Vol.2019-NL-241 No.6
2019/8/29

トピック間の階層構造を考慮したGaussian LDAの構成

吉田 崇裕[1,a)]　久野 遼平[1,b)]　大西 立顕[1,c)]

概要：
トピックモデルは自然言語処理を始めとして多くの分野で用いられる手法である．トピックモデルの基本形である Latent Dirichlet Allocation(LDA) の提唱後，様々な LDA の改良モデルが提案されてきた．例えば Correlated Topic Model(CTM) は LDA が文書中のトピック間の相関を十分に考慮できない点に注目したモデルであり，汎化性能が向上すると報告されている．Gaussian LDA は LDA が単語間の意味的な近さを十分に考慮できない点に注目したモデルであり，トピックの意味一貫性が向上すると報告されている．両者を組み合わせた Correlated Gaussian Topic Model(CGTM) と呼ばれるモデルは上記二つの欠点を同時に補うのみならず，単語の埋め込み空間上でトピックの相関構造を可視化することができ革新的である．しかし，文書内におけるトピックの関係性は，CGTM が対象とする単純な相関構造だけで表現できるものではない．実際日常生活においても，例えば「経済」-「金融政策」-「出口戦略」のように話題の階層性を意識し会話をすることは多々ある．そこで本稿では階層的トピックモデルとして最も単純な PAM(Pachinko Allocation Model) と Gaussian LDA を組み合わせたモデルを提案することで，トピックの階層構造を単語埋め込みベクトル空間上で分析する一歩としたい．

1. 序論

トピックモデルは自然言語処理を始めとして多くの分野で用いられる手法である [11]．トピックモデルは自然言語処理の中では文書中の単語の出現頻度を表した Bag of Words 行列 [1] に適用され，単語を生成するトピックと各文書におけるトピックの構成確率という潜在構造をデータから推定する際に使用される．トピックモデルの基本形とも呼ばれる Latent Dirichlet Allocation(LDA)[2] はそれまでの Latent Semantic Indexing(LSI)[7] や probabilistic Latent Semantic Indexing(pLSI)[9] に比べ，ベイズモデル化し事前分布にディリクレ分布を置いたことで，解釈性の高い疎なトピック分布を推定できるようになると同時に，未知の単語と文書に対しても容易に確率を計算できるようになったことでモデルの利便性が格段に向上した．かかるメリットと Bag of Words 行列に代表される非負行列の遍在性もあいまって，LDA は自然言語処理の世界のみならず幅広く用いられるようになった．

もっとも，先述の通り LDA は文書トピック分布の事前分布にディリクレ分布を仮定しているところ，ディリクレ分布にしたがう確率分布は各成分ごとに相関を持っていないため，文書内におけるトピック間の相関を十分に考慮することができないという欠点を有している．かかる欠点に対しては，CTM(Correlated Topic Model)[4] という手法が提案されており，かかる手法は，文書トピック分布の事前分布としてディリクレ分布でなく多変量正規分布（を正規化したもの）を用いることで，各成分ごとの相関を表現している．そして，CTM が元の LDA と比べ，テストデータに含まれる各単語の尤度の幾何平均が大きくなる（すなわち Perplexity が低くなる）との報告が [4] によってなされており，CTM の汎化性能の高さを示唆している．

他にも，LDA は文書の単語の出現頻度のみに注目しているモデルであることから，単語間の意味的な近さを十分に考慮することができていないという欠点も有している．かかる欠点に対しては，Gaussian LDA[6], [10] と呼ばれる手法が提案されており，これは，word2vec[12] などで事前に学習された単語の分散表現を用い，単語を連続空間に埋め込まれたベクトルとして考えるという手法である．これにより，単語間の意味的な近さを事前情報として数値的に考慮することができ，同じトピックに割当てられた単語集合の意味一貫性が LDA に比べ向上する（すなわち Coherence が高くなる）との報告が [6] によってなされている．

そして，これらの欠点を共に改善すべく CTM と Gaussian LDA を組み合わせた CGTM(Correlated Gaussian Topic

[1] 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
[a)] takahiro_yoshida@mist.i.u-tokyo.ac.jp
[b)] hisano.ryohei@sict.i.u-tokyo.ac.jp
[c)] ohnishi.takaaki@i.u-tokyo.ac.jp

Model)[17] という手法も提案されている．CGTM のメリットは文書分類の精度向上やトピック内の単語の意味的一貫性の向上のみならず，単語の埋め込み空間上でトピックの相関構造を可視化でき（図 1 参照），単語間の意味的な近さを考慮したトピック間の関係を考察できるようになっている点が革新的であるといえる．

しかし，トピックの関係性は単純な相関構造だけで表現できるものではない．例えば LDA の拡張モデルの一つである PAM(Pachinko Allocation Model)[16] は，トピックの上位概念であるスーパートピックを導入し，上位概念であるスーパートピックと下位概念であるサブトピックのように，トピック間に階層性を導入することでトピック同士の関係性を考慮している．他にも，hLDA(hierarchical LDA)[3] や hPAM(hierarchical PAM)[13] なども，トピックの階層構造をモデリングに組み入れている．ここで PAM は，あくまで下位の層に属するサブトピックからしか単語が生起しないと考える一方で，hLDA や hPAM は，下位層に限らず上位の層に属するトピックからも単語が生起する可能性があることを前提にモデリングされており，これによって単語自体の階層性も考慮することが可能になっている．このように，トピック間の関係を考慮するにあたっては，CTM のような横並びの相関関係として表現するやり方のみならず，トピックを上下に配置して階層構造として考察するやり方も十分に考えられるところである．さらに，実際に我々が言葉を使うときにも，例えば「経済」-「アベノミクス」,「スポーツ」-「野球」等のように，トピックを包括的なトピックとその包括的なトピックの一部である個別具体的なトピックに分類することができると考えられ，トピックの考察において階層構造を入れることは，むしろ自然言語のモデリングとして自然であるとも考えうる．そのため，トピックの階層構造を考慮した上で単語埋め込み空間上でトピック間の関係を可視化することができれば，より自然かつコンパクトにトピックの関係性を捉えられることが期待される．

