ガウス過程に基づく連続空間トピックモデル

持橋大地

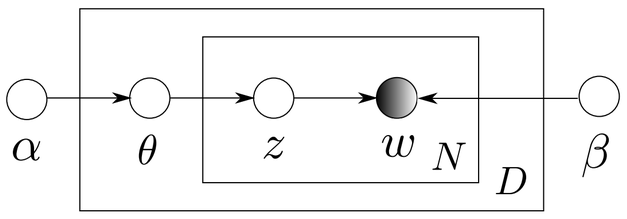
統計数理研究所 数理・推論研究系 学習推論グループ

daichi@ism.ac.jp

1. 背景: トピックモデル

トピックモデル…文書(離散データ)の確率的生成モデル
代表的なもの: Latent Dirichlet Allocation (Blei+ 2003)
生成プロセス: 各文書ごとに,

1. 潜在トピック分布 $\theta \sim \mathrm{Dir}(\alpha)$ を生成
2. 各語ごとに,
   (a) 潜在トピック $z \sim \mathrm{Mult}(\theta)$ を選択.
   (b) 単語 $w \sim p(w|z) = \beta_{zw}$ を生成.



- θ: 多項分布
- β: 潜在トピック別単語確率分布
- w: 観測値(単語)

応用と拡張

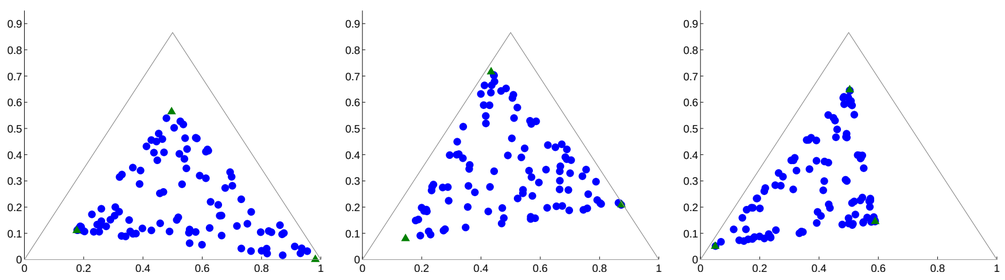
- 非常に多くの分野で使われている
  - グラフィックス, ビジョン, バイオインフォマティクス, 推薦モデル, リンク解析, ネットワークモデル, …
- 多数の改良の試み(sLDA, CTM, DHDP, Mixed Random Measures (Kim+ 2012), dHNRM (Chen+ 2012), …)

2. 問題点

- 混合モデルのため, $\beta = p(w|z)$ は一度決まると固定
  - 語彙の情報を細かく制御できない＝ラベルがトピックにしか影響しない(sLDA, DHNRM等の拡張全て)

実際の例
- 女性特有の言い回し, 語彙選択
- 会話文の文体 (トピック／内容とは独立)
- ひらがな／漢字ばかりの文章
- 著者の語彙的な癖 (-eous や re- が多い, など)

- 単語Simplexの全領域をモデル化できない



⇓

単語Simplexの全領域をモデル化し, 語彙に柔軟に適応する統計モデル.

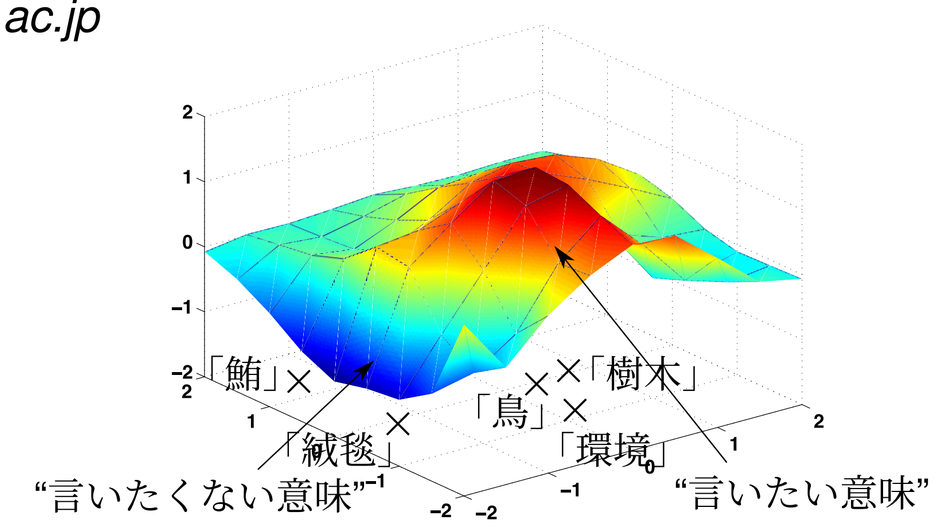
3. CSTM: Continuous Space Topic Model

基本的なアイデア: **潜在単語空間**上のガウス過程 $f \sim \mathrm{GP}(0, K)$ によって,「この文書で言いたいこと」を表現: 単語確率を, 下の積でModify.

