \qquad x_{d,s} \sim \mathrm{Cat}(\mathrm{softmax}(\boldsymbol{a}_{y_{d,s-1},\cdot}))$$

メイントピック→サブトピック因果構造行列 $\boldsymbol{C}$ の生成

$$c_{j,j} = u_c \; \forall j = 1, \cdots, K \qquad c_{j,k \neq j} \sim \mathcal{TN}(\mu_c, \sigma_c^2, -\infty, c_{j,j})$$

各文のサブトピック分布 $\boldsymbol{\theta}_{d,s}$ の生成

$$\boldsymbol{\theta}_{d,s} = \mathrm{softmax}(\boldsymbol{y}_{d,s}) \qquad \boldsymbol{y}_{d,s} \sim \mathcal{N}(\boldsymbol{c}_{x_{d,s},\cdot}, \gamma^2)$$

各単語のサブトピック z および単語 w の生成

$$z_{d,s,n} \sim \mathrm{Cat}(\boldsymbol{\theta}_{d,s}) \qquad w_{d,s,n} \sim \mathrm{Cat}(\boldsymbol{\phi}_{x_{d,s} \to z_{d,s,n}})$$

【実験】

データ：NIPS corpus． Titleに ”bayes” を含む論文のAbstract.

データ詳細：文書数 $D = 330$，語彙数 $V = 1789$，総文数 $\sum_{d=1}^{D} S_d = 5373$，総トークン数 $\sum_{d=1}^{D} \sum_{s=1}^{S_d} N_{d,s} = 31204$.

パラメータ：トピック数 $K = 30$，その他 $\gamma = 1$，$\mu_a = 0, \sigma_a = 3$，$u_a = 5$，$\mu_c = 0$, $\sigma_c = 3$，$u_c = 5$.

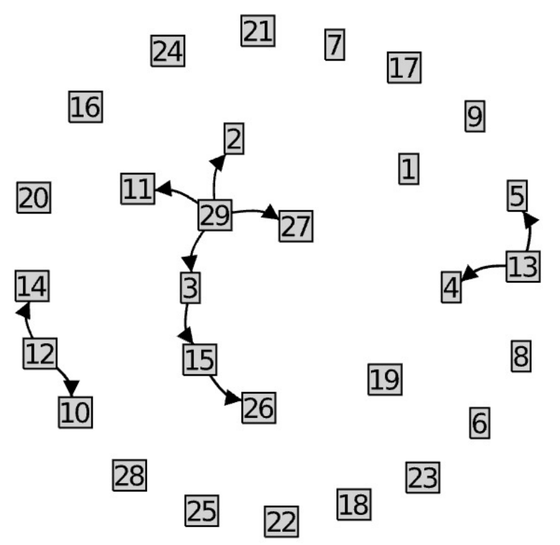
推論：MCMCで1000イテレーション推論を実行

【結果】

推定されたトピックの代表単語の例
（正規化自己相互情報量を指標として代表単語を選出）

id	代表単語
2	social networks, state-of-the-art, artifact, markov decision processes, model fitting
3	complicated, superiority, new york, strengthen, probe, two-fold, hinton, communicate
4	pose, inference scheme, effectiveness, window, factorization, minor, begin, page
5	hdp, review, state-of-the-art, devise, implicit, non-trivial, stochastic processes
10	good approximation, resonance, technology, university, non, precede, feasible
11	dramatic, stand, social networks, hamiltonian monte carlo, opposite
12	indian buffet process, elegant, fundamental, recovery, free energy, gather, statement
13	new york, justify, raise, empirical error, previous approaches, neuroscience, integer
14	means, operation, norm, one-dimensional, variational bayesian method, adjacent
20	springer, power-law, ap, exploitation, pp, suffer, organization, tion, act, count
27	dimensionality reduction, ganglion, high-quality, participate, true posterior, hybrid
29	inaccurate, feedback, approximate posterior, computer vision, variational bayesian method

- パラメータ設定や各確率項の尤度スケールのバランス調整により，推論結果の改善を検討中
- トピック単語分布のモデリングについては，他の方法も検討中
- メイントピック系列やメイントピック→サブトピックという単語の生成経路情報を用いることで，特に長い文書のコンテンツの解釈を容易にする手法の確立を目指す



トピック因果構造の例（頻度上位 10 件）

参考文献

[1] Blei, D., and Lafferty, J. Correlated topic models. In Advances in Neural Information Processing Systems 18. 2006.

[2] Wang, Chong and Blei, David M. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In Neural Information Processing Systems, 2009..