そこで本稿では，トピックの階層性を考慮したモデルの中でも最も単純な PAM と Gaussian LDA を組み合わせたモデリング・推論手法を提案することで，トピックの階層構造を単語埋め込み空間上で可視化する一歩としたい．本稿ではモデリング・推論手法の提案にとどめるが，発表時には数値実験並びに hLDA や hPAM との違いの考察に関しても紹介することを予定している．

2. 確率分布の準備

定義 2.1 (多項分布)．パラメータ $\pi = (\pi_1, \pi_2, \cdots \pi_n), (\sum_{i=1}^{n} \pi_i = 1)$，確率変数 $x \in \{1, 2, \cdots n\}$ に対し，

$$p(x = i|\pi) = \pi_i \ (i = 1, 2, \cdots, n) \tag{1}$$

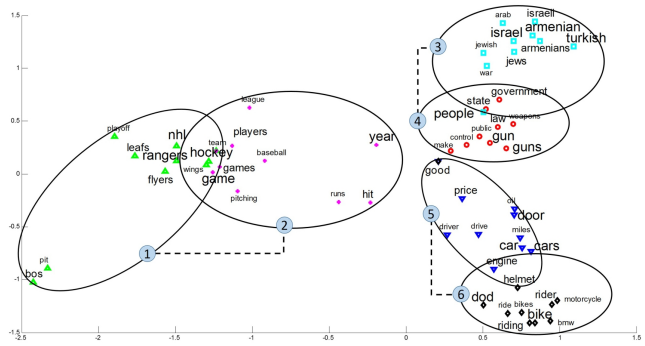

図 1 CGTM における単語埋め込み空間上でのトピックの相関関係（[17] より引用）

で定義される確率分布のことを多項分布という．x がパラメータ π の多項分布にしたがって生成される時，$x \sim \mathrm{Multi}(\pi)$ と書く．

定義 2.2 (ディリクレ分布)．パラメータ $\alpha = (\alpha_1, \alpha_2, \cdots \alpha_n)\,(\alpha_i > 0)$，確率変数 $\pi = (\pi_1, \pi_2, \cdots \pi_n)$（ただし各成分の和は 1）に対し，

$$p(\pi|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{n} \alpha_i\right)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} \pi_i^{\alpha_i - 1} \tag{2}$$

で定義される確率分布のことをディリクレ分布という．ただし，$\Gamma(x)$ はガンマ関数を表す．π がパラメータ α のディリクレ分布にしたがって生成される時，$\pi \sim \mathrm{Dir}(\alpha)$ と書く．

また，ディリクレ分布は多項分布の共役事前分布である．すなわち，

$$\pi \sim \mathrm{Dir}(\alpha), \tag{3}$$

$$x_i \sim \mathrm{Multi}(\pi)\ (i = 1, 2, \cdots, N) \tag{4}$$

とデータが生成されたとすると，π の事後確率 $p(\pi|x)$ はディリクレ分布（パラメータの k 成分は $\alpha_k + n_k$）になる（ただし，n_k は $x_i = k$ となる i の個数を表す）．

定義 2.3 (M 次元正規分布)．パラメータ $\mu \in \mathbb{R}^M, \Sigma \in \mathbb{R}^{M \times M}$（$\Sigma$ は正定値対称行列），確率変数 $x \in \mathbb{R}^M$ に対し，確率密度関数が

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{M}{2}} \sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2}(x - \mu)^{\mathsf{T}} \Sigma^{-1} (x - \mu)\right\} \tag{5}$$

で定義される確率分布のことを M 次元正規分布という．x がパラメータ μ, Σ の正規分布にしたがって生成されるとき，$x \sim \mathcal{N}(\mu, \Sigma)$ と書く．

定義 2.4 (逆ウィシャート分布・正規逆ウィシャート分布)．パラメータ $\Psi \in \mathbb{R}^{M \times M}, v \in \mathbb{R}$，確率変数 $\Sigma \in \mathbb{R}^{M \times M}$（$\Psi, \Sigma$ は正定値対称，$v > M - 1$）に対し，確率密度関数が

$$p(\Sigma|\Psi,v)=\frac{1}{Z_{IW}}|\Sigma|^{-(v+M+1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\Psi\Sigma^{-1}\right)\right\}, \tag{6}$$

$$Z_{IW}=|\Psi|^{-v/2}2^{vM/2}\Gamma_M(v/2), \tag{7}$$

$$\Gamma_M(v/2)=\prod_{i=1}^{M}\Gamma(\frac{v+1-i}{2}) \tag{8}$$

で定義される確率分布のことを逆ウィシャート分布という．Σ がパラメータ Ψ, v の逆ウィシャート分布にしたがって生成されるとき，$\Sigma\sim\mathcal{IW}(\Psi,v)$ と書く．

また，(μ,Σ) に関する分布が，

$$p(\mu,\Sigma)=p(\mu|\Sigma)p(\Sigma), \tag{9}$$

$$p(\Sigma)=\mathcal{IW}(\Psi,v), \tag{10}$$

$$p(\mu|\Sigma)=\mathcal{N}(u,\frac{1}{\kappa}\Sigma) \tag{11}$$

のように表されるとき，この分布のことをパラメータ (u,κ,Ψ,v) の正規逆ウィシャート分布といい，$\mathcal{NIW}(u,\kappa,\Psi,v)$ と書く．

ここで，逆ウィシャート分布は，正規分布の平均・分散共分散行列に対する共役事前分布であるという性質を有する．すなわち，μ,Σ が上記のような正規逆ウィシャート分布を事前分布として持ち，

$$x_i\sim\mathcal{N}(\mu,\Sigma)\ (i=1,2,\cdots,N) \tag{12}$$

とデータが生成されたとすると，μ,Σ の事後確率 $p(\mu,\Sigma|x)$ も正規逆ウィシャート分布になる（パラメータは x から定まる）．

3. 先行研究のモデリング手法

3.1 LDA のモデリング

LDA[2] は文書の生成過程を確率的に記述したモデルの一つである．具体的には，各文書にはトピック割合が潜在的に定まっていると仮定し，そのトピック割合を確率分布を用いてモデリングする．

