Dirichlet Process Unigram Mixture Modelに対するCollapsed Variational Bayes Inferenceの適用

佐藤一誠[†] 中川裕志[††]

Unigram Mixtureは教師なし文書分類などで幅広く使われている確率的生成モデルである．Unigram Mixtureは，混合モデルであり，実際の適用にはユーザは混合数決定問題をつねにかかえている．近年，このような混合モデルにおいて，Dirichlet Processを用いたノンパラメトリックベイズモデルが注目を集めている．Dirichlet Processを用いることでデータに合わせてモデル構造（混合数）を変化させることができる．本研究では，Dirichlet Processにより拡張したUnigram Mixtureに対して，Collapsed変分ベイズ法を用いてモデル学習する手法を示す．対数尤度とF-scoreによる評価により従来手法に対する有効性を確認した．

Collapsed Variational Bayes Inference for Dirichlet Process Unigram Mixture Model

Issei Sato[†] and Hiroshi Nakagawa[††]

Unigram Mixture is a probabilistic generative model that is widely used in unsupervised clustering of documents. Unigram Mixture is a mixture model and have a problem of how to determine the number of clusters. Recently, a nonparametric Bayes model using Dirichlet Process has gotten a lot of attention in this problem. Models using Dirichlet Process can determine the number of cluster corresponding to data. In this paper, we expand Unigram Mixture by Dirichlet Process and present a scheme that learns the model by Collapsed Variational Bayes inference.

1. はじめに

Unigram Mixture[1]は教師なし文書分類や情報検索などで幅広く使われている確率的生成モデルである．Unigram Mixtureは，混合モデルであり，実際の適用には混合数決定問題をユーザはつねに意識しなければならない．近年，このような混合モデルにおいて，Dirichlet Process[2]を用いたノンパラメトリックベイズモデル[3]が注目を集めている．Dirichlet Processを用いることでデータに合わせてモデル構造（混合数）を変化させることができる．このようなモデルは総称してDirichlet Process Mixtureと呼ばれている[3]．これまで，Dirichlet Process Mixtureの決定的な学習手法に変分ベイズ法[4),14)]の適用が提案されており，Gaussian Mixtureにおいてその効果が確認されている[5),6)]．Dirichlet Process Mixtureの変分ベイズ法は，混合する分布が指数分布族の場合には一般化されているため適用が容易である．変分ベイズ法の適用には，近似分布を導入するが，この近似分布に対して因子化という仮定（近似）を行う．この因子化仮定（近似）が，実際のデータのモデルによく合致していない場合は性能が低下する可能性を含んでおり，モデルの本来の性能を引き出せない可能性がある．このような問題に対して，近年，Collapsed Variational Bayes inference（Collapsed変分ベイズ法）が，Tehらによって提案された[8)]．Tehらは，LDA（Latent Dirichlet Allocation）というUnigram Mixtureを多重トピックへ拡張したモデルに対してCollapsed変分ベイズ法を適用しており，その効果が確認されている．ただし，ここで扱われているLDAは，Dirichlet Processを用いない有限混合モデルである．Dirichlet Processを用いたLDAの拡張は，Tehらによって提案されているが，学習はGibbs samplingにより行われている[7)]．また，Dirichlet Process MixtureのCollapsed変分ベイズ法は，Gaussian Mixtureに対しては適用

† 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
†† 東京大学情報基盤センター
Interfaculty Initiative in Information Studies

されており，その効果が確認されている[9]．しかし，Unigram Mixture は，Gaussian Mixture に比べ共分散構造をモデル化していないため，超高次元のデータに対してオーバフィットが起こりやすく，Gaussian Mixture における Dirichlet Process の効果がそのまま Unigram Mixture にもあてはまるとは限らない．また，Gaussian Mixture は連続モデルであり，離散モデルである Unigram Mixture とは性質が異なる．

よって，Unigram Mixture において，Dirichlet Process Mixture への拡張に対する Collapsed 変分ベイズ法の効果を確認することは重要であると考えられる．

したがって，本研究では，まず，従来の研究における Dirichlet Process Mixture の変分ベイズ法による学習手法を Unigram Mixture に対して適用し，その効果を評価する．次に，Collapsed 変分ベイズ法を用いた学習における更新式を導出し，その有効性を評価する．

以下，2 章では，本稿で扱う記号について説明する．3 章で，Dirichlet Process について説明する．4 章で，Dirichlet Process Unigram Mixture の定式化を行う．5 章では，変分ベイズ法について説明する．6 章では，Collapsed 変分ベイズ法の適用方法について説明する．7 章で，実験結果を示し，8 章でまとめを行う．

2. Terminology

本稿で扱う記号について以下説明する．W を語彙の総数とする．$\mathscr{W} = \{1, 2, \cdots, W\}$ を語彙のインデックス集合とする．N_d を文書 d における単語数とする（重複を含む）．$\mathbf{w}_d = (w_1, w_2, \cdots, w_{N_d})$ を文書 d 中の単語列とし，単語ベクトルと呼ぶことにする．w_n は，文書中の n 番目の単語を示す．つまり，$\mathbf{w}_d$ は文書 d そのものである．$\mathbf{x}_d = (x_1, x_2, \cdots, x_W)$ は，文書 d における各語彙の出現頻度を示す．x_v は，語彙 v の出現回数を示す．つまり，$\mathbf{x}_d$ は文書 d を BOW（Bag Of Words）表現に変換したベクトルである．

3. Dirichlet Process

3.1 Dirichlet Process の概要

Dirichlet Process は，1973 年に Ferguson によって提案された[2]．Ferguson による Dirichlet Process の定義（付録 A.2）は難解であるため，本節では，簡単な説明のみ行う．Dirichlet Process は，確率測度の（可算無限次元の）離散分布化を可能にする確率測度であると考えられる．もしくは，確率測度の（可算無限次元の）離散分布上の確率測度であると考えられる．Dirichlet Process は，集中度パラメータ α_0 と基底測度 G_0 からなる．集中度パラメータ α_0 は，離散化の程度・様相を変化させるパラメータである．基底測度 G_0 は，離散化するもとの確率測度である．つまり，Dirichlet Process は，「どの確率測度（基底測度 G_0）を」，「どの程度忠実に離散化するか（α_0）」を入力とすることで確率測度 G_0 の（可算無限次元の）離散分布 G を生成する．**図 1** は，Dirichlet Process によって，確率測度 G_0 が離散化され，離散確率測度 G が生成されている例を示している．

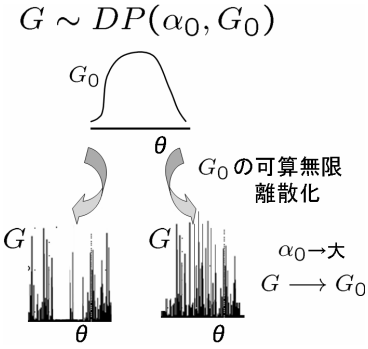

図 1 Dirichlet Process の例
Fig. 1 Example of Dirichlet Process.

