ガウス過程に基づく連続空間トピックモデル

持橋 大地 [1]　　吉井 和佳 [2]　　後藤 真孝 [2]

[1] 統計数理研究所
〒 190-8562 東京都立川市緑町 10-3
daichi@ism.ac.jp

[2] 産業技術総合研究所
〒 305-8568 茨城県つくば市梅園 1-1-1
{k.yoshii,m.goto}@aist.go.jp

概要

本論文では, 単語に潜在空間における座標を明示的に与え, その上でのガウス過程を考えることで, 通常の混合モデルに基づくトピックモデルより高精度なテキストモデルが得られることを示す. 提案法は潜在層が二値ではなく, ガウス分布に従う RBM の生成モデルともみることができ, MCMC 法により RBM に匹敵する性能を局所解に陥ることなく, 容易に達成できる. こうして単語や文書の潜在座標を学習することは他の多くの応用や, 可視化にも自然に繋がる基本的なモデルである.
キーワード: ガウス過程, トピックモデル, 潜在意味解析, MCMC, RBM

Modeling Text through Gaussian Processes

Daichi Mochihashi[1]　Kazuyoshi Yoshii[2]　Masataka Goto[2]

[1] The Institute of Statistical Mathematics　[2] National Institute of Advanced Industrial Science and Technology

Abstract

This paper proposes a continous space text model based on Gaussian processes. Introducing latent coordinates of words over which the Gaussian process is defined, we can encode word correlations directly and lead to a model that performs better than mixture models. Our model would serve as a foundation of more complex text models and also as a statistical visualization of texts.
Keywords: Gaussian process, Topic models, Latent semantic analysis, MCMC, RBM

1　はじめに

テキストや語に潜む見えない “意味” を表現するための統計モデルとして, 1999 年の PLSI [1] によって提案され, 2001 年の LDA [2][3] に受け継がれた確率的潜在意味解析, またはトピックモデルは自然言語処理の内外で広範に適用されてきた.

LDA は統計的には混合モデルの一種[1]であり, トピック別単語分布 $p(w|k)$ $(k=1\cdots K)$ を, 文書が持つ各トピックへの所属確率 $\theta=(\theta_1,\cdots,\theta_K)$ $(\sum_{k=1}^{K}\theta_k=1)$ で混合することで, 文書での単語分布を表現する. 新聞などの文書集合について適用すると, 「政治」「国際経済」「国内経済」「芸術」…などのトピックに対応する単語分布が教師なしで得られることが確かめられている.

いっぽう, 制限ボルツマンマシン (RBM) に基づく一種のニューラルネットである RaP [4], RSM [5] などのモデルは, 混合モデルではなく積モデル (Product of Experts [6]) であり, 文書の話題を国際経済=‘国際’×‘経済’, 民俗音楽=‘民俗’×‘音楽’ などのより基本的な要素の積として表現できるため, 一般に LDA を超える性能を持つといわれている. これらは多層化することで “Deep Learning” といわれ, より高性能となることも報告されている [7][8].

しかしながら, 潜在層が 1/0 の二値のベクトルである RBM は最適化がきわめて難しく, 学習率などの多数のパラメータの組み合わせを各データに合わせて発見的に調節しなければ, 品質の低い局所解しか得られない. また, 生成モデルを持たない無向グラフの “ニューラルネットワーク” であり,

- 何が学習されているのかが分からない
- RBM 以外の他のモデルとの接続が難しい

という大きな問題を持っている. こうした理由から, 実際には RBM は研究の場以外ではほとんど使われてこなかった.

そこで本論文では, Latent space models [9][10] の考え方に基づいて, 単語に潜在空間における座標を明示的に与え, この上でのガウス過程を考えることで, 単語および文書の意味をデータから学習することのできる連続空間トピックモデル (CSTM) を提案する.

RBM と異なり, CSTM では文書の潜在表現は二値ではなく, ガウス分布に従う連続値のベクトルである. このため他のモデルとの接続や, 共変量の導入も容易であり, さらに可視化にも自然に繋がる. 特に, その構造から LDA と異なり, 語彙の確率を直接制御することができるため, 文書や単語自体の素性に基づいた, よりきめ細かいモデル化が可能になる. CSTM は完全なベイズ生成モデルであるため, 学習は局所解

[1] 正確には, 混合モデルの混合モデル.

の影響がなく, MCMC で最適化を行うことで RBM に匹敵する性能を容易に達成できる.

以下では, 2 章で背景を説明し, 3 章で CSTM のモデルを導入する. 4 章で学習について, 5 章で共変量を用いた拡張について述べた後, 6 章で実験を行い, 7 章で全体をまとめる.

2 トピックモデルと潜在意味表現

確率的潜在意味解析, またはトピックモデルとは, 文書ごとの単語出現のばらつきを表現するための確率モデルであり, その代表である LDA (Latent Dirichlet Allocation [3]) では, 離散データ, すなわち単語の集合である文書 $\mathbf{w} = w_1 w_2 \cdots w_N$ が, 次のようにして生成されたと仮定する.

1. トピック (話題) 分布 $\theta \sim \mathrm{Dir}(\alpha)$ を生成.
2. For $n = 1 \cdots N$,
   (a) トピック $z \sim \mathrm{Mult}(\theta)$ を選択.
   (b) 単語 $w_n \sim p(w|z)$ を生成.

ここで $\mathrm{Mult}(\theta)$ は多項分布, $\mathrm{Dir}(\alpha)$ はディリクレ分布であり, θ および $\beta = \{p(w|z)\}$ $(z = 1 \cdots K)$ が LDA のパラメータである.