以下, 具体的な定式化を行う．まず, D を文書の総数とし，文書 d の総単語数を N_d とおく．そして，トピックの総数を K とし，これは固定された数であるとする．単語の語彙の総数を V とし，各単語は $1,2,\cdots,V$ と数字が割り振られているとする．$\theta_{d,k}$ を文書 d においてトピック k が現れる確率とし，文書 d のトピック分布を $\theta_d=(\theta_{d,1},\theta_{d,2},\cdots,\theta_{d,K})$ とおく．また，$\phi_{k,v}$ をトピック k から単語 v が生成される確率とし，単語出現分布として $\phi_k=(\phi_{k,1},\phi_{k,2},\cdots,\phi_{k,V})$ を定義する．さらに，文書 d の n 番目の単語トークンに割りあてられたトピックを $z_{d,n}$，具体的な単語を $w_{d,n}$ で表す．

このとき，α を K 次元ベクトルのパラメータ，β を V 次元ベクトルのパラメータとして，

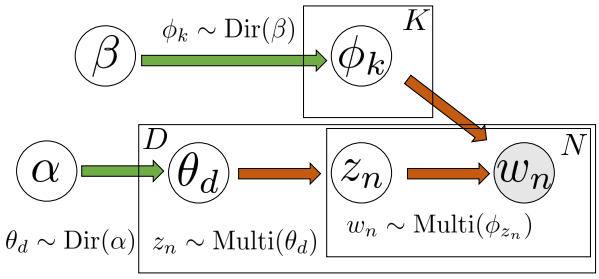

図 2　LDA のグラフィカルモデル

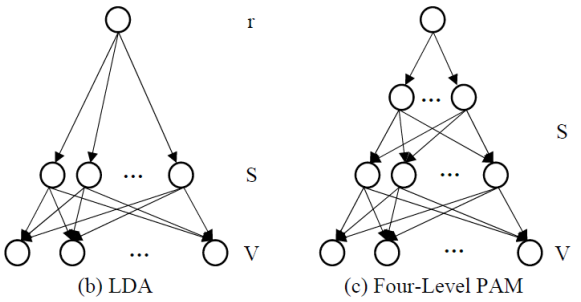

図 3　LDA と PAM の比較（[16] より引用）

$$\theta_d\sim\mathrm{Dir}(\alpha)\ (d=1,2,\cdots,D), \tag{13}$$

$$\phi_k\sim\mathrm{Dir}(\beta)\ (k=1,2,\cdots,K) \tag{14}$$

のようにディリクレ分布から θ や ϕ を定め，各文書 d に対して，

$$z_{d,n}\sim\mathrm{Multi}(\theta_d)\ (n=1,2,\cdots,N_d), \tag{15}$$

$$w_{d,n}\sim\mathrm{Multi}(\phi_{z_{d,n}}) \tag{16}$$

として文書生成する確率モデルが LDA(Latent Dirichlet Allocation，潜在的ディリクレ配分) である．LDA のグラフィカルモデルを図 2 に示す．

3.2 PAM のモデリング

Pachinko Allocation Model(PAM)[16] は，トピックの上位概念としてスーパートピックを導入し，スーパートピックからの各サブトピックの生起確率の大小によってトピック間の相関を考慮する．LDA と PAM の違いを表した図を図 3 に示す．

PAM においては，各文書ごとにスーパートピック分布と，各スーパートピックにおけるサブトピック分布を定め，それにしたがって各単語トークンのトピックや具体的な単語を定めていく．このような，スーパートピック・サブトピックを経由して単語トークンに割りあてられる語彙が定まるという階層性が，“Pachinko”と名づけられたゆえんである．

以下 PAM の定式化を行う．スーパートピックの数を S とする．また，3.1 節で定義した変数のほかに，$\hat{\theta}_d$ を文書

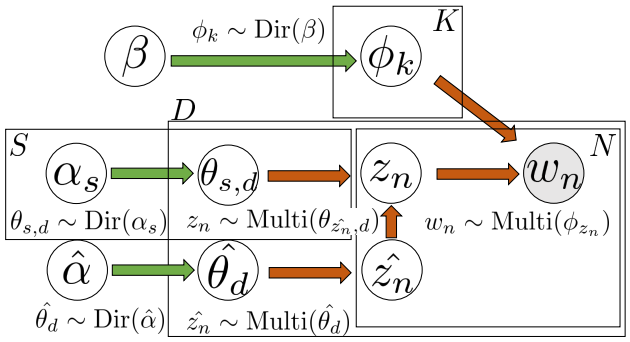

図 4　PAM のグラフィカルモデル

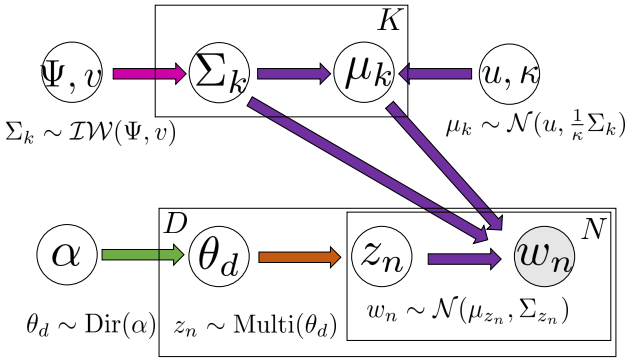

図 5　Gaussian LDA のグラフィカルモデル

d のスーパートピック分布を表す確率変数と定め，$\theta_{s,d}$ を文書 d のスーパートピック s の元でのサブトピック分布を表す確率変数とする．そして，$\hat{z}_{d,n}$ を文書 d の n 番目の単語トークンのスーパートピックを表す変数とし，$z_{d,n}$ を文書 d の n 番目の単語トークンのサブトピックを表す変数とする．$\hat{\alpha}$ を S 次元のパラメータ，$\alpha_s (s = 1, 2, \cdots, S)$ を K 次元のパラメータとするとき，