集中度パラメータ α_0 を大きくすると，得られる離散分布 G は，基底測度 G_0 に近い離散分布となる．確率測度 G が Dirichlet Process に従うとき，$G \sim DP(\alpha, G_0)$ と表記する．Dirichlet Process の主な構成方法として，次節で説明する 2 つの構成方法が提案されている．

3.2 Dirichlet Process の構成方法

Dirichlet Process は，主に以下の 2 つの構成方法が提案されている．1 つ目の Dirichlet Process の構成方法は，Stick-breaking representation（SB）[10] である．SB は，次に示す可算無限個の確率変数 $\{\pi_k^{'}\}_{k=1}^{\infty}$ と $\{\phi\}_{k=1}^{\infty}$ から構成されている．

$$\pi_k^{'} \mid \alpha_0 \sim Beta(1, \alpha_0) \tag{1}$$
$$\phi_k \mid G_0 \sim G_0 \tag{2}$$

$Beta$ はベータ分布である．上記を用いて確率測度 $G \sim DP(\alpha, G_0)$ は以下のようにして構成される．

$$\pi_k = \pi_k^{'} \Pi_{i=1}^{k-1} (1 - \pi_i^{'}) \tag{3}$$
$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \tag{4}$$

本稿では，式 (1)，(3) を用いて π を構成する場合，$\pi\ (= \{\pi\}_{k=1}^{\infty}) \sim SB(\pi; \alpha_0)$ と表記する．ただし，変分ベイズ法などの決定的な方法における実際の実装には，無限を扱うことはできないので，上限を決める必

要がある[5]．SB におけるこの上限を T と表記する．

2 つ目の構成方法は，Dirichlet-Multinomial allocation（DMA）[11),12)] である．DMA は，有限次元（K）の確率測度 G_K を仮定し，π に対して以下に示す K 次元の symmetric Dirichlet 分布を用いる．

$$\pi \sim Dir(\pi; \frac{\alpha_0}{K}, \cdots, \frac{\alpha_0}{K}) \tag{5}$$

$$G_K = \sum_{k=1}^{K} \pi_k \delta_{\phi_k} \tag{6}$$

$K \to \infty$ のとき，DMA は Dirichlet Process に収束する[11),12)]．すなわち，$G_K \to G(K \to \infty)$．変分ベイズ法などの決定的な方法における実際の実装では，この K の上限を決める必要がある[6)]．本稿では，式 (5) を用いて π を構成する場合，π $(= \{\pi\}_{k=1}^{K}) \sim DMA(\pi; \alpha_0)$ と表記する．

4. Dirichlet Process Unigram Mixture

本章では，まず Unigram Mixture について説明し，次に，Dirichlet Process Unigram Mixture を説明する．

4.1 Unigram Mixture

Unigram Mixture は文章の生成過程を以下のようにモデル化した確率的生成モデルである．まず，各文書には K 個の潜在的なトピック（クラス）があると仮定する．各トピックの出現する確率を $\pi = (\pi_1, \cdots, \pi_K)$ とする．各トピックは，単語を生成する固有の確率分布 η を保持している．つまり，η_{kv} は，トピック k で語彙 v が生成される確率 $p(v|k)$ である．文書は以下のように生成されると仮定する．

(1) π からトピック z を生成：$z \sim p(z|\pi)$
(2) z に対応する η から単語 w を生成：$w \sim p(w|\eta, z)$

よって文書 $\mathbf{w}_d$ が生成される確率 $p(\mathbf{w}_d)$ は，以下のように定式化される．

$$p(\mathbf{w}_d|\eta, \pi) = \sum_{z} p(\mathbf{w}_d|\eta, z) p(z|\pi) \tag{7}$$

Unigram Mixture では，文書中の単語は相互に独立に生成されると仮定する．つまり，データ $\mathbf{w}_d$ 中の要素 w_{dn} 間に共分散構造をモデル化しない．よって，単語の結合分布を以下のように仮定する．

$$p(\mathbf{w}_d|\eta, z) = \Pi_{n=1}^{Nd} p(w_{dn}|\eta, z) \tag{8}$$

このように Unigram Mixture は，Gaussian Mixture と比べ共分散構造をモデル化していないため，超高次元のデータに対してオーバフィットが起こりやすいという性質を持つ．

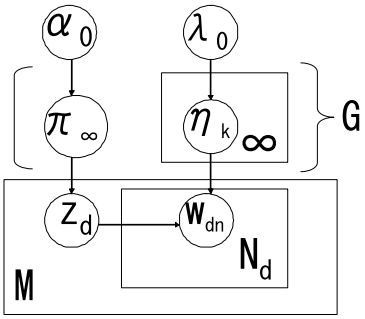

図 2 Dirichlet Process Mixture Model のグラフィカルモデル
Fig. 2 Graphical model of Dirichlet Process Mixture Model.

4.2 Dirichlet Process Unigram Mixture の定式化

Dirichlet Process Unigram Mixture は，文書集合 $D = \{\mathbf{w_d}\}_{d=1}^{M}$，混合比 π，$\mathbf{z} = \{z_d\}_{z=1}^{M}$（$z_d = k$ のとき文書 d がトピック k に属する），トピック k において語彙 v が生成される確率 $\eta = \{\eta_{kv}\}$（Multinomial 分布のパラメータ）とハイパーパラメータ α_0，λ を用いて，以下のように定式化される．

$$p(D|\alpha_0, \lambda) = \int \sum_{\mathbf{z}} p(D, \pi, \eta, \mathbf{z}|\alpha_0, \lambda) d\pi\eta \tag{9}$$

$$p(D, \pi, \eta, \mathbf{z}|\alpha_0, \lambda) = p(\pi|\alpha_0) p(\eta|\lambda) \Pi_{d=1}^{M} p(\mathbf{w}_d|z_d, \eta) p(z_d|\pi) \tag{10}$$