LDA は広く使われており, 時間や文書のラベルなどの共変量の利用といった無数の拡張があるが, そのほぼ全てがトピック分布 θ の制御のみに関わっており, 実際に単語を生成する確率分布 β は固定となっている. これはトピックの数を K, 語彙の数を V としたとき, KV 個の値を持つ β を直接パラメータとする LDA では, これを動的に変更することが難しいためである. 基本となる β から地域別の β' を生成する [11] のような研究もあるが, 有限個のバリエーションを生成するのみであり, 各文書に応じて β を適切に変更することはできない.

ここで上の生成モデルを考えてみると, ステップ 2 は (a)(b) をあわせて

$$w_n \sim p(w) = \textstyle\sum_{k=1}^{K} \theta_k p(w|k) \qquad (1)$$

とできるから, これは図 1 のように単語の出現確率 $\mathbf{p} = (p_1, p_2, \cdots, p_V)$ の存在する単語単体の内部に $\beta_1 \cdots \beta_K$ の張る K 次元のトピック単体を考え, これを θ で内分する点を $\mathbf{p}$ としていることがわかる. 図 3 に, LDA からランダムに生成した多項分布 $\mathbf{p}$ の例を示した. これから明らかなように, LDA では β に

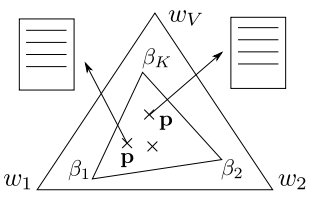


図 1: LDA による多項分布 $\mathbf{p}$ の生成. $\beta = \{p(w|z)\}$ で張られるトピック単体の内部を θ の割合で内分した点が $\mathbf{p}$ であり, $\mathbf{p}$ から各文書の単語が生成される.

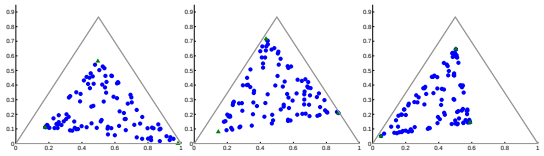


図 2: LDA の生成モデルからランダムに生成した単語分布 $\mathbf{p}$. 混合モデルである LDA では, 単語単体中の低次元トピック単体の内部しかモデル化できない.

よって決まるトピック単体の内部しかモデル化できない. しかし, 上の議論から $\mathbf{p}$ をモデル化する際にはトピック z は局外母数 (nuisance parameter) であり, 必ずしも必要ではない. ゆえに, 何らかの方法で直接 $\mathbf{p}$ を求めることが考えられる.

意味の分散表現 もう一つの観点は, 意味の分散表現である. LDA は混合モデルであり, 学習の際にはデータの各単語にそれを生成した潜在トピックを割り当ててゆくが, これは単語をどれかのクラスに確率的にクラスタリングしていることを意味する. しかし実際は, should は助動詞+過去という要素に分解できたり, 計算言語学の論文は言語学+数学+統計学の要素を持っていたりするように, 文書や単語の意味は複数の素性の束としても捉えられるはずである.

こうした方向のモデルとして, 図 3 のように出力層 $\mathbf{v} = (v_1, \cdots, v_V)^T$ と通常は二値の潜在層 $\mathbf{h} = (h_1, \cdots, h_H)^T$, $h_i \in \{0, 1\}$ が重み W を持つ二分グラフで結びついている制限ボルツマンマシン (Restricted Boltzmann Machine, RBM) は因子分解による表現を持っており, 次の式で $\mathbf{v}$ と $\mathbf{h}$ の同時確率が表される.

$$p(\mathbf{v}, \mathbf{h}) = \frac{\exp(\mathbf{v}^T W \mathbf{h})}{\sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp(\mathbf{v}^T W \mathbf{h})} \qquad (2)$$

(2) 式の分子は

$$\exp(\textstyle\sum_i \sum_j w_{ij} v_i h_j) = \prod_i \prod_j e^{w_{ij} v_i h_j} \qquad (3)$$

と書けるから, これは混合モデルではなく積モデル (Product of Experts [6]) であることに注意されたい.

実際に, $\mathbf{v}$ を期待値としたポアソン分布から単語が生成されたとする Rate Adapting Poisson (RAP) モデル [4] や, それを正規化して多項分布にすることで可変長の文書を表現できる Replicated Softmax Model (RSM) [5] は, 混合モデルより良い性能を示すと報告されている.

しかしながら, RBM は一般に最適化が非常に難し

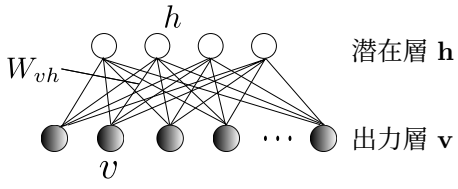


図 3: RBM による因子分解表現. 出力層 $\mathbf{v}$ と $\{0, 1\}$ の値をとる潜在層 $\mathbf{h}$ が重み W_{vh} で結びついている.

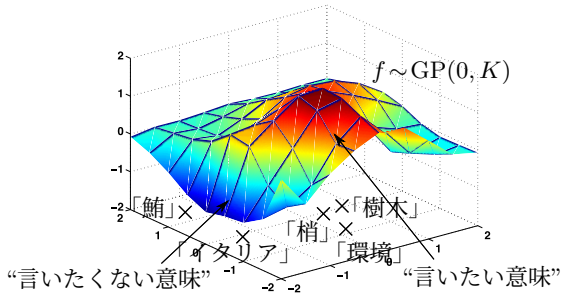


図 4: 単語の潜在座標とガウス過程に従う関数 f.