$$\hat{\theta}_d \sim \mathrm{Dir}(\hat{\alpha}) \ (d = 1, 2, \cdots, D), \tag{17}$$

$$\theta_{s,d} \sim \mathrm{Dir}(\alpha_s) \ (d = 1, 2, \cdots, D, \ s = 1, 2, \cdots, S), \tag{18}$$

$$\phi_k \sim \mathrm{Dir}(\beta) \ (k = 1, 2, \cdots, K) \tag{19}$$

のようにディリクレ分布から $\hat{\theta}$，θ や ϕ を定め，各文書 d に対して，

$$\hat{z}_{d,n} \sim \mathrm{Multi}(\hat{\theta}_d) \ (n = 1, 2, \cdots, N_d), \tag{20}$$

$$z_{d,n} \sim \mathrm{Multi}(\theta_{\hat{z}_{d,n},d}), \tag{21}$$

$$w_{d,n} \sim \mathrm{Multi}(\phi_{z_{d,n}}) \tag{22}$$

とする．これが PAM のモデリングである．PAM のグラフィカルモデルを図 4 に示す．

3.3　Gaussian LDA のモデリング

Gaussian LDA[6], [10] も LDA と同様文書の生成過程を確率的に記述したモデルの一つであるが，単語の分散表現を用い，各単語をベクトルとして扱う点が LDA と異なる．

以下，具体的な定式化を行う．まず，$D, N_d, K, \theta_d, z_{d,n}$ については 3.1 節と同様に定義する．また，文書 d における n 番目の単語トークンに割りあてられた具体的な単語（これは分散表現によりベクトルで表示されたものを考える．）を $w_{d,n}$ で表す．

このとき，α を K 次元ベクトルのパラメータ，u, κ, Ψ, v をそれぞれ，M 次元ベクトル，スカラー値，$M \times M$ 行列，スカラー値のパラメータとして，

$$\theta_d \sim \mathrm{Dir}(\alpha) \ (d = 1, 2, \cdots, D), \tag{23}$$

$$\mu_k, \Sigma_k \sim \mathcal{NIW}(u, \kappa, \Psi, v) \ (k = 1, 2, \cdots, K) \tag{24}$$

のようにディリクレ分布から θ を定め，正規逆ウィシャート分布から μ_k, Σ_k を定める．そして，各文書 d に対して，

$$z_{d,n} \sim \mathrm{Multi}(\theta_d) \ (n = 1, 2, \cdots, N_d), \tag{25}$$

$$w_{d,n} \sim \mathcal{N}(\mu_{z_{d,n}}, \Sigma_{z_{d,n}}) \tag{26}$$

として文書を生成する確率モデルが Gaussian LDA(Gaussian Latent Dirichlet Allocation) である．

Gaussian LDA のグラフィカルモデルを図 5 に示す．

4.　GLDA の拡張

4.1　概要

LDA や Gaussian LDA では文書のトピック分布の事前分布をディリクレ分布としている点に大きな特徴があった．もっともディリクレ分布に従う確率変数は各成分間で相関がなく，それは文書のトピックの決まり方としてトピック間に相関がないということを暗に仮定してしまっている．しかし，実際には，政治というトピックと経済というトピックは同じ一つの文書で同時に出てきやすいといったトピック間の相関は存在すると考えられる．したがって，トピック間の相関を考慮したモデリングを行うことができれば，元の LDA や Gaussian LDA よりもよい予測精度を示すことが期待される．

そして，LDA においては，トピック間の相関を考慮できるように LDA を拡張したモデルとして，Correlated Topic Model(CTM)[4] や Pachinko Allocation Model(PAM)[16] などが提案されており，特に Gaussian LDA と CTM を組み合わせた手法（CGTM）がすでに提案されている [17]．もっとも，[16] によって，PAM の方が CTM よりも尤度が高くなることが報告されており，より現実に即したモデリングになっていると考えられる点，[5] によると，CTM は，生成しうる単語の選択肢の数を表す Perplexity（定義は式 (27)）という指標の下では高い性能を発揮しているが，人間の解釈する単語同士の意味関連性を数値化した Coherence という指標の下では高い性能を発揮しているとは言えず，CTM は人間にとって可読性の高い (辻褄の合った) トピック分布を返すとは必ずしも言えないと考えられ

る点，そして 1 章でも述べたとおり，PAM は CTM にはないトピック間の階層性を考慮できる点などを踏まえ，本研究ではトピック間の階層性を表現することができる PAM を Gaussian LDA に適用することを考えた．

なお，テスト文書 D_{test} に対する perplexity は以下のように定義される [2]．

$$\text{perplexity}(D_{\text{test}}) = \exp\left(-\frac{\sum_{d\in D_{\text{test}}}\sum_n \log p(w_{d,n})}{\sum_d N_d}\right). \quad (27)$$

これは各単語の生起確率 $p(w_{d,n})$ の逆数の幾何平均を表している．確率の逆数を選択肢の数と観念できるところ，perplexity は「モデリングをすることで選択肢の数をどこまで減らせるか」を表す指標であり，トピックモデルの汎化能力を表していると考えることができる．

また，Coherence は，トピックの分類が人間にとって可読性高くなされているかを示す指標であり，[5] は人間の被験者を集めトピックの可読性を数値化する手法を提案している．さらに，[15] は，Wikipedia コーパスにおける単語の共起回数を用いる PMI(Pointwise mutual information) という指標を提案している．