$p(\pi|\alpha_0)$ は，SB が適用される場合，$SB(\pi|\alpha_0)$ であり，DMA が適用される場合，$DMA(\pi|\alpha_0)$ である．$p(\mathbf{w}_d|z_d, \eta)$ は，トピック z_d のとき，文書 w_d が生成される（Multinomial 分布に従う）確率であり，式 (8) によって求まる．Unigram Mixture では，$p(\mathbf{w}_d|z_d, \eta) = \Pi_{n=1}^{N_d} p(w_{dn}|z_d, \eta)$ と仮定する．$p(z_d|\pi)$ は，$p(z_d = k|\pi) = \pi_k$ を示す．$p(\eta|\lambda)$ は，Multinomial 分布のパラメータの事前分布であり，λ をパラメータとする Dirichlet 分布である．**図 2** に，Dirichlet Process Unigram Mixture のグラフィカルモデルを示す．

5. 変分ベイズ法

式 (10) より事後確率分布 $p(\pi, \eta, \mathbf{z}|D, \alpha_0, \lambda)$ を求めることで，文書 d がトピック k に属する確率 $p(z_d = k)$ を推定することができる．$p(\pi, \eta, \mathbf{z}|D, \alpha_0, \lambda)$ は，ベイズの定理により求めることができるが，計算量が現実的ではないため，変分ベイズ法[4),14)] では，この事後分布の近似分布 $q(\pi, \eta, \mathbf{z})$ を求める．この近似分布

は，一般的には以下のような仮定をおく（因子化仮定（近似））．

$$\tilde{q}(\pi,\eta,\mathbf{z}) \simeq \tilde{q}(\pi)\tilde{q}(\eta)\Pi_{d=1}^{M}\tilde{q}(z_d) \tag{11}$$

$D(q,p)$ を Kullback-Leibler Divergence とすると，対数尤度 $\log p(D|\alpha_0,\lambda)$ は近似分布を用いて以下のように表すことができる．

$$\log p(D|\alpha_0,\lambda) = \mathscr{F}(\tilde{q}) + D(\tilde{q},p) \tag{12}$$

$$\mathscr{F}(\tilde{q}) = E_{\tilde{q}}[\log p(D,\pi,\eta,\mathbf{z}|\alpha_0,\lambda)] - E_{\tilde{q}}[\log \tilde{q}(\pi,\eta,\mathbf{z})] \tag{13}$$

$E_{\tilde{q}}$ は $\tilde{q}$ による期待値を示す．Kullback-Leibler Divergence は，確率分布間の距離を測る尺度であるので，$D(q,p)$ を最小にする q が望ましい．したがって，$\mathscr{F}(\tilde{q})$ を最大にする q を求めればよい．Dirichlet Process Mixture における変分ベイズ法の適用手法は，SB の場合は，Blei らが導出しており[5]，DMA の場合は，Yu らが導出している[6]．

$\theta = \{\eta,\pi\}$ とすれば，Dirichlet Process Unigram Mixture 更新式は，以下の式で表せる．

$$\tilde{q}(z_k) \sim \exp(E_{\tilde{q}(\theta)}[\log p(w_d, z_d = k|\theta,\alpha_0,\lambda)]) \tag{14}$$

$$\tilde{q}(\theta_i) \sim p(\theta_i)\exp(E_{\tilde{q}(\mathbf{z})\tilde{q}(\theta^{-i})}[\log p(D,\mathbf{z}|\theta,\alpha_0,\lambda)]) \tag{15}$$

6. Collapsed 変分ベイズ法

変分ベイズ法の因子化仮定（近似）は，実際のデータのモデルによく合致していない場合は性能が低下する可能性を含んでおり，モデルの本来の性能を引き出せない可能性がある．このような問題に対して，Collapsed 変分ベイズ法が近年提案された[8]．Collapsed 変分ベイズ法では，対象とするモデルにおいて因子化仮定（近似）が好ましくない変数の近似分布を真の事後分布と置き換えて消去（“collapse”）することで，近似分布をより真の事後分布に近づけることができる．どの変数を “collapse” するかは，モデルに依存するため一般的な議論はできないが，本研究で扱うモデルをもとに Collapsed 変分ベイズ法を次に説明する．

本研究では，Collapsed 変分ベイズ法における近似分布の因子化仮定（近似）を以下のように仮定する．

$$\hat{q}(\pi,\eta,\mathbf{z}) \simeq \hat{q}(\pi,\eta|\mathbf{z})\Pi_{d=1}^{M}\hat{q}(z_d) \tag{16}$$

π，η と $\mathbf{z}$ は強く依存すると考えられるので，$q(\pi,\eta|\mathbf{z})$ を真の事後分布 $p(\pi,\eta|\mathbf{z},D,\alpha_0,\lambda)$ と置き換える（代入する）と

$$D(\tilde{q}(\pi)\tilde{q}(\eta)\tilde{q}(\mathbf{z}),p) > D(p(\pi,\eta)\hat{q}(\mathbf{z}),p) \tag{17}$$

より

$$\tilde{F}(\tilde{q}) < \tilde{F}(\hat{q})\ (=\tilde{F}(p(\pi,\eta)\hat{q}(\mathbf{z}))) \tag{18}$$

となる．ここで

$$\tilde{F}(\hat{q}) = \tilde{F}(p(\pi,\eta)\hat{q}(\mathbf{z})) \tag{19}$$

$$= E_{\hat{q}(\mathbf{z})}[\log p(D,\mathbf{z}|\alpha_0,\lambda)] - E_{\hat{q}(\mathbf{z})}[\log \hat{q}(\mathbf{z})] \tag{20}$$

より，$\tilde{F}(\hat{q})$ を最大にする $\hat{q}(z_d)$ を求めればよい．以上により，π，η を “collapse” することができる．$\tilde{F}(\hat{q})$ を最大にする $\hat{q}(z_d)$ は，ラグランジュ乗数法により，以下のように求まる．

$$\hat{q}(z_d = k) \propto \exp(E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\log p(D,\mathbf{z}^{-\mathbf{d}},z_d = k|\alpha_0,\lambda)]) \tag{21}$$

$\mathbf{z}^{-\mathbf{d}}$ を，$\{z_d\}_{d=1}^{M} \setminus \{z_d\}$ とする．

また，

$$p(D,\mathbf{z}^{-\mathbf{d}},z_d = k|\alpha_0,\lambda) \propto p(z_d = k|D^{-d},\mathbf{z}^{-\mathbf{d}},\alpha_0,\lambda) \cdot \Pi_{n=1}^{N_d} p(w_{dn}|z_d = k,D^{-d},\mathbf{z}^{-\mathbf{d}},\alpha_{\mathbf{0}},\lambda) \tag{22}$$

