1W-02

# ハウスホルダーフローを導入したEmbedded Topic Modelに関する一考察

松苗亮汰† 山極綾子† 後藤正幸†

早稲田大学†

## 1. 背景と目的

トピックモデルは，単語の共起関係をモデル化し，文書から自動でトピックを抽出することが可能な統計モデルであり，様々な目的で用いられている．トピックモデルの一つであるEmbedded Topic Model（以下，ETM）[1]は，トピックと単語を同一空間上に埋め込むモデルであり，文書データの分析手法として有効である．ただし，ETMのトピック分布の推定には償却推論が用いられ，各変数間の共分散が0として変分分布が推論される．しかし実際には，トピック間に相関が存在すると考えられる．

上記の問題を解決するため，著者らはETMにハウスホルダーフロー[2]を適用し，トピック間の相関を表現可能なFlow-ETM[3]を提案した．Flow-ETMはETMよりも，より柔軟に入力文書に沿うトピック分布の生成が期待される．本稿においては，文書データセットを用いたより詳細な評価実験を行い，Flow-ETMの特性を解析し，その有効性を検証する．

## 2. 準備

### 2.1 ETM

ETM[1]は深層学習を用いたトピックモデルであり，単語とトピックを同一空間上に埋め込む．全$V$個の語彙の埋め込み表現を$\boldsymbol{P}=(\boldsymbol{\rho}_1,\cdots,\boldsymbol{\rho}_V)\in\mathbb{R}^{L\times V}$，全$K$個のトピックの埋め込み表現を$\boldsymbol{A}=(\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_K)\in\mathbb{R}^{L\times K}$と表す．ETMではトピック$k$の近くに埋め込まれた単語はトピック$k$のもとで生起しやすいことを意味し，トピック$k$の単語分布$\boldsymbol{\beta}_k$を埋め込み表現の内積を用いて$\boldsymbol{\beta}_k=\boldsymbol{P}^{\mathrm{T}}\boldsymbol{\alpha}_k$と記述する．また，全$D$個のうち$d$番目の文書を$\boldsymbol{w}_d\in\mathbb{R}^V$，文書$\boldsymbol{w}_d$のトピック分布を$\boldsymbol{\theta}_d\in\mathbb{R}^K$，文書$\boldsymbol{w}_d$の$N_d$個の単語のうち$n$番目を$w_{dn}\in\{1,\cdots,V\}$，$w_{dn}$のトピックを$z_{dn}$とする．ETMにおける文書$\boldsymbol{w}_d$の生成過程を以下に示す．なお，$\mathrm{Cat}(\cdot)$はカテゴリカル分布を表す．

Embedded topic model with householder flow

Ryota Matsunae† Ayako Yamagiwa† Masayuki Goto†
Waseda University†

なお，$\mathrm{Cat}(\cdot)$はカテゴリカル分布を表す．

I. トピック割合を生成 $\boldsymbol{\delta}_d\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$
II. トピック分布に変換 $\boldsymbol{\theta}_d=\mathrm{softmax}(\boldsymbol{\delta}_d)$
III. $N_d$回以下の処理を繰り返す
  i. トピックを生成 $z_{dn}\sim\mathrm{Cat}(\boldsymbol{\theta}_d)$
  ii. 単語を生成 $w_{dn}\sim\mathrm{Cat}(\boldsymbol{\beta}_{z_{dn}})$

ETMの学習では，変分分布$q(\boldsymbol{\delta}_d|\boldsymbol{w}_d)$を用いた変分下界を最大化する．変分分布$q(\boldsymbol{\delta}_d|\boldsymbol{w}_d)$には正規分布が用いられ，文書$\boldsymbol{w}_d$を入力とする推論ネットワークによって償却推論される．

### 2.2 ハウスホルダーフロー

ハウスホルダーフロー[2]とは，式(1)のように$\boldsymbol{T}$回のハウスホルダー変換を繰り返し行う変数変換手法である．ここで一般に，$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$に従う確率変数に対して行列$\boldsymbol{H}$で線形変換を行った場合，変換後の変数は$\mathcal{N}(\boldsymbol{H\mu},\boldsymbol{H\Sigma H}^{\mathrm{T}})$に従う．したがって，各変数が独立な正規分布$\mathcal{N}\left(\boldsymbol{\mu},\mathrm{diag}(\boldsymbol{\sigma}^2)\right)$に従う確率変数$\boldsymbol{\delta}^{(0)}$に対してハウスホルダーフローを適用することで，変数間で相関を持つ正規分布$\mathcal{N}(\boldsymbol{U\mu},\boldsymbol{U}\mathrm{diag}(\boldsymbol{\sigma}^2)\boldsymbol{U}^{\mathrm{T}})$に従う確率変数$\boldsymbol{\delta}^{(T)}$を得ることができる．ここで，$\boldsymbol{U}=\boldsymbol{H}^{(T)}\boldsymbol{H}^{(T-1)}\cdots\boldsymbol{H}^{(1)}$である．

$$\boldsymbol{\delta}^{(T)}=\boldsymbol{H}^{(T)}\cdots\boldsymbol{H}^{(2)}\boldsymbol{H}^{(1)}\boldsymbol{\delta}^{(0)}=\boldsymbol{U}\boldsymbol{\delta}^{(0)}$$
$$\boldsymbol{H}^{(t)}=\boldsymbol{I}-2\frac{\boldsymbol{v}^{(t)}\boldsymbol{v}^{(t)^{\mathrm{T}}}}{\|\boldsymbol{v}^{(t)}\|^2} \quad (1)$$