4.2 提案手法

本節では，Gaussian LDA と PAM を組み合わせた提案手法について述べる．各変数を 3.1 節，3.2 節，3.3 節で定めたものと同様に設定する．このとき，

$$\hat{\theta}_d \sim \text{Dir}(\hat{\alpha})\ (d = 1, 2, \cdots, D), \quad (28)$$

$$\theta_{s,d} \sim \text{Dir}(\alpha_s)\ (d = 1, 2, \cdots, D,\ s = 1, 2, \cdots, S), \quad (29)$$

$$\mu_k, \Sigma_k \sim \mathcal{NIW}(u, \kappa, \Psi, v)\ (k = 1, 2, \cdots, K) \quad (30)$$

のようにディリクレ分布から $\hat{\theta}$，θ を定め，正規逆ウィシャート分布に従って μ_k, Σ_k を定める．そして，各文書 d に対して，

$$\hat{z}_{d,n} \sim \text{Multi}(\hat{\theta}_d)\ (n = 1, 2, \cdots, N_d), \quad (31)$$

$$z_{d,n} \sim \text{Multi}(\theta_{\hat{z}_{d,n},d}), \quad (32)$$

$$w_{d,n} \sim \mathcal{N}(\mu_{z_{d,n}}, \Sigma_{z_{d,n}}) \quad (33)$$

とする．これが提案手法のモデリングである．提案手法のグラフィカルモデルを図 6 に示す．

4.3 提案手法の推論方法

本節では周辺化ギブスサンプリングに基づく，各単語トークンのスーパートピック分布・サブトピック分布を推定する方法について述べる．

そもそも，この提案手法においては，$\hat{z}_{d,n}, z_{d,n}, \theta_d, \mu_k, \Sigma_k$ の事後確率を求めることにより，未知の文書に対するトピック分類などが可能になるが，事後確率 $p(\hat{z}, z, \theta, \mu, \Sigma|w)$ を解析的に求めることは難しいという問題点がある．そこで，LDA や Gaussian LDA での推論手法として主によく用いられるものとして，変分ベイズ法 [2], [10] と周辺化ギブスサンプリング [6], [8] の二つがある．変分ベイズ法は，計算が容易な分布で事後分布を近似する手法であるが，分布の形状を指定した上で（具体的には因子分解可能性を課して）近似するため，精度が下がりやすいという欠点がある．そのため今回は，変分ベイズ法よりは計算量が多いものの精度よく推論できるとされている，周辺化ギブスサンプリングを用いて事後確率の近似分布を求めることとした．

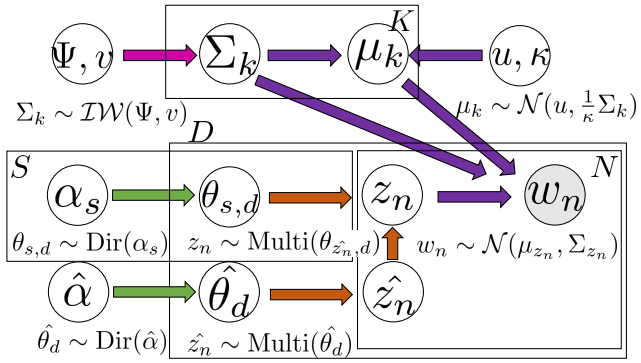

図 6 提案手法のグラフィカルモデル

サンプリングによる近似手法では，$p(\hat{z}, z, \theta, \mu, \Sigma|w)$ を具体的に求めることを諦め，$p(\hat{z}, z, \theta, \mu, \Sigma|w)$ から生成した T 個のサンプル点 $\hat{z}^{(t)}, z^{(t)}, \theta^{(t)}, \mu^{(t)}, \Sigma^{(t)}$ によって，事後確率の近似分布を求める．もっとも，サンプリングの際，変数が多い同時確率分布からそれぞれの変数を同時にサンプリングするのは難しい．そのため，ギブスサンプリングという手法では，1 つの変数を除いて他の変数を全て固定し，その条件付確率分布から 1 つずつ変数をサンプリングしていくという方針をとる．

また，周辺化ギブスサンプリングは θ, μ, Σ で周辺化した $p(\hat{z}, z|w) = \int p(\hat{z}, z, \theta, \mu, \Sigma|w) d\theta d\mu d\Sigma$ から，$\hat{z}_{d,n}, z_{d,n}$ を一つ一つサンプリングしていく手法である．周辺化することでサンプリングすべき変数が少なくなる分，より高速でサンプリングすることが可能になると共に，$\hat{z}_{d,n}, z_{d,n}$ がサンプリングできていれば，残りの θ, μ, Σ は簡単に近似できるため，事後分布を近似するという目的は果たされることになる．以下，$w_{-(d,n)}$ などの表記は，全ての文書における w のうち，$w_{d,n}$ のみを取り除いたものを表すとする．

以降，周辺化ギブスサンプリングを行うための確率分布 $p(\hat{z}_{d,n} = s, z_{d,n} = k|w_{d,n}, \hat{z}_{-(d,n)}, z_{-(d,n)}, w_{-(d,n)})$ の計算式を導出する．まずベイズの定理を用い，グラフィカルモデルに基づいて展開した上で，定数部分 $p(\hat{z}_{-(d,n)}, z_{-(d,n)}, w_{-(d,n)})$ を無視すると，

$$
\begin{aligned}
&p(\hat{z}_{d,n}=s, z_{d,n}=k|w_{d,n}, \hat{z}_{-(d,n)}, z_{-(d,n)}, w_{-(d,n)})\\
&\propto p(w_{d,n}|z_{d,n}=k, \hat{z}_{-(d,n)}, z_{-(d,n)}, w_{-(d,n)})\\
&\quad\times p(z_{d,n}=k|\hat{z}_{d,n}=s, \hat{z}_{-(d,n)}, z_{-(d,n)}, w_{-(d,n)})\\
&\quad\times p(\hat{z}_{d,n}=s|\hat{z}_{-(d,n)}, z_{-(d,n)}, w_{-(d,n)})
\end{aligned}
\tag{34}
$$