より，式 (21) は積に分解して考えることができる．

ここで，$z_d = k$（$z_d \neq k$）のとき，$n_{dk} = 1(0)$ とし，$\sum_{d=1}^{M} n_{dk} = n_{\cdot k}$，$n_{\cdot k}^{-d} = n_{\cdot k} - n_{dk}$ とすると

SB の場合，

$$p(z_d = k|D^{-d},\mathbf{z}^{-\mathbf{d}},\alpha_0,\lambda) \propto \frac{1+n_{\cdot k}^{-d}}{1+\alpha+\sum_{t=k}^{T} n_{\cdot t}^{-d}} \Pi_{t=1}^{k-1} \frac{\alpha+\sum_{l=t+1}^{T} n_{\cdot l}^{-d}}{1+\alpha+\sum_{l=t}^{T} n_{\cdot l}^{-d}} \tag{23}$$

となり（付録 A.3），DMA の場合，

$$p(z_d = k|D^{-d},\mathbf{z}^{-\mathbf{d}},\alpha_0,\lambda) \propto \frac{\alpha/T + n_{\cdot k}^{-d}}{\alpha+\sum_{t=1}^{T} n_{\cdot t}^{-d}} \tag{24}$$

となる（付録 A.3）．

また，$n_{dkv} = \sharp\{n : w_{dn} = v, z_d = k\}$ とし，添え字を $\cdot$ としたものをその添え字での和とすると，

$$p(w_{dn}|z_d = k,D^{-d},\mathbf{z}^{-\mathbf{d}},\alpha_{\mathbf{0}},\lambda) \propto \frac{\lambda + n_{\cdot k w_{dn}}^{-d}}{W\lambda + n_{\cdot k \cdot}^{-d}} \tag{25}$$

となる（付録 A.4）．

式 (21) に，式 (23)，(24)，(25) を代入することで $\hat{q}(z_d = k)$ を得ることができる．ただし，Teh ら[8] と同様に実際の計算には Taylor 展開や中心極限定理による近似を用いた（付録 A.5）．

7. 実　　験

本章では，Dirichlet Process Unigram Mixture に対する Collapsed 変分ベイズ法の有効性を評価する．

7.1 データセット

7.1.1 人工データ

Multinomial 分布からサンプリングし，実験用のデータセットを以下のように作成した．作成したデー

タセットは，生成過程が既知の擬似的な文書集合と見なせる．データセットは，以下のように R☆を用いて作成した．

データセット 5000 は，5 つの 5000 次元 Multinomial 分布（トピック）から作成した．各トピックにおける文書数は，後述する BBC コーパスの各トピックにおける文書数の分布と同じものとした．つまり，トピック 1（510），トピック 2（386），トピック 3（417），トピック 4（511），トピック 5（401）とした．1 文書あたりの単語数を，500 に固定したものを，**データセット 5000.1**，100 から 900 の間の値をとる一様分布に従い可変としたものを，**データセット 5000.2** とする．

データセット 10000 は，5 つの 10000 次元 Multinomial 分布（トピック）から作成した．各トピックにおける文書数は，データセット 5000 と同様に，トピック 1（510），トピック 2（386），トピック 3（417），トピック 4（511），トピック 5（401）とした．1 文書あたりの単語数を，500 に固定したものを，**データセット 10000.1**，100 から 900 の間の値をとる一様分布に従い可変としたものを，**データセット 10000.2** とする．

データセット Dim (=5000 or 10000).1 における，トピック k に従う文書の作成を行う R のコードを以下示す

$$topic_k \leftarrow rdirichlet(1, rgamma(Dim, 1));$$
$$doc \leftarrow rmultinom(1, 500, topic_k)$$

データセット Dim (=5000 or 10000).2 における，トピック k に従う文書の作成を行う R のコードを以下示す

$$topic_k \leftarrow rdirichlet(1, rgamma(Dim, 1));$$
$$n \leftarrow round(runif(1, 0.1, 0.9) * 1000)$$
$$doc \leftarrow rmultinom(1, n, topic_k)$$

7.1.2 BBC コーパス

実際の文書セットとして BBC コーパス☆☆を用いた．文書数は 2225 文書で，語彙数は 9636 語である．トピックは，Business（510），Entertainment（386），Politics（417），Sport（511），Tech（401）の 5 つである（カッコ内は文書数）．

データセットの作成以外は，すべて C++で実装した．

7.2 評　　価

Dirichlet Process Mixture のような確率的生成モデルでは，一般的にはテストセットの対数尤度（もしくは Perplexity）を評価として用いる[8),9)]．

☆ http://www.r-project.org/
☆☆ http://mlg.ucd.ie/content/view/21/

以下のグラフ上の数値は，5 回の実験における平均値である．Dirichlet Process における集中度パラメータ α_0 は，すべて 1 として実験を行った．

図 3 は，データセット 5000.1 において各手法の負の対数尤度（negative loglikelihood）を比較したものである．SB，DMA におけるトピック数の上限（横軸）を変えて実験を行った．負の対数尤度（縦軸）は，低い値ほど良いモデルおよび学習手法であることを示す．図 3 が示すとおり，変分ベイズ法に比べ Collapsed 変分ベイズ法を用いた学習は，トピック数の上限にかかわりなく安定して負の対数尤度が低いことが分かる．また，学習方法の差は顕著に現れており，学習方法の違いに比べれば，2 つの構成方法 SB，DMA における違いはさほど見られない．

次に，情報検索や教師なし分類で用いられる F-score[13)] を用いて評価を行った．**図 4** は，データセット 5000.1 において各手法の F-score を比較したものである．F-score（縦軸）は，[0,1] 間の実数値をとり，1 ほど良い分類であることを示す．特に，F-score が 1 のときは，データの生成元の構造（分類）と完全に一致することを示す．変分ベイズ法では，比較的高い F-score を保持するものの，F-score に上下のばらつきが見られる．これに対し，Collapsed 変分ベイズ法

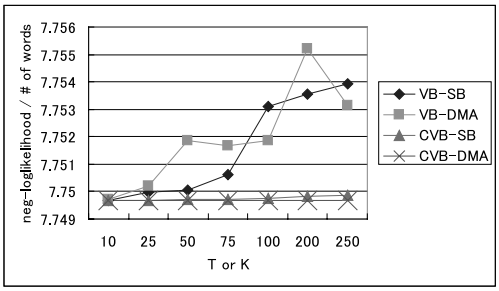

図 3 データセット 5000.1 を用いた負の対数尤度による比較
Fig. 3 Comparison of models with respect to negative loglikelihood of testset using Dataset 5000.1.