く, 学習率, 初期値, モーメント, ミニバッチの大きさなどの最適化パラメータをすべて試行錯誤により適切に設定しなければ, 品質の悪い局所解しか得られない. また, 何が学習されているのか分からないという理論的な側面に加え, 一種のニューラルネットである RBM はそれ以外のモデルとの連携が難しいという問題も持っている. こうした理由から, RBM は研究の場を離れてはあまり使用されてこなかった.

しかし, RBM の基礎となっている分散表現の概念は重要であり, これを, より自然な統計モデルとする仕組みが望まれる.

3 連続空間トピックモデル

ここで考え直してみると, 実は上のトピックモデルでは, 結果的に単語 w の潜在意味表現として, LDA では確率分布 $p(k|w) \propto p(w|k)p(k)$ を, RBM では重み W_{wk} を持っており, これらが最も重要なパラメータとなっていることに気づく.

そこで我々は, 各単語 w が d 次元の潜在座標 $\phi(w) \sim \mathrm{N}(0, I_d)$ を持っていると仮定する. このとき, 意味的に関連のある単語の確率を同時に大きくするためには, 図 4 のように, この上に平均 0 のガウス過程に従う関数 (曲面)

$$f \sim \mathrm{GP}(0, K) \tag{4}$$

を生成し, 単語の確率を

$$p(w) \propto e^{f(w)} G_0(w) \tag{5}$$

とモデル化する. $G_0(w)$ は単語 w の「デフォルト」確率であり, 以下では最尤推定 $G_0(w) = n(w) / \sum_w n(w)$ ($n(w)$ は w の頻度) を考える.

ガウス過程 [12] とは, ランダムな回帰関数を生成する確率過程であり, 実際には単純に無限次元のガウス分布のことである. カーネル行列 K の要素 $K_{ij} = k(w_i, w_j)$ が近いほど, 対応する出力 $f(x_i)$, $f(x_j)$ も近くなり, 本研究では線形カーネル $k(w_i, w_j) = \phi(w_i)^T \phi(w_j)$ を用いる. f は直感的には, 「この文書で言いたいこと」を表している.

(5) 式はポアソン過程の Multiplicative intensity モデル [13] または対数ガウス Cox 過程 [14] と同じ考え方に基づくもので, 確率 G_0 が文書によって e^f 倍される. 後に述べるように, f はほぼ $-9 < f < 9$ 前後の値をとるので, $f > 0$ の領域では意味的に関連のある単語の確率は $e^1 \sim 2.72$ 倍から最大で $e^9 \sim 8100$ 倍程度になり, $f < 0$ の意味的に無関係な語の確率は $e^{-1} \sim 0.37$ 倍から $e^{-9} \sim 0.0001$ 倍程度に抑えられる. f は平均 0 のため, 期待値は $E[p(w)] \propto e^0 G_0(w) = G_0(w)$ であることに注意しよう.

実際のテキストはどうなっているのだろうか. 一つの目安として, Brown コーパスの各文書の単語について, 最尤推定による文書内での単語確率 $\hat{p}(w|d)$ と全体での確率 $\hat{p}(w)$ の対数尤度比 $\log(\hat{p}(w|d)/\hat{p}(w))$ をプロットしたものを図 5(a) に示す. これから, (5) 式を変形した $f(w) \propto \log(p(w)/G_0(w))$ はほぼ正規分布に従っており, ガウス過程でモデル化することが妥当であることがわかる.

ただし, よく見ると, 図 5(b) の Cranfield コーパス [15] での分布は右に裾が長い. これは, 言語には単語が一度出現すると, その後現れやすくなるというバースト性がある [16] ためであり,[2] この影響をモデル化するために, 我々は式 (5) の $p(w)$ の代わりに, 次の DCM (Polya) 分布を用いる.

$$p(\mathbf{w}|\alpha) = \mathrm{DCM}(\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\sum_k \alpha_k + n_k)} \prod_k \frac{\Gamma(\alpha_k + n_k)}{\Gamma(\alpha_k)} \tag{6}$$

ここで, n_k は $\mathbf{w}$ の中での単語 k の頻度である. これは, 次の 2 つの過程

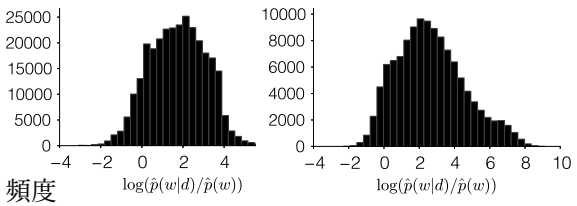
1. Draw $\mathbf{p} \sim \mathrm{Dir}(\alpha)$.
2. For $n = 1 \cdots N$, Draw $w_n \sim \mathbf{p}$.

から $\mathbf{p}$ を積分消去して得られるもので, ステップ 1 でそのテキストに適した多項分布をディリクレ分布から生成しているために, バースト性を表現できることが知られている [17].

これらを通して, CSTM では図 6 に示した生成モデルを仮定する. 図 7 に, CSTM から生成した多項分布と単語の潜在座標 $\phi(w)$ を示した. 図 2 と異なり, CSTM は単語単体のほぼ全域をモデル化できていることがわかる.