を得る．式 (34) の 2 つ目について，

$$
\begin{aligned}
&p(z_{d,n}=k|\hat{z}_{d,n}=s, \hat{z}_{-(d,n)}, z_{-(d,n)}, w_{-(d,n)})\\
&=\int p(z_{d,n}=k|\hat{z}_{d,n}=s, \theta_{s,d})p(\theta_{s,d}|\hat{z}_{-(d,n)}, z_{-(d,n)}, \alpha_s)d\theta_{s,d}\\
&=E_{p(\theta_{s,d}|\hat{z}_{-(d,n)}, z_{-(d,n)}, \alpha_s)}[(\theta_{s,d})_k]
\end{aligned}
\tag{35}
$$

と変形でき，式 (34) の 3 つ目について，

$$
\begin{aligned}
&p(\hat{z}_{d,n}=s|\hat{z}_{-(d,n)}, z_{-(d,n)}, w_{-(d,n)})\\
&=\int p(\hat{z}_{d,n}=s|\hat{\theta}_d)p(\hat{\theta}_d|\hat{z}_{-(d,n)}, z_{-(d,n)}, \hat{\alpha})d\hat{\theta}_d\\
&=E_{p(\hat{\theta}_d|\hat{z}_{-(d,n)}, z_{-(d,n)}, \hat{\alpha})}[(\hat{\theta}_d)_s]
\end{aligned}
\tag{36}
$$

と変形できる．ここで，ディリクレ分布が多項分布の共役事前分布であるという性質を用いると，$p(\theta_{s,d}|\hat{z}_{-(d,n)}, z_{-(d,n)}, \alpha_s)$ と $p(\hat{\theta}_d|\hat{z}_{-(d,n)}, z_{-(d,n)}, \hat{\alpha})$ はそれぞれディリクレ分布になる．また，パラメータについても 2 章で述べたことを踏まえれば簡単に計算でき，

$$
p(\theta_{s,d}|\hat{z}_{-(d,n)}, z_{-(d,n)}, \alpha_s)=\mathrm{Dir}(\alpha_s+N_{ds(-n)}),
\tag{37}
$$

$$
p(\hat{\theta}_d|\hat{z}_{-(d,n)}, z_{-(d,n)}, \hat{\alpha})=\mathrm{Dir}(\hat{\alpha}+\hat{N}_{d(-n)})
\tag{38}
$$

であることが分かる．ただし，$N_{ds(-n)}$ の第 k 成分 $(N_{ds(-n)})_k$ と，$\hat{N}_{d(-n)}$ の第 s 成分 $(\hat{N}_{d(-n)})_s$ はそれぞれ，

$$
(N_{ds(-n)})_k=\#\{i|\hat{z}_{d,i}=s\wedge z_{d,i}=k\ (1\le i\le N, i\neq n)\},
\tag{39}
$$

$$
(\hat{N}_{d(-n)})_s=\#\{i|\hat{z}_{d,i}=s\ (1\le i\le N, i\neq n)\}
\tag{40}
$$

を表す．また，一般に，x がパラメータ $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_n)$ のディリクレ分布 $\mathrm{Dir}(\alpha)$ にしたがっているとしたとき，

$$
E[x]=\frac{\alpha_i}{\sum_j\alpha_j}
\tag{41}
$$

となることが知られている．式 (41) を式 (35)，(36) に代入することで，結局，

$$
\begin{aligned}
&p(z_{d,n}=k|\hat{z}_{d,n}=s, \hat{z}_{-(d,n)}, z_{-(d,n)}, w_{-(d,n)})\\
&=\frac{(\alpha_s)_k+(N_{ds(-n)})_k}{\sum_{k'}((\alpha_s)_{k'}+(N_{ds(-n)})_{k'})},
\end{aligned}
\tag{42}
$$

$$
\begin{aligned}
&p(\hat{z}_{d,n}=s|\hat{z}_{-(d,n)}, z_{-(d,n)}, w_{-(d,n)})\\
&=\frac{\hat{\alpha}_s+(\hat{N}_{d(-n)})_s}{\sum_{s'}((\hat{\alpha}_{s'}+(\hat{N}_{d(-n)})_{s'})}
\end{aligned}
\tag{43}
$$

を得る．

また，式 (34) の 1 つ目については，

$$
\begin{aligned}
&p(w_{d,n}|z_{d,n}=k, \hat{z}_{-(d,n)}, z_{-(d,n)}, w_{-(d,n)})\\
&=\int p(w_{d,n}|\mu_k,\Sigma_k)p(\mu_k,\Sigma_k|z_{-(d,n)}, w_{-(d,n)})d\mu_k d\Sigma_k
\end{aligned}
\tag{44}
$$

と変形することができる．

ここで，正規逆ウィシャート分布の共役性を用いると，$p(\mu_k,\Sigma_k|z_{-(d,n)}, w_{-(d,n)})$ は，パラメータ $(m_k,\kappa_k,\Psi_k,v_k)$ の正規逆ウィシャート分布になる．パラメータの各値は以下の通りである．

$$
m_k=\frac{\kappa u+N_k\bar{x}_k}{\kappa_k},
\tag{45}
$$

$$
\kappa_k=\kappa+N_k,
\tag{46}
$$

$$
C_k=\sum_{(d',n')\in\mathrm{DN}_{-(d,n)}}\delta(z_{d',n'}=k)(w_{d',n'}-\bar{x}_k)(w_{d',n'}-\bar{x}_k)^{\mathsf{T}},
\tag{47}
$$

$$
\Psi_k=\Psi+\frac{\kappa N_k}{\kappa_k}(\bar{x}_k-u)(\bar{x}_k-u)^{\mathsf{T}}+C_k,
\tag{48}
$$

$$
v_k=v+N_k.
\tag{49}
$$

ただし，$N_k,\bar{x}_k,\mathrm{DN}_{-(d,n)}$ は以下の通り定義されるものである．

$$
N_k=\sum_{(d',n')\in\mathrm{DN}_{-(d,n)}}\delta(z_{d',n'}=k),
\tag{50}
$$

$$
\bar{x}_k=\frac{1}{N_k}\sum_{(d',n')\in\mathrm{DN}_{-(d,n)}}\delta(z_{d',n'}=k)w_{d',n'},
\tag{51}
$$

$$
\mathrm{DN}_{-(d,n)}=\{(d',n')|1\le d'\le D, 1\le n'\le N_{d'}\}\backslash(d,n).
\tag{52}
$$