### 2.3. Flow-ETM

筆者らはETMにハウスホルダーフローを適用したFlow-ETMを提案している[3]．Flow-ETMでは，償却推論で得られたトピック割合$\boldsymbol{\delta}_d^{(0)}$に対し，長さ$\boldsymbol{T}$のハウスホルダーフローを適用することで，トピック間の相関を考慮した新たなトピック割合$\boldsymbol{\delta}_d^{(T)}$を得る．学習時には，式(2)の変分下界を最大化するように，埋め込み表現$\boldsymbol{P}$と$\boldsymbol{A}$，推論ネットワーク，ハウスホルダーフローのパラメータを推論する．また，事前分布$p\left(\boldsymbol{\delta}_d^{(T)}\right)$は標準正規分布$\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$とする．

$$\mathcal{L} = \sum_{d=1}^{D}\sum_{n=1}^{N_d} \mathbb{E}_q\left[\log p\left(w_{dn}\middle|\boldsymbol{\delta}_d^{(T)}\right)\right] - \sum_{d=1}^{D} \mathrm{KL}\left[q\left(\boldsymbol{\delta}_d^{(0)}\middle|\boldsymbol{w}_d\right)\middle\|p\left(\boldsymbol{\delta}_d^{(T)}\right)\right] \quad (2)$$

図 1 に Flow-ETM の全体像を示す．推論ネットワークの出力は変分分布の平均$\boldsymbol{\mu}_d$と分散$\boldsymbol{\sigma}_d$であり，分岐直前の層の出力を$\boldsymbol{\pi}_d$としている．

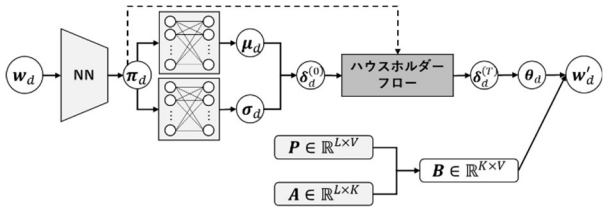


図 1　Flow-ETM の全体像

## 3. 実験と考察

ニュース記事データセットの 20 Newsgroups を用いて，テストデータに対する Perplexity を比較する評価実験を行う．本実験では全記事をランダムに分割することで，学習用は 11,214 件，検証用は 100 件，テスト用は 7,531 件とした．また，前処理後の単語数は 3,072，平均文書長は 127.2 であった．

はじめに，トピック数を$K = 10,20,30,40,50$，フローの長さを$T = 0,3,5,10,15$と変化させたときの Perplexity の値を図 2 に示す．ここで，$T = 0$は従来の ETM を意味する．図 2 より，$K = 10$の場合には$T = 5$が，それ以外のトピック数の場合には$T = 3$が最も良い性能を示しており，フローの導入によって Perplexity の値は改善したといえる．

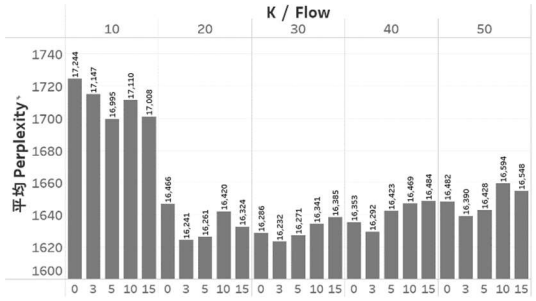


図 2　トピック数/フローの長さと Perplexity の関係

次に，Flow-ETM $(K = 20, T = 3)$から得られた，全テストデータに対する相関係数行列の平均を図 3 に示す．また，同一のモデルから得られた単語分布の出現確率上位 5 単語を，一部抜粋して表 1 に示す．図 2 よりトピック 1, 7, 10, 18 は互いに相関係数が低く，表 1 よりこれらのトピックの単語分布上位にはほぼ同一の単語が出現している．したがって，Flow-ETM は似たトピック間の相関を低く推論しており，一つの文書に関して，例えば“posting”や“writes”といった単語の背景に複数のトピックを仮定しづらい特性を持つことが分かる．

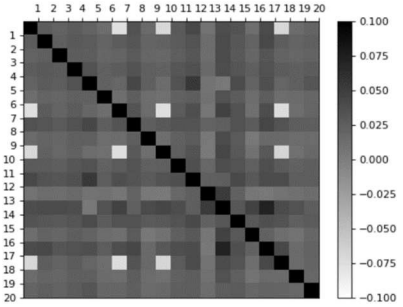

図 3　相関係数行列のヒートマップ

表 1　単語分布上位（一部抜粋）

| Topic 1 | Topic 7 | Topic 10 | Topic 18 |
|---|---|---|---|
| posting | writes | nntp | writes |
| host | posting | posting | posting |
| nntp | article | writes | host |
| writes | university | host | article |
| university | nntp | university | nntp |

## 4. 結論と今後の課題

本稿では，Flow-ETM に対する詳細な条件での実験により，文書分析における有効性を検証した．また相関係数行列の可視化により，Flow-ETM で推論される相関の特徴を確認した．今後の課題としては，より大規模なデータセットによる評価や，トピック相関係数行列に関するより詳しい考察が挙げられる．

**謝辞**

本研究の一部は，日本学術振興会（JSPS）科学研究費科研費 No.21H04600 の助成を受けたものです．

**参考文献**

[1] Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2019). Topic Modeling in Embedding Spaces. *arXiv:1907.04907*.

[2] Tomczak, J. M., and Welling, M. (2016). Improving variational auto-encoders using householder flow. *arXiv preprint arXiv:1611.09630*.

[3] 松苗亮汰, 山極綾子, 後藤正幸. (2021). Flow-ETM －トピック感の相関を表現した Embedded Topic Model. 日本計算機統計学会第 35 回シンポジウム, 学生研究発表セッション 2.