4 学習

CSTM では, 単語の潜在座標 $\phi(w)$ および各文書のもつ, ガウス過程に従う関数 f が未知の確率変数である. これはどうやって推定すればよいのだろうか.



(a) Brown (b) Cranfield

図 5: コーパスにおける各文書での単語出現確率 (最尤推定値) の, 平均値との対数尤度比 $\log(\hat{p}(w|d)/\hat{p}(w))$ のプロット.

[2] Cranfield は航空工学の論文抄録で専門用語が多いため, 均衡コーパスである Brown コーパスよりこの傾向が強い.

1. Draw $\alpha_0 \sim \mathrm{Ga}(a_0, b_0)$.
2. Draw $G_0 \sim \mathrm{PY}(\beta, \gamma)$. (実際には最尤推定値)
3. For $w = 1 \cdots W, \phi(w) \sim N(0, I_d)$.
4. For $d = 1 \cdots D$,
   - Draw $f \sim \mathrm{GP}(0, K)$.
   - $\alpha(w) = \alpha_0 G_0(w) e^{f(w)}$ for $w = 1 \cdots W$.
   - Draw $\mathbf{w} \sim \mathrm{DCM}(\alpha)$.

図 6: CSTM の生成モデル.

ここで, f は図 4 のように文書ごとに原理的には無限次元, 実際には語彙次元のベクトルであることに注意されたい. こうした f を直接求めることは難しいが, LDA の枠組でトピック間に相関を与えるために潜在座標を導入する DILN [18] と同様に, 補助変数

$$u \sim \mathrm{N}(0, I_d) \tag{7}$$

を導入する. このとき, $\phi(w)$ をまとめて $\Phi = (\phi(w_1), \phi(w_2), \cdots, \phi(w_V))^T$ とおけば, $f = \Phi u$ の分布は u を積分消去して

$$f|\Phi \sim N(0, \Phi^T \Phi) = N(0, K) \tag{8}$$

となる. これは, f が式 (4) と同じガウス過程に従うことを意味する. よって, 式 (6) の DCM に基づくモデルでは, 式 (5) は α_0 を定数として

$$\alpha(w) = \alpha_0 G_0(w) e^{f(w)} = \alpha_0 G_0(w) e^{\phi(w)^T u} \tag{9}$$

と書けることになり, 観測値 $\mathbf{w}$ から, 潜在変数 α_0, $\phi(w)$, および u を学習する問題となる.

なお (9) 式はベクトル表記すれば, exp を要素ごとに適用するものとして

$$\frac{\alpha}{\alpha_0 G_0} = \exp(\Phi U) \tag{10}$$

とも書け, 潜在行列 Φ と U を更新する, 非線型な NMF に似た問題ともみなせる. またこれは, LSI において語彙空間を使った後付けの説明 [19] を, 潜在空間での確率的生成モデルで置き換えたようなものとみることもできる.

MCMC 法による学習 CSTM の潜在変数, 特に $\phi(w)$ の間には非常に高い相関があるため, 本研究では局所解の問題のない MCMC 法によって学習を行った. このアルゴリズムを図 8 に示す.

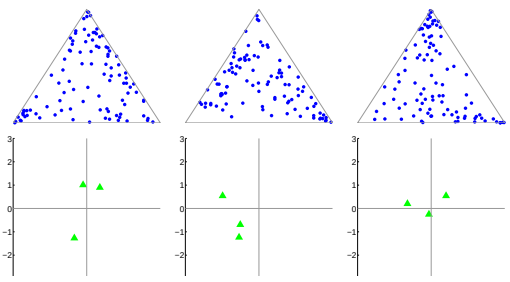


図 7: CSTM からランダムに生成した多項分布 $\mathbf{p}$ と単語の潜在座標 $\phi(w)$. ($\mathbf{p} = \alpha / \sum_w \alpha(w)$)

$\phi(w)$, u_i を $N(0, I_d)$ の乱数で初期化
$\alpha_0 = 1$; /* 任意の初期値 */
for $j = 1 \cdots J$ **do** /* MCMC sweeps */
  for $i = \mathrm{randperm}(1 \cdots D)$ **do** /*文書座標*/
    Draw $u_i' \sim \mathrm{N}(u_i, \sigma^2_{(u)})$
    if MH-accept(u_i') **then**
      $u_i = u_i'$
      $Z_i = \mathrm{Update}(Z_i, u_i)$
    end if
  end for
  for $w = \mathrm{randperm}(1 \cdots W)$ **do** /*単語座標*/
    Draw $\phi(w)' \sim N(\phi(w), \sigma^2_{(\phi)})$
    if MH-accept($\phi(w)'$) **then**
      $\phi(w) = \phi(w)'$
      for $i = 1 \cdots D$ **do**
        $Z_i = \mathrm{Update}(Z_i, \phi(w))$
      end for
    end if
  end for
  $z \sim \mathrm{N}(0, \sigma^2_\alpha)$; $\alpha_0' = \alpha_0 \cdot \exp(z)$ /* α_0 */
  if MH-accept(α_0') **then**
    $\alpha_0 = \alpha_0'$
  end if
end for

図 8: CSTM の MCMC 法による学習アルゴリズム. 実際にはさらに, $\phi(w)$ と u_i の更新を混合してランダムな順番で行っている.