これを式 (44) に代入すると，式 (44) は，

$$
\int\mathcal{N}(w_{d,n}|\mu_k,\Sigma_k)\mathcal{NIW}(\mu_k,\Sigma_k|m_k,\kappa_k,\Psi_k,v_k)d\mu_k d\Sigma_k
\tag{53}
$$

と変形できる．そして，[14] によると，式 (53) の形の積分は解析的に求めることができ，式 (53) は以下のように M 次元スチューデントの t 分布の確率密度関数に $w_{d,n}$ を代入したものという形で書くことができる．

$$
\begin{aligned}
&\int\mathcal{N}(w_{d,n}|\mu_k,\Sigma_k)\mathcal{NIW}(\mu_k,\Sigma_k|m_k,\kappa_k,\Psi_k,v_k)d\mu_k d\Sigma_k\\
&=\mathcal{T}_{v_k-M+1}(w_{d,n}|m_k,\frac{\kappa_k+1}{\kappa_k(v_k-M+1)}\Psi_k).
\end{aligned}
\tag{54}
$$

したがって，

$$
\begin{aligned}
&p(w_{d,n}|z_{d,n}=k, \hat{z}_{-(d,n)}, z_{-(d,n)}, w_{-(d,n)})\\
&=\mathcal{T}_{v_k-M+1}(w_{d,n}|m_k,\frac{\kappa_k+1}{\kappa_k(v_k-M+1)}\Psi_k)
\end{aligned}
\tag{55}
$$

を得る．

最後に，式 (42),(43),(55) を式 (34) に代入して，

$$p(\hat{z}_{d,n}=s, z_{d,n}=k|w_{d,n}, \hat{z}_{-(d,n)}, z_{-(d,n)}, w_{-(d,n)}) \propto \frac{\hat{\alpha}_s + (\hat{N}_{d(-n)})_s}{\sum_{s'}((\hat{\alpha}_{s'} + (\hat{N}_{d(-n)})_{s'})} \cdot \frac{(\alpha_s)_k + (N_{ds(-n)})_k}{\sum_{k'}((\alpha_s)_{k'} + (N_{ds(-n)})_{k'})} \cdot \mathcal{T}_{v_k-M+1}(w_{d,n}|m_k, \frac{\kappa_k+1}{\kappa_k(v_k-M+1)}\Psi_k) \quad (56)$$

を得る．$\hat{z}_{d,n}, z_{d,n}$ はこの値に従いサンプリングをし，d, n を変えて次々と $\hat{z}_{d,n}, z_{d,n}$ をサンプリングしていくことで，結果 $\hat{z}, z$ 全体についてサンプリングをすることが可能になる．

4.4 ハイパーパラメータ α_s の更新式

本節ではハイパーパラメータの定め方について述べる．トピックモデルにおいては，ハイパーパラメータも推定する場合もあるが，最初から固定して推論する場合も少なくなく，実際ハイパーパラメータを固定してもよい結果が得られることが知られている．しかし，提案手法中のハイパーパラメータ α_s については，スーパートピックとサブトピックの関係性を定めているから，トピック間の相関を考慮するうえで極めて重要なパラメータといえる．そのため，本稿ではハイパーパラメータ α_s について，[16] でも用いられているモーメントマッチングという手法によって更新することとする．

モーメントマッチングとは，実際のデータのモーメントとパラメータから求めたモーメントが一致するようにパラメータを調整することでパラメータを更新する手法である．これを提案手法に適用すると，以下のような流れで更新式が求まる．すなわち，提案手法においては，パラメータ $\alpha_s = (\alpha_{s,1}, \alpha_{s,2}, \cdots, \alpha_{s,k}, \cdots, \alpha_{s,K})\,(\alpha_{s,k} > 0)$，確率変数 $\theta_{s,d} = (\theta_{s,d,1}, \theta_{s,d,2}, \cdots, \theta_{s,d,k}, \cdots, \theta_{s,d,K})$ に対し，$\theta_{s,d} \sim \mathrm{Dir}(\alpha_s)$ である．そして，$\theta_{s,d,k}$ の 1 次モーメントと 2 次モーメントは $\alpha_{s,0} = \sum_j \alpha_{s,j}$ を用いて以下のように書けることが知られている．

$$E[\theta_{s,d,k}] = \frac{\alpha_{s,k}}{\alpha_{s,0}}, \quad (57)$$

$$E[\theta^2_{s,d,k}] = \frac{\alpha_{s,k}(1+\alpha_{s,k})}{\alpha_{s,0}(1+\alpha_{s,0})}. \quad (58)$$

これは d に関わらず成り立つから，d で平均を取って

$$\frac{1}{D}\sum_d E\left[\theta_{s,d,k}\right] = \frac{\alpha_{s,k}}{\alpha_{s,0}}, \quad (59)$$

$$\frac{1}{D}\sum_d E\left[\theta^2_{s,d,k}\right] = \frac{\alpha_{s,k}(1+\alpha_{s,k})}{\alpha_{s,0}(1+\alpha_{s,0})} \quad (60)$$

を得る．そして，$\theta_{s,d,k}$ は「文書 d 中のスーパートピック s に属する単語トークンがサブトピック k に属する確率」を表しているところ，これの 1 次モーメントと 2 次モーメントは，文書全体のデータからは以下のように計算できる．

$$\mathrm{first}_{skd} = \frac{n^{(d)}_{sk}}{n^{(d)}_s}, \quad (61)$$

$$\mathrm{second}_{skd} = \left(\frac{n^{(d)}_{sk}}{n^{(d)}_s}\right)^2. \quad (62)$$