$$p(w|d) = e^f G_0(w)$$

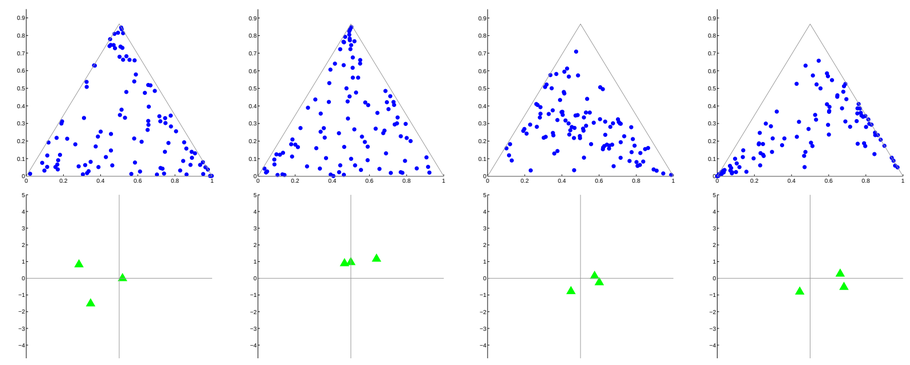
生成モデル

1. For $w = 1 \cdots W$, $\phi(w) \sim \mathrm{N}(0, I_d)$. ($w$の潜在座標)
2. for $d = 1 \cdots D$
   - $f \sim \mathrm{GP}(0, K)$; $K(v, w) = \phi(v)^T \phi(w)$
   - $\alpha(w|d) = \alpha_0 G_0(w) \exp(f)$
   - $G \sim \mathrm{DP}(\alpha(\cdot|d))$, $\mathbf{w} \overset{\mathrm{iid}}{\sim} G$. (Polya分布)



特徴:

- 語彙の確率を直接制御できる
  - 拡張: 文書ラベル $\alpha(w|d) = \alpha_0 G_0(w) e^f e^{\theta^T c(d)}$
  - 拡張: 単語の特徴 $\alpha(w|d) = \alpha_0 G_0(w) e^f e^{\eta^T c(w)}$
- Product model ($\leftrightarrow$ Mixture model)
- 単語Simplexのほぼ全域をモデル化できる



4. 計算と推論

GP f は直接表現が難しい→補助変数 $u \sim \mathrm{N}(0, I_d)$ のとき, $f = \Phi u$ の分布は u を消去して

$$f|\Phi \sim N(0, \Phi^T \Phi) = N(0, K).$$

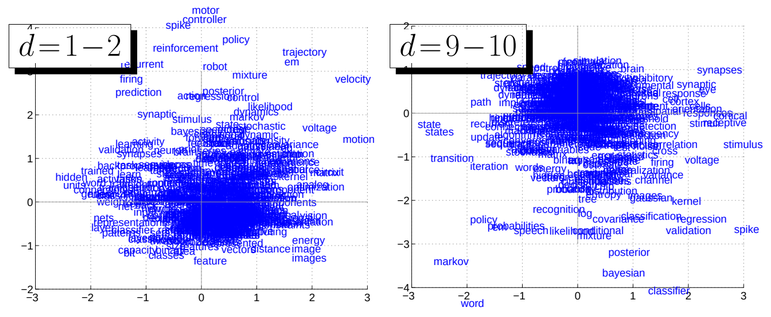
よって, 文書d毎に $f = \Phi u_d$ として u_d とΦをMCMCで更新. $\alpha_0 \sim \mathrm{Ga}(a_0, b_0)$, $\eta_{kv} \sim \mathrm{N}(0, I_d)$, $\theta_{cv} \sim \mathrm{Exp}(\beta)$.

5. 実験結果

Perplexity

データ	CSTM	CSTM(+lex)	LDA
NIPS	1410.96	–	1648.3
KOS	1632.35	–	1730.7
毎日新聞	473.954	473.273	507.3855

Latent Embedding (NIPS, $d=10$)



Covariate modeling

η(カタカナ)

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

テキスト2364:

#日公開の映画ではウォンカーウ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ウ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

テキスト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…

Future Work

Random Poisson intensityとして, HDP/HPYの拡張→ PCFG, nグラムモデル等の潜在空間埋め込み拡張