このアルゴリズムは, $\phi(w)$ および u についてランダムな順番で更新する MH 法 [20] であり[3], 関数 MH-accept() ではパラメータの事前分布および式 (6), (9) から得られる尤度を用いて受理を判定する. $\sigma_{(u)}$, $\sigma_{(\phi)}$, $\sigma_{(\alpha)}$ は正規分布のランダムウォーク幅で, 計算効率に影響するが, 本研究では予備実験の結果から $\sigma_{(u)} = 0.01, \sigma_{(\phi)} = 0.02, \sigma_{(\alpha)} = 0.2$ とおいた.

計算効率のため, 各文書ごとに式 (6) で用いる

$$\sum_{w=1}^{W} \alpha(w) = \alpha_0 \sum_{w=1}^{W} G_0(w) e^{\phi(w)^T u_i} \equiv Z_i \tag{11}$$

を保持し, パラメータの変化に従って $\mathrm{Update}(Z_i)$ で更新してゆく.

5 拡張

5.1 語彙共変量

トピックモデルは文書をモデル化するものであるが, 通常は観測値は単語というより, その ID の整数値であり, 自然言語処理というより整数処理に近い. これは, 中間的な「トピック」を持つ LDA では, 文書の持つ情報をトピックに集約せざるを得ないからである. しかし, 提案法は単語の確率を直接変更するもの

[3] 対数尤度の勾配を用いる Hamiltonian MCMC 法 [21] なども試みたが, 計算量が増える一方で結果が不安定になり, ここでは単純なランダムウォーク MH 法の方が優れていた.

であるため, 語彙の情報もモデルに自然に含めることができる.

具体的には, 単語 w の持つ特徴 (共変量) ベクトルを $c(w) = (c_1, \cdots, c_n)$ とおくと, 文書ごとに対応する重み $\zeta_d \sim \mathrm{N}(0, I_n)$ を用いて, (9) 式を次のように変更し, MCMC 法の中で u_d と同様に更新する.

$$\alpha(w) = \alpha_0 G_0(w) e^{\phi(w)^T u_d} e^{c(w)^T \zeta_d} \qquad (12)$$

ζ_d の期待値は 0 なので, 影響がなければ最後の項は $e^0 = 1$ となって無視できることに注意されたい. もし, 文書を通じて現れる単語の多くに共通する特徴 c_n があれば, 上式によりその特徴を持つ単語の確率が $e^{\zeta_{dn} c_n}$ 倍される.

重要な点は, これは意味内容とは独立だということである. 例えば, 単語を漢字またはひらがなで表記する, saxon word ではなく roman word を用いる等の語彙選択は文書内で通常一貫しており, これをモデル化することが可能になる. 文書-単語の関係ではやや分かりにくいが, トピックモデルの重要な応用である協調フィルタリング [22] の文脈で考えてみよう.

この場合, 文書=人, 単語=商品となり, 誰が何の商品を買ったか, という購買データがモデル化される. このとき, 単語=商品が単なる番号で管理されるだけでなく, 「菓子」「電気製品」「高価格帯」「資生堂」「暖色系の色」…などのさまざまな特徴を同時に持つことは明らかであり, この上で, ある人が同じ条件での選択において「寒色系の色」「Sony 製品」を好む, すなわちこれらの特徴を持つ商品の購買確率が高まることが考えられる. CSTM は, こうした傾向もモデル化することができる.

5.2 語彙選択と文書共変量

文書に付与されたラベルや時間といった共変量を利用してトピックモデルを高精度化する研究は無数に存在するが, 2 章で述べたトピックモデルの制約のため, それらはすべて共変量を文書のトピック分布 θ に影響させるものであった. これに対し, CSTM では文書の共変量と単語を直接関連づけることができる.

文書 d の共変量を $c(d)$ とおくと, (9) を

$$\alpha(w) = \alpha_0 G_0(w) e^{\phi(w)^T u_d} e^{c(d)^T \eta_w} \qquad (13)$$

とし, 同様に重み $\eta_w \sim \mathrm{N}(0, I_V)$ を学習する. これにより, 例えば書き言葉では “get” と言いたい場合でも “obtain” が選ばれ, 女性は “nice” より “adorable” の方を好む [23], といった語彙選択を扱うことができる.

なお, (12) 式も同様であるが, この式は対数をとれば

$$\log \alpha(w) = \log \alpha_0 + \log G_0(w) + \phi(w)^T u_d + c(d)^T \eta_w \qquad (14)$$

となり, 共変量の影響を, 切片 $\log \alpha_0, \log G_0(w)$ および説明変数による回帰 $\phi(w)^T u_d$ からの残差として表現する対数線形回帰モデルともいえることに注意されたい. ただし, CSTM では説明変数 $\phi(w)$ 自体も推定すべき潜在変数となっている.

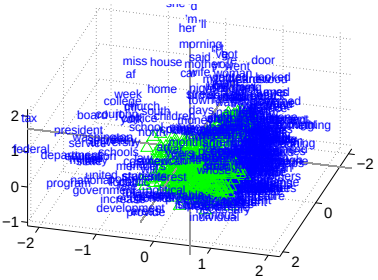


図 9: Brown コーパスで学習された単語と文書の潜在座標. △が文書の座標 u を表す.

6 実験

英語および日本語の標準的なコーパスを用いて実験を行った. 使用したデータは, NIPS[4], KOS ブログテキスト[5], 毎日新聞 (2000 年度から 10,000 記事をランダムに選択), および CSJ 話し言葉コーパス [24] である. データの統計量を以下に示した.