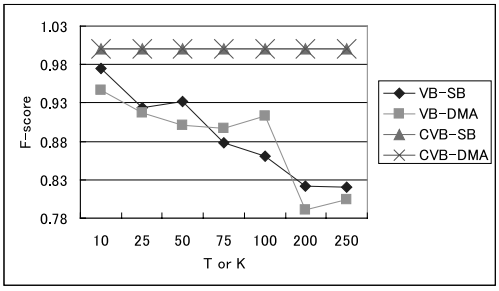

図 4 データセット 5000.1 における F-score による比較
Fig. 4 Comparison of models with respect to F-score using Dataset 5000.1.

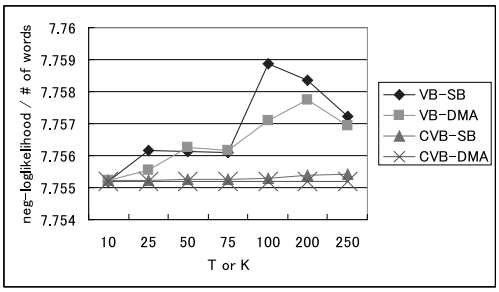

図 5 データセット 5000.2 を用いた負の対数尤度による比較
Fig. 5 Comparison of models with respect to negative loglikelihood of testset using Dataset 5000.2.

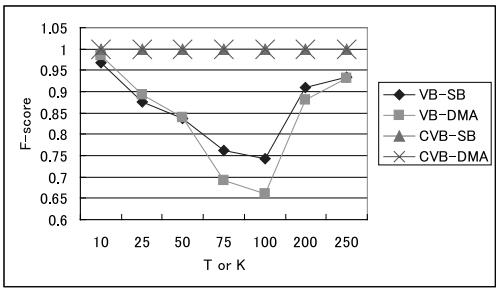

図 8 データセット 10000.1 における F-score による比較
Fig. 8 Comparison of models with respect to F-score using Dataset 10000.1.

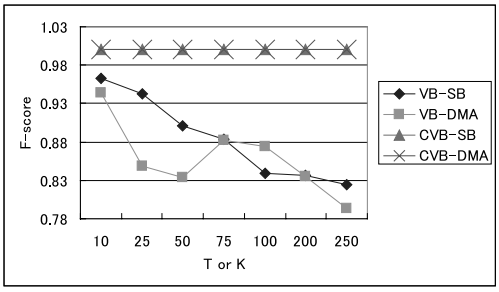

図 6 データセット 5000.2 における F-score による比較
Fig. 6 Comparison of models with respect to F-score using Dataset 5000.2.

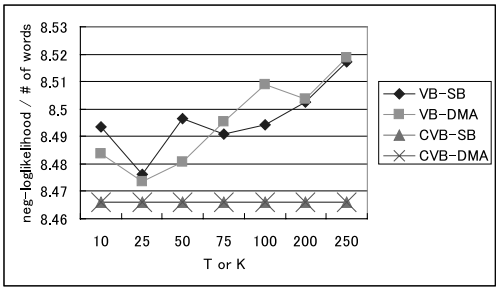

図 9 データセット 10000.2 を用いた負の対数尤度による比較
Fig. 9 Comparison of models with respect to negative loglikelihood of testset using Dataset 10000.2.

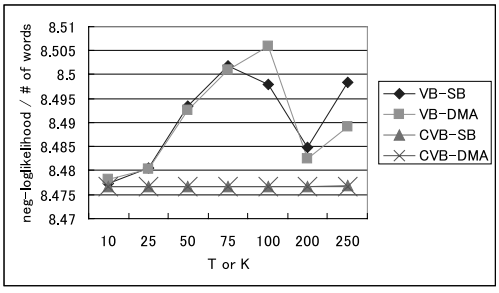

図 7 データセット 10000.1 を用いた負の対数尤度による比較
Fig. 7 Comparison of models with respect to negative loglikelihood of testset using Dataset 10000.1.

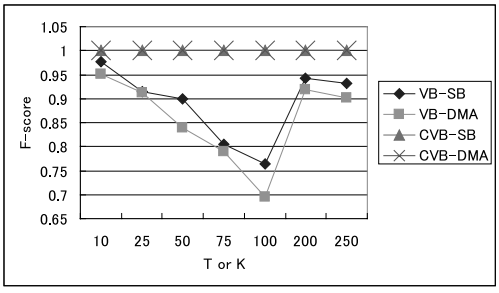

図 10 データセット 10000.2 における F-score による比較
Fig. 10 Comparison of models with respect to F-score using Dataset 10000.2.

では，どのようなトピック数の上限（横軸）に対しても F-score が 1 となる．つまり，Collapsed 変分ベイズ法では，完全にモデル推定ができている．

図 5 は，データセット 5000.2 において各手法の負の対数尤度（negative loglikelihood）を比較したものである．データセット 5000.1 と同様に Collapsed 変分ベイズ法の効果が確認できる．**図 6** の F-score においても，データセット 5000.1 と同様に Collapsed 変分ベイズ法の効果が確認できる．

同様に，**図 7**，**図 8**，**図 9**，**図 10** は，データセット 10000.1 および 2 における負の対数尤度と F-score を比較したものである．データセット 5000.1 および 2 の場合とほぼ同様に Collapsed 変分ベイズ法の効果が確認できる．

最後に，実際の文書のデータセットとして BBC コーパスを用いた．**図 11** は，各手法の負の対数尤度（negative loglikelihood）を比較したものである．テストセットは，全体の 10%とした．人工データと比べて，トピック数の上限による安定性は少ないものの，変分ベイズ法に比べ Collapsed 変分ベイズ法を用いた方が負の対数尤度は低いことが分かる．

図 12 は，各手法の F-score を比較したものであ

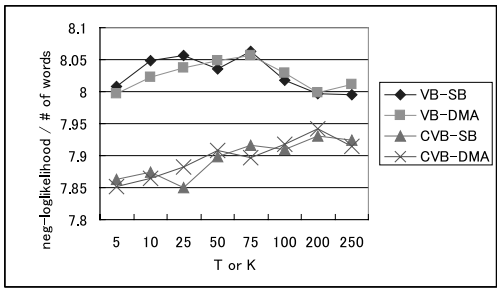

図 11 BBC コーパスを用いた負の対数尤度による比較
Fig. 11 Comparison of models with respect to negative loglikelihood of testset using BBC corpus.