ただし，$n^{(d)}_s, n^{(d)}_{sk}$ は，それぞれ，文書 d 中のスーパートピック s に属する単語トークン数と，文書 d 中のスーパートピック s に属しサブトピック k に属する単語トークン数を表す．そしてこれについて文書間で平均を取ると以下を得る．

$$\mathrm{first}_{sk} = \frac{1}{D}\sum_d \frac{n^{(d)}_{sk}}{n^{(d)}_s}, \quad (63)$$

$$\mathrm{second}_{sk} = \frac{1}{D}\sum_d \left(\frac{n^{(d)}_{sk}}{n^{(d)}_s}\right)^2. \quad (64)$$

そして，式 (59) と (63)，式 (60) と (64) が一致するように α_s を定めればよい．任意の k について成り立つ関係式

$$\alpha_{s,0} = \frac{E[\theta_{s,d,k}] - E[\theta^2_{s,d,k}]}{E[\theta^2_{s,d,k}] - E[\theta_{s,d,k}]^2} \quad (65)$$

を用いると，

$$\alpha_{s,k} \propto \mathrm{first}_{sk}, \quad (66)$$

$$\alpha_{s,0} = \frac{1}{K}\sum_k m_{sk} \quad (67)$$

とすれば式 (59) と (63)，式 (60) と (64) が一致することがわかる．ただし，

$$m_{sk} = \frac{\mathrm{first}_{sk} - \mathrm{second}_{sk}}{\mathrm{second}_{sk} - \mathrm{first}^2_{sk}} \quad (68)$$

である．したがって，式 (66),(67) にしたがい，α_s を更新していけばよい．もっとも，原論文 [16] では違う導出結果を得ており，今後実験を通じて検証していきたいと考えている．

4.5 推論アルゴリズム

以下に，事後確率の推論・ハイパーパラメータの更新アルゴリズムをまとめる．

Algorithm 1 提案手法の推論アルゴリズム

```
Initialize ẑ, z randomly
for t = 1, 2, ··· , T do
    for d = 1, 2, ··· , D do
        for n = 1, 2, ··· , N_d do
            Update N̂_{d(−n)}, N_{ds(−n)}
            Sample ẑ^{(t)}_{d,n}, z^{(t)}_{d,n} from (56)
        end for
    end for
    Update α_s from (66) and (67)
end for
```

5. 結論

本稿では，文書におけるトピックの階層構造を単語埋め込みベクトル空間上で可視化し，単語埋め込み空間上でトピックの階層構造を分析する一歩として PAM と Gaussian LDA を組み合わせるモデルと推論手法を提案した．しかし，階層構造を持つトピックモデルは何も PAM だけではない．そのため今後は，実データを用いた数値実験によって提案モデルの性能を検討しつつ，PAM モデルを改変することや hPAM や hLDA など他の階層的トピックモデルも比較検討することで，より自然かつコンパクトに単語埋め込みベクトル空間上でトピック間の関係をとらえられるようにしていきたい．

参考文献

[1] R. Alghamdi, and K. Alfalqi, “A survey of Topic Modeling in Text Mining,” *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, pp. 147–153, 2015.
[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993—1022, Feb. 2003.
[3] D. M. Blei, and T. L. Griffiths, and M. I. Jordan, and J. B. Tenenbaum, “Hierarchical Topic Models and the Nested Chinese Restaurant Process,” *Neural Information Processing Systems*, 2004.
[4] D. M. Blei, and J. D. Lafferty, “A Correlated Topic Model of Science,” *The Annals of Applied Statistics*, vol. 1, no. 1, pp. 17–35, Apr. 2007.
[5] J. Chang, J. B-Graber, S. Gerrish, C. Wang, and D. M. Blei, “Reading tea leaves: How humans interpret topic models,” *Neural Information Processing Systems*, 2009.
[6] R. Das, M. Zaheer, and C. Dyer, “Gaussian LDA for Topic Models with Word Embeddings,” *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics*, pp. 795–804, Beijing, China, Jul. 2015.
[7] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis,” *Journal of the American Society of Information Science*, vol. 41, issue 6, pp. 391–407, Sep. 1990.
[8] T. L. Griffiths, and M. Steyvers, “Finding scientific topics,” *Proceedings of the National Academy of Science*, vol. 101, suppl. 1, pp. 5228–5235, Apr. 2004.
[9] T. Hofmann, “Probabilistic latent semantic indexing,” *Proceedings of the Twenty-Second Annual International SIGIR Conference*, pp. 289–296, 1999.
[10] P. Hu, W. Liu, W. Jiang, and Z. Yang, “Latent Topic Model Based on Gaussian-LDA for Audio Retrieval,” *Chinese Conference on Pattern Recognition*, vol. 321, pp.556–563, 2012.
[11] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, “Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey,” *Multimedia Tools and Applications*, pp.15169–15211, 2019.
[12] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” *Neural Information Processing Systems*, 2013.
[13] D. Mimno, “Mixtures of Hierarchical Topics with Pachinko Allocation,” *Proceedings of the 24th International Conference on Machine Learning*, pp. 363–371, 2007.
[14] K. P. Murphy, “Machine Learning: A Probabilistic Perspective,” The MIT Press, 2012.
[15] D. Newman, S. Karimi, and L. Cavedon, “External Evaluation of Topic Models,” *Proceedings of the 14th Australasian Document Computeing Symposium*, Sydney, Australia, Dec. 2009.
[16] W. Li, and A. McCallum, “Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations,” *Proceedings of the 23rd International Conference on Machine Learning*, pp. 577–584, Pittsburgh, PA, USA, 2006
[17] G. Xun, Y. Li, W. X. Zhao, J. Gao, and A. Zhang, “A Correlated Topic Model Using Word Embeddings,” *Proceedings of the 26th International Joint Conference on Artificical Intelligence*, pp. 4207–4213, 2017.