データ	文書数	語彙数	総単語数
NIPS	1,740	13,649	2,301,375
KOS	3,430	6,906	467,714
CSJ	3,302	14,993	6,658,503
毎日新聞	10,000	16,496	2,697,996

6.1 予測パープレキシティ

提案手法と RSM [5], LDA, および (6) 式の混合モデルである Smoothed DM (SDM) [25] におけるパープレキシティを計算した. [26] にならい, データの各文書のランダムな 80%の単語頻度を用いてモデルを計算し, 残りの 20%の単語の予測確率を計算する. この幾何平均の逆数がパープレキシティとなり, 高性能なモデルほど小さい値となる. こうした予測は, 協調フィルタリングやリンク解析の際に実際に文書モデルが使われる状況を模したものともなっている.

表 1 にパープレキシティの結果を示す. CSTM は次元数を 10,20,$\cdots$,50 まで, LDA と SDM はトピック数を 10,20,50,100 と変化させた. RSM には後に述べるような多数のハイパーパラメータがある. これらのうち, 最高の性能となったものを示した.

CSTM は LDA および SDM と比べ, 常に高い性能を見せることがわかる. 最高性能では RSM にやや及ばないが, RSM は最適化が難しいため, 平均的には非常に悪い性能 (NIPS データで 2736.74) であることに注意されたい. 図 10 に様々な最適化パラメータの組み合わせに対する, RSM の NIPS データにおけるテストセットパープレキシティを示す. これは潜在次元数 $\in \{20, 50, 100\}$, 学習率 $\in \{0.05, 0.01, 0.005, 0.001\}$, ミニバッチのサイズ $\in \{100, 50, 20\}$, Contrastive divergence の繰り返し数 $\in \{1, 3, 5\}$ の合計 108 個のパ

[4] http://www.cs.nyu.edu/~roweis/data.html
[5] http://archive.ics.uci.edu/ml/datasets/Bag+of+Words

データ	CSTM	RSM	SDM	LDA
NIPS	1383.66	1290.74	1638.94	1648.3
KOS	1632.35	1396.61	1936.25	1730.7
毎日新聞	466.83	622.69	582.37	507.39

表 1: 各データでのテストセットパープレキシティ. ここでは最高値のみを示した. RSM の性能については, 図 10 を参照のこと.

ラメータの組み合わせに対し, それぞれ 400 回の繰り返しで計算した結果である. この中で CSTM の性能を超えたのはわずか 4 通りであり, これは事前には予測できない.[6] RBM の取扱いの難しさが見てとれる.

6.2 潜在座標

図 11 に, NIPS データを $d=20$ 次元で学習した場合の単語の潜在座標を示す. 単語は分散表現されているが, 次元毎に近い意味の単語がまとまっていることがわかる. (‘image’,‘images’),(‘state’,‘states’) のような複数形が, 自動的にほぼ同じ座標となっていることに注意されたい. (9) 式の形から, 潜在座標は文書間に偏りのない, 無情報な単語 (機能語など) ほど原点近くに位置し, 意味の強い単語 (専門用語など) ほど原点から遠い場所に位置することとなる.

なお, CSTM は文書の潜在座標も補助変数 u_d として計算するため, 図 9 に示すように, 単語と同様に文書も可視化することができる. ここでは上と異なり, $d=3$ 次元に縮約して学習を行った. 従来の PLSV[27] のような手法と異なり, 「トピック」という中間単位を仮定しないため, 文書の座標を単語と直接関連して求められるのが特徴といえる.

6.3 単語の女性らしさ, 男性らしさ

語彙選択の典型的な実例として, 5.2 節で述べたように, 女性と男性の言葉遣いの違いが挙げられる. 最近になって, こうした差も計算言語学の俎上に乗るようになったが [28][29], これらは通常のトピックモデルを基にしているために, トピック選択への影響という

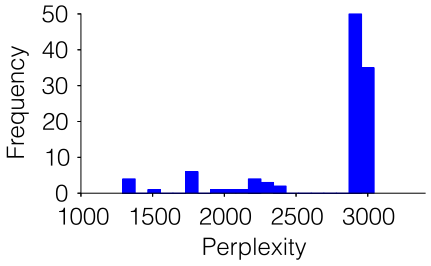


図 10: 様々な最適化パラメータの組み合わせによる, RSM の NIPS データにおけるテストセットパープレキシティ. 性能が CSTM (1383.66) を超えたのは, このうち 3.7%だけであった. この計算には Xeon 2.4GHz の計算機で 2 週間を要している.

[6]学習データのパープレキシティが低くとも, RBM は簡単にオーバーフィットするため, 実際の性能を示す指標とはならない.

上位語		下位語	
e^{ζ}	単語	e^{ζ}	単語
5.189397	会える	0.022626	長尾
4.789601	おとなしい	0.022603	求まる
4.734041	混ぜ合わせ	0.022252	パープレキシティー
4.653134	いらし	0.021917	与党
4.575240	敷き	0.021602	公明
4.490396	っぽく	0.021542	サーチ
4.379417	嫁い	0.021403	速報
4.363100	開ける	0.021295	ディフェンダー
4.287748	寄せ	0.021107	データー量
4.258699	出掛ける	0.020954	徳島
4.152641	美しく	0.020551	ビションフリーゼ
4.137955	大ヒット	0.020187	トルシエ
4.089389	乾い	0.019836	晴郎
3.993580	過ごせる	0.019825	共振
3.985330	治ら	0.019105	鯉のぼり
3.976584	こんにち	0.017528	独占
3.965575	味付け	0.017237	仕様
3.912359	かわいらしい	0.016810	カバレージ
3.903488	かわいかっ	0.016356	真紀子
3.860548	素晴らしかっ	0.016326	芦田
3.829211	素敵	0.014809	しゅ
3.810095	飲み物	0.014667	スラッシュ
3.802053	飲み会	0.014512	湾曲
3.796178	菜園	0.014346	ハイパーリンク
3.761855	敏感	0.012725	韻書

表 2: CSJ コーパスから計算した単語の「女性度」.