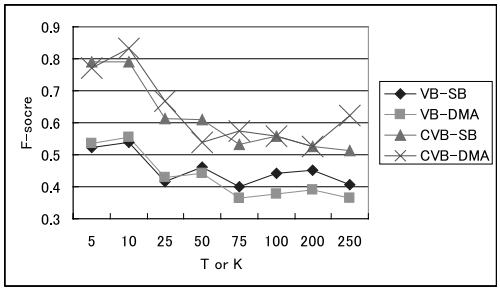

図 12 BBC コーパスを用いた F-score による比較
Fig. 12 Comparison of models with respect to F-score using BBC corpus.

る．BBC コーパスにおけるトピックを正解セットとして F-score を算出した．トピック数の上限を上げると，基本的にはどのモデルも F-score は正解トピック数（= 5）の場合と比べ減少していく．この理由は，トピック分類の良さに関する視点の違いによると考えられる．確率的生成モデルは，データが生成される確率，つまり尤度を高くする分類を良い分類であると仮定する．このため，モデルが尤度最大化を目標とした学習によりモデル構造を変化させるにつれて，現実の人の行った分類との違いが出てくる可能性が生じる．つまり，人によって決定されたトピック数との間に差が生じるため F-score が下がってしまう．むしろここで重要なのは，4 つのモデルともに，トピック数上限の 50 前後から F-score が安定していることである．これは，推定したトピック数が収束してきているためと考えられる．**図 13** に，各手法が推定したトピック数を示す．縦軸が推定したトピック数で，横軸が指定したトピックの上限数である．図 13 から，BBC コーパスの生成モデルの構造（トピック数）が 60 前後に収束していることが分かる．つまり，Dirichlet Process によるモデル選択が行われていると考えられる．変分ベイズ法においても，Dirichlet Process による効果は見られるが，Collapsed 変分ベイズ法は，変分ベイズ法よりも

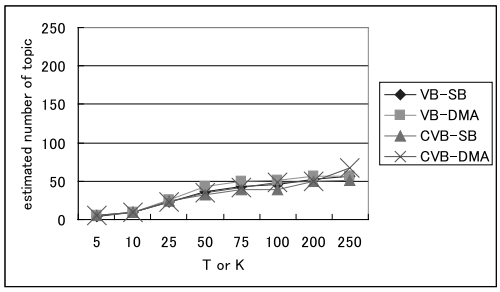

図 13 BBC コーパスにおいて推定されたトピック数
Fig. 13 Estimated number of topics in BBC corpus.

F-score が高いことから，より効果的な学習方法であることが分かる．

7.3 考　　察

前節のデータセットにおいて，対数尤度，F-score ともに Collapsed 変分ベイズ法の有効性が確認された．どの人工データにおいても，変分ベイズ法による学習では，Dirichlet Process による構造推定が，トピック数の上限に対してロバストに行えていない．

一方，Collapsed 変分ベイズ法では，トピック数の上限に対してロバストな Dirichlet Process による構造推定が可能である．また，Dirichlet Process Unigram Mixture では，構成方法（SB，DMA）の違いにはほとんど依存せず，学習方法（VB，CVB）による効果が結果に顕著に現れている．これは本実験で用いたどのデータセットに対してもあてはまる．

これに対して，Dirichlet Process Gaussian Mixture における Collapsed 変分ベイズ法の適用では，学習方法（VB，CVB）よりも構成方法（SB，DMA☆）に対数尤度による評価の結果が強く依存している[9]．

この理由として考えられるのは，Unigram Mixture が共分散構造をモデル化していないため，データに対するオーバフィットが起こりやすいことである．Collapsed 変分ベイズ法は，式 (21) からも分かるように，$q(z_d = k)$ の推定に $q(\mathbf{z}^{-\mathbf{d}})$ を用いる．つまり，$q(z_d = k)$ の推定に，1 ステップ前の $q(z_d = k)$ の情報を用いない．変分ベイズ法は，式 (14)，(15) からも分かるように，$q(z_d = k)$ の推定に $q(\eta)$，$q(\pi)$ を用いる．$q(\eta)$，$q(\pi)$ は，$q(z_d = k)$ も用いて推定するため，$q(z_d = k)$ の推定に，1 ステップ前の $q(z_d = k)$ の情報も用いていることになる．したがって，Collapsed 変分ベイズ法は，従来の変分ベイズ法よりも $q(z_d = k)$ の学習における汎化能力が高いと考えられ，データに対してオーバフィットを起こしやすい Unigram Mixture

☆ 文献 9) では FSD と表記されている．

には有効であると考えられる．

8. おわりに

本研究では，Dirichlet Process により拡張した Unigram Mixture に対して，Collapsed 変分ベイズ法を用いてモデル学習する手法とその有効性を示した．

今後は，Unigram Mixture の上位モデルである Dirichlet Mixture への適用方法を検討する予定である．

謝辞 本研究は，文科省科学研究費特定領域研究「情報爆発」の補助を得て行われた．

参考文献

1) Nigam, K., McCallum, A.K., Thrun, S. and Mitchell, T.M.: Text Classification from Labeled and Unlabeled Documents using EM, *Machine Learning*, Vol.39, pp.103–134 (2000).
2) Ferguson: A Bayesian Analysis of Some Nonparametric Problems, *The Annals of Statistics*, Vol.1, No.2 (1973).
3) Antoniak: Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems, *The Annals of Statistics*, Vol.2, No.6 (1974).
4) Attias, H.: Learning parameters and structure of latent variable models by Variational Bayes, *Proc. Uncertainty in Artificial Intelligence* (1999).
5) Blei, D.M. and Jordan, M.I.: Variational inference for Dirichlet process mixtures, *Journal of Bayesian Analysis*, Vol.1, No.1, pp.121–144 (2005).
6) Yu, S., Yu, K., Tresp, V. and Kriegel, H.P.: Variational Bayesian Dirichlet-Multinomial Allocation for Exponential Family Mixtures, *European Conference on Machine Learning* (*ECML2006*), pp.841–848 (2006).
7) Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M.: Hierarchical Dirichlet Processes, Technical Report 653, Department of Statistics, University of California, Berkeley (2004).
8) Teh, Y.W., Newman, D. and Welling, M.: A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation, *Neural Information Processing Systems* (*NIPS 2006*) (2006).
9) Kurihara, K., Welling, M. and Teh, Y.W.: Collapsed Variational Dirichlet Process Mixture Models, *The 20th International Joint Conference on Artificial Intelligence* (*IJCAI 2007*) (2007).
10) Sethuraman: A Constructive Definition of Dirichlet Priors, *Statistica Sinica*, Vol.4, pp.639–650 (1994).
11) Green, P.J. and Richardson, S.: Modelling Heterogeneity With and Without the Dirichlet Process, *Scandinavian Journal of Statistics*, Vol.28, No.2, pp.355–375 (2001).
12) Ishwaran, H. and Zarepour, M.: Exact and approximate sum representations for the Dirichlet process, *The Canadian Journal of Statistics*, Vol.30, No.2, pp.269–283 (2002).
13) Larsen, Bjornar, Aone and Chinatsu: Fast and effective text mining using lineartime document clustering, *International Conference on Knowledge Discovery and Data Mining* (1999).
14) 上田修功：計算統計 I（統計科学のフロンティア 11），第 1 版，III 章 4 節 (2003).