面に留まっていた. これに対し, 提案手法では (13) 式により, 文書ラベルに示された性別の, 語彙への影響を直接モデル化することができる.

本研究では, 講演者の性別の付与された CSJ 話し言葉コーパス [24] を用いた. 3302 講演のうち, 1381 講演が女性, 残りの 1921 講演が男性によるものである. 同内容のテーマが指定されているため, 内容の影響を (14) 式により除去した上での語彙選択の効果が計算できる.

表 2 に, 得られた単語確率の修正係数 e^{ζ} の上位および下位語を示した. 話題選択に関係する語も多少はあるが, 女性の選択しやすい/し難い語が品詞をまたいで, 定量的に取り出せていることがわかる.

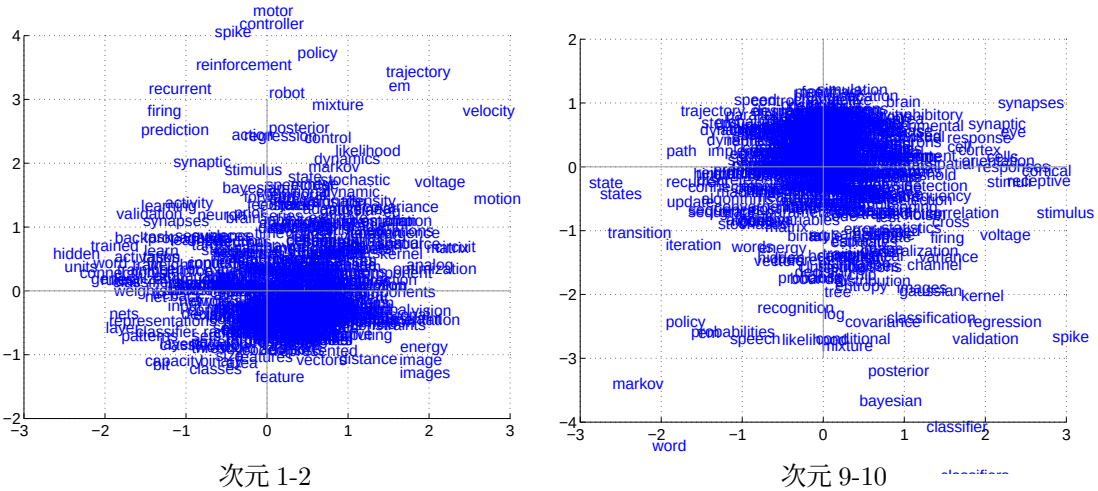
6.4 単語の特徴とトピックモデル

5.1 節で述べた語彙共変量として, 毎日新聞記事の各単語が数字, 漢字, ひらがな, カタカナのいずれかのみで成る場合を 1/0 で符号化し, $c(\text{テーブル})=(0,0,0,1)$ のような共変量の下で式 (12) を最適化する実験を行った. 各文書 d に含まれる語が共通の特徴を持つ場合に, 文書の潜在座標に加えて, 対応する重み

$$\eta_d=(\eta(\text{数字}),\eta(\text{漢字}),\eta(\text{ひらがな}),\eta(\text{カタカナ}))$$

が学習される.

図 12 に, η(カタカナ) および η(ひらがな) の値の高かった文書番号を示す. 実際に文書 2364 はカタカナの人名の多い映画評, 文書 4580 はほぼ全体がひらがなで書かれた絵本のテキストであり, 共通する特徴が教師なしで確かに学習されていた. なお, この際のパープレキシティは 473.27 と, 語彙共変量を用いな



次元 1-2　　次元 9-10

図 11: NIPS データを $d=20$ 次元で学習した場合の単語の潜在座標 $\phi(w)$.

い表 1 の場合と予測性能はほぼ同等であった.

η(カタカナ)

文書	e^η
2364	1.498
4597	1.471
442	1.440
4633	1.433
1520	1.422

文書 2364:

#日公開の映画ではウォンカーヴァイ監督の花様年華かようねんかがカンヌ国際映画祭最優秀男優賞トニーレオン高等技術院賞受賞のかくかくたる戦果をあげての香港凱旋がいせんだあまりにも古風な映画でカーヴァイ監督ファンはびっくりするかも日本映画は連弾がはじけるおもしろさ人間の屑もけっこう見せるデニーロのくせ者ぶりが楽しめるミートザペアレンツ小粒でも ···

η(ひらがな)

文書	e^η
4580	1.720
9961	1.501
5238	1.494
7420	1.470
8375	1.452

文書 4580:

文 小森香折 こもりかおり 絵 広瀬弦 ひろせげん ゆうが押おし入いれをかたづけているとこちらへどうぞという父とうさんの声こえがきこえてきました押し入れのむこうはうらないの部屋へやですゆうは押し入れにもぐりこんで耳みみをあてました女のお客きゃくさんが入はいってきて母かあさんとあいさつしているのがきこえてきますあのうへびがみさまはこれがおすきだとうかがいまして ···

図 12: 単語の特徴に基づく効果の学習. 各文書内で, ある特徴を共有する単語の確率がさらに e^η 倍される.

7　まとめと今後の展望

本論文では, 単語に潜在空間における座標を与え, その上でのガウス過程を考えることで, 混合モデルに基づく通常のトピックモデルより高精度なテキストモデルが得られることを示した. ベイズ学習の際に補助変数を導入することにより, CSTM は潜在層が二値ではなく, 連続なガウス分布に従う RBM ともみなせ, MCMC 法により容易に最適化することができる. 通常の RBM は無向グラフのため学習の意味が不明であり, シグモイド関数のみによって強引な正則化が行われるが, 提案法はガウス過程による生成モデルを与え, 事前分布によって自然に正則化されることが特徴といえる.

CSTM は混合モデルではないため, 潜在次元の選択に HDP [30] のような方法を用いることができない. 次元が大きすぎると過適応が起こるため, 積モデルにおいて次元数決定をどのようにベイズ的に行うかは今後の研究課題である.

自然言語は離散であるが, 本研究では確率の比を考えることで, ガウス分布によりモデル化できることを示した. 基本となるアイデアは基底測度の Exponential tilting であり, 今後これを階層モデルに適用していきたい.

単語に潜在座標を与えることで, 単語間の相関を「トピック」という外生変数に頼らず, 直接与えることができる. 埋め込めるものは単語に限らず, 離散データ一般であり, またその距離の定義も本研究の線形カーネルには留まらない. 空間の文脈適応化も含め, さらなる可能性を追及していきたい.

謝辞

本研究は JST CREST OngaCREST プロジェクト, NII Science 3.0 プロジェクトおよび科研費若手 (B)(課題番号 24700152) の補助を受けて行ったものである.

参考文献

[1] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *SIGIR 1999*, pages 50–57, 1999.
[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. In *Neural Information Processing Systems 14*, 2001.
[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
[4] Peter V. Gehler, Alex D. Holub, and Max Welling. The Rate Adapting Poisson model for Information Retrieval and Object Recognition. In *ICML 2006*, pages 337–344, 2006.
[5] Ruslan Salakhutdinov and Geoffrey Hinton. Replicated Softmax: an Undirected Topic Model. In *NIPS 2009*, pages 1607–1614, 2009.
[6] G. E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14:1771–1800, 2002.
[7] Nitish Srivastava, Ruslan Salakhutdinov, and Geoffrey Hinton. Modeling Documents with a Deep Boltzmann Machine. In *UAI 2013*, 2013.

[8] Mac'Aurelio Ranzato and Martin Szummer. Semi-supervised Learning of Compact Document Representations with Deep Networks. In *ICML 2008*, pages 792–799, 2008.
[9] Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97:1090–1098, 2002.
[10] Purnamrita Sarkar and Andrew Moore. Dynamic Social Network Analysis using Latent Space Models. In *NIPS 2005*, pages 1145–1152, 2005.
[11] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A Latent Variable Model for Geographic Lexical Variation. In *EMNLP 2010*, pages 1277–1287, 2010.
[12] Carl Edward Rasmussen and Christopher K. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
[13] Hemant Ishwaran and Lancelot F. James. Computational Methods for Multiplicative Intensity Models Using Weighted Gamma Processes: Proportional Hazards, Marked Point Processes, and Panel Count Data. *JASA*, 99(465):175–190, 2004.
[14] J. Møller, A. R Syversveen, and R. P. Waagepetersen. Log Gaussian Cox Processes. *Scandinavian Journal of Statistics*, 25:451–482, 1998.
[15] C. J. van Rijsbergen and W. Bruce Croft. Document clustering: An evaluation of some experiments with the cranfield 1400 collection. *Information Processing and Management*, 11(5–7):171–182, 1975. http://ir.dcs.gla.ac.uk/resources/test_collections/cran/.
[16] Kenneth W. Church. Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to $p/2$ than p^2. In *COLING 2000*, pages 173–179, 2000.
[17] Gabriel Doyle and Charles Elkan. Accounting for burstiness in topic models. In *ICML 2009*, pages 281–288, 2009.
[18] John Paisley, Chong Wang, and David M. Blei. The Discrete Infinite Logistic Normal Distribution. *Bayesian Analysis*, 7(2):235–272, 2012.
[19] Hinrich Schütze. Dimensions of Meaning. In *Proceedings of Supercomputing'92*, pages 787–796, 1992.
[20] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall / CRC, 1996.
[21] Radford M. Neal. *Bayesian Learning for Neural Networks*. Number 118 in Lecture Notes in Statistics. Springer-Verlag, 1996.
[22] 神嶌敏弘. 推薦システムのアルゴリズム (1)〜(3). **人工知能学会誌**, 22–23(6–2), 2007–2008.
[23] Robin Lakoff. Language and Woman's Place. *Language in Society*, 2(1):45–80, 1973.
[24] 国立国語研究所. 日本語話し言葉コーパス, 2008. http://www.kokken.go.jp/katsudo/seika/corpus/.
[25] 貞光 九月, 待鳥 裕介, 山本 幹雄. 混合ディリクレ分布パラメータの階層ベイズモデルを用いたスムージング法. **情報処理学会研究報告** *2004-SLP-53*, pages 1–6, 2004.
[26] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation Methods for Topic Models. In *ICML 2009*, pages 1105–1112, 2009.
[27] Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda. Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents. In *KDD 2008*, pages 363–371, 2008.
[28] Adam Vogel and Dan Jurafsky. He said, she said: Gender in the ACL anthology. In *ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, 2012.
[29] Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre. In *CoNLL 2011*, pages 78–86, 2011.
[30] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *JASA*, 101(476):1566–1581, 2006.