付 録

A.1 Dirichlet 分布

和が 1 になる K 個の確率変数の分配比率 π を考える．$\pi = (\pi_1, \pi_2, \cdots, \pi_K) \in [0,1]^K, \pi_i > 0, \sum_{i=1}^K \pi_i = 1$ とすると，π は，$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_K) : \alpha_i > 0$ をパラメータとして以下の確率分布に従う．

$$\pi \sim Dir(\alpha) \tag{26}$$

$$\Longleftrightarrow p(\pi|\alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\Pi_{i=1}^K \Gamma(\alpha_i)} \Pi_{i=1}^K \pi_i^{\alpha_i - 1} \tag{27}$$

上記の確率分布を K 次元の Dirichlet 分布という．特に，2 次元（$K = 2$）の場合，Beta 分布となる．$\Gamma(x)$ はガンマ関数である．

π の期待値は以下のようにして求めることができる．

$$E[\pi] = \frac{\alpha_i}{\sum_{i=1}^K \alpha_i} \tag{28}$$

A.2 Dirichlet Process

Dirichlet Process の正確な定義は以下のとおりである[2)]．

Ω を全事象，B を Borel 集合族とする．確率測度 G_0 が定義された可測空間（Ω，B）を考える．α_0 を正の実数とする．このとき確率測度 G に対して，任意の m について，Ω の任意の分割 $A_1, \cdots, A_m$ を考えたとき，$(G(A_1), \cdots, G(A_m))$ が $(\alpha_0 G_0(A_1), \cdots, \alpha_0 G_0(A_m))$ をパラメータとする Dirichlet 分布に従うとき，G は Dirichlet Process に従う（$G \sim DP(\alpha_0, G_0)$）という．α_0 は集中度パラメータ，G_0 は基底測度と呼ばれる．

A.3

$$p(z_d = k|D^{-d}, \mathbf{z}^{-\mathbf{d}}, \alpha_0, \lambda) = \int p(z_d = k|\pi) p(\pi|D^{-d}, \mathbf{z}^{-\mathbf{d}}, \alpha_0, \lambda) d\pi \tag{29}$$

確率変数間の条件付き独立性により

$$p(z_d = k|D^{-d}, \mathbf{z}^{-\mathbf{d}}, \alpha_0, \lambda) = \int p(z_d = k|\pi)p(\pi|\mathbf{z}^{-\mathbf{d}}, \alpha_0)d\pi \quad (30)$$

SB の場合

$$p(z_d = k|\pi) = \pi_k = \pi_k^{'}\Pi_{i=1}^{k-1}(1-\pi_k^{'}) \quad (31)$$

$$p(\pi|\mathbf{z}^{-\mathbf{d}}, \alpha_0) \propto p(\mathbf{z}^{-\mathbf{d}}|\pi)p(\pi|\alpha_0) \quad (32)$$
$$\propto \Pi_{k=1}^{T}\pi_k^{n_{\cdot k}^{-d}}\ \Pi_{k=1}^{T}Beta(\pi^{'};1,\alpha_0) \quad (33)$$
$$\propto \Pi_{k=1}^{T}\pi_k^{'n_{\cdot k}^{-d}}(1-\pi_k^{'})^{\alpha_0-1+\sum_{i=k+1}^{T}n_{\cdot k}^{-d}} \quad (34)$$
$$\propto \Pi_{k=1}^{T}Beta(\pi^{'};n_{\cdot k}^{-d}+1,\alpha_0+\sum_{i=k+1}^{T}n_{\cdot k}^{-d}) \quad (35)$$

上記，式 (30)，(31)，(35) により，式 (23) を導くことができる（付録 A.1 の式 (28)）．

DMA の場合

$$p(z_d = k|\pi) = \pi_k \quad (36)$$
$$p(\pi|\mathbf{z}^{-\mathbf{d}}, \alpha_0) \propto p(\mathbf{z}^{-\mathbf{d}}|\pi)p(\pi|\alpha_0) \quad (37)$$
$$\propto \Pi_{k=1}^{K}\pi_k^{n_{\cdot k}^{-d}}\ Dir(\pi;\frac{\alpha_0}{K},\cdots,\frac{\alpha_0}{K}) \quad (38)$$
$$\propto \Pi_{k=1}^{K}\pi_k^{n_{\cdot k}^{-d}}\ \Pi_{k=1}^{K}\pi_k^{\frac{\alpha}{K}-1} \quad (39)$$
$$\propto \Pi_{k=1}^{K}\pi_k^{\frac{\alpha}{K}-1+n_{\cdot k}^{-d}} \quad (40)$$
$$\propto Dir(\pi;\frac{\alpha_0}{K}+n_{\cdot 1}^{-d},\cdots,\frac{\alpha_0}{K}+n_{\cdot K}^{-d}) \quad (41)$$

上記，式 (30)，(36)，(41) により，式 (24) を導くことができる（付録 A.1 の式 (28)）．

A.4

$$p(w_{dn}|z_d = k, D^{-d}, \mathbf{z}^{-\mathbf{d}}, \alpha_{\mathbf{0}}, \lambda) = \int p(w_{dn}|z_d = k, \eta)p(\eta|D^{-d}, \mathbf{z}^{-\mathbf{d}}, \alpha_{\mathbf{0}}, \lambda) \quad (42)$$

確率変数間の条件付き独立性により

$$p(w_{dn}|z_d = k, D^{-d}, \mathbf{z}^{-\mathbf{d}}, \alpha_{\mathbf{0}}, \lambda) = \int p(w_{dn}|z_d = k, \eta)p(\eta|D^{-d}, \mathbf{z}^{-\mathbf{d}}, \lambda)d\eta \quad (43)$$
$$p(w_{dn}|z_d = k, \eta) = \eta_{kw_{dn}} \quad (44)$$
$$p(\eta|D^{-d}, \mathbf{z}^{-\mathbf{d}}, \lambda) \propto p(D^{-d}|\mathbf{z}^{-\mathbf{d}}, \eta)p(\eta|\lambda) \quad (45)$$
$$\propto \Pi_k\Pi_{d'\neq -d}\Pi_v\eta_{kv}^{n_{d'kv}} \cdot Dir(\eta;\lambda) \quad (46)$$
$$\propto \Pi_k\Pi_v\eta_{kv}^{n_{\cdot kv}^{-d}} \cdot \Pi_k\Pi_v\eta_{kv}^{\lambda-1} \quad (47)$$
$$\propto \Pi_k\Pi_v\eta_{kv}^{\lambda-1+n_{\cdot kv}^{-d}} \quad (48)$$
$$\propto \Pi_k Dir(\eta_{\mathbf{k}};\lambda+n_{\cdot k1}^{-d},\cdots,\lambda+n_{\cdot kW}^{-d}) \quad (49)$$

上記，式 (43)，(44)，(49) により，式 (25) を導くことができる（付録 A.1 の式 (28)）．

A.5

V を分散とする．対数関数の 2 次の Taylor 展開により以下のように近似できる．

$$E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\log(\alpha n_{\cdot k}^{-d})] \quad (50)$$
$$\approx \log(\alpha + E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot k}^{-d}]) - \frac{V_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot k}^{-d}]}{2(\alpha + E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot k}^{-d}])^2} \quad (51)$$
$$E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\log p(z_d = k|D^{-d}, \mathbf{z}^{-\mathbf{d}}, \alpha_0, \lambda)] \quad (52)$$

は，SB の場合，式 (23) と Taylor 展開による近似により

$$\begin{aligned}
&E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\log p(z_d = k|D^{-d}, \mathbf{z}^{-\mathbf{d}}, \alpha_0, \lambda)] \\
&\propto log\frac{1+E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot k}^{-d}]}{1+\alpha+E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\sum_{t=k}^{T}n_{\cdot t}^{-d}]} \\
&+\sum_{t=1}^{k-1}log\frac{\alpha+E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\sum_{l=t+1}^{T}n_{\cdot l}^{-d}]}{1+\alpha+E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\sum_{l=t}^{T}n_{\cdot l}^{-d}]} \\
&-\frac{V_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot k}^{-d}]}{2(1+E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot k}^{-d}])^2} \\
&+\frac{V_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\sum_{t=k}^{T}n_{\cdot t}^{-d}]}{2(1+\alpha+E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\sum_{t=k}^{T}n_{\cdot t}^{-d}])^2} \\
&-\sum_{t=1}^{k-1}\frac{V_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\sum_{l=t+1}^{T}n_{\cdot l}^{-d}]}{2(\alpha+E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\sum_{l=t+1}^{T}n_{\cdot l}^{-d}])^2} \\
&+\sum_{t=1}^{k-1}\frac{V_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\sum_{l=t}^{T}n_{\cdot l}^{-d}]}{2(1+\alpha+E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\sum_{l=t}^{T}n_{\cdot l}^{-d}])^2} \quad (53)
\end{aligned}$$

となる．DMA の場合も同様に，式 (24) と Taylor 展開による近似により

$$\begin{aligned}
&E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\log p(z_d = k|D^{-d}, \mathbf{z}^{-\mathbf{d}}, \alpha_0, \lambda)] \\
&\propto log(\alpha/K + E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot k}^{-d}]) \\
&-\frac{V_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot k}^{-d}]}{2(\alpha/K+E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot k}^{-d}])^2} \quad (54)
\end{aligned}$$

となる（式 (24) の分母は定数なので k とは無関係）．

$E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\log p(w_{dn}|z_d = k, D^{-d}, \mathbf{z}^{-\mathbf{d}}, \alpha_{\mathbf{0}}, \lambda)]$ も同様に式 (25) と Tayhor 展開による近似により

$$\begin{aligned}
&E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[\log p(w_{dn}|z_d = k, D^{-d}, \mathbf{z}^{-\mathbf{d}}, \alpha_{\mathbf{0}}, \lambda)] \\
&\propto log\frac{\lambda+E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot kw_{dn}}^{-d}]}{W\lambda+E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot k\cdot}^{-d}]} \\
&-\frac{V_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot kw_{dn}}^{-d}]}{2(\lambda+E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot kw_{dn}}^{-d}])^2} \\
&+\frac{V_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot k\cdot}^{-d}]}{2(W\lambda+E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot k\cdot}^{-d}])^2} \quad (55)
\end{aligned}$$

となる．

上記で用いられている期待値 E や分散 V の計算は高い計算コストがかかる．しかし，n_{dk} は，確率を $q(z_d = k)$ とする Bernoulli 分布に従う確率変数と見

なすことができるので，以下のような中心極限定理を用いた近似を行うことができる[8]．

$$E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot k}^{-d}] = \sum_{d' \neq d} q(z'_d = k) \tag{56}$$

$$V_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot k}^{-d}] = \sum_{d' \neq d} q(z'_d = k)(1 - q(z'_d = k)) \tag{57}$$

n_{dkv} は，文書 d がトピック k に属するときの語彙 v の出現頻度と見なせるので，同様に

$$E_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot kv}^{-d}] = \sum_{d' \neq d} q(z'_d = k) \cdot x_{d'v} \tag{58}$$

$$V_{\hat{q}(\mathbf{z}^{-\mathbf{d}})}[n_{\cdot kv}^{-d}] = \sum_{d' \neq d} q(z'_d = k)(1 - q(z'_d = k)) \cdot x_{d'v} \tag{59}$$

と近似できる．

(平成 19 年 4 月 19 日受付)
(平成 19 年 6 月 7 日再受付)
(平成 19 年 7 月 20 日採録)

佐藤 一誠

2007 年現在，東京大学大学院情報理工学系研究科に在籍．半構造データマイニングやノンパラメトリックベイズモデルを中心とした機械学習，自然言語処理の研究に従事．

中川 裕志（正会員）

1975 年東京大学工学部卒業，1980 年東京大学大学院博士課程修了（工学博士），同年より横浜国立大学勤務．1999 年より東京大学情報基盤センター教授，現在に至る．機械学習，自然言語処理，Web 情報アクセスの研究に従事．2004 年より 2006 年まで言語処理学会会長．2006 年より情報処理学会自然言語研究会主査．