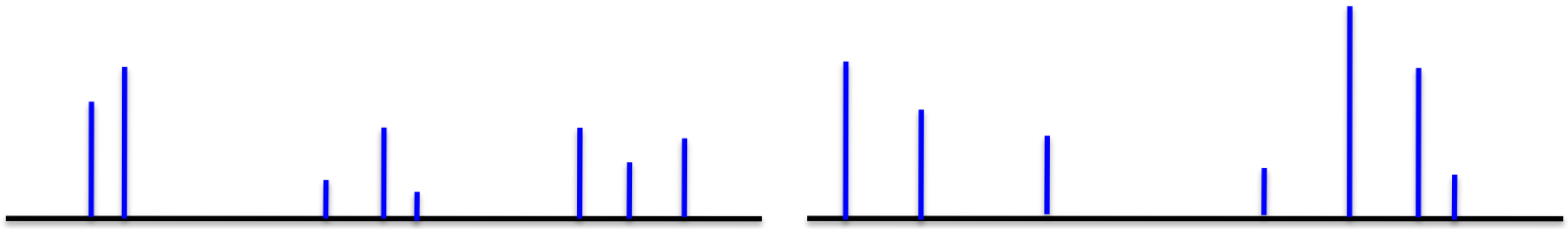
ガウス過程に基づく連続空間トピックモデル

持橋大地 (統数研), 吉井和佳, 後藤真孝 (産総研)
daichi@ism.ac.jp

IPSJ SIGNL-213
2013-9-13(Fri), 山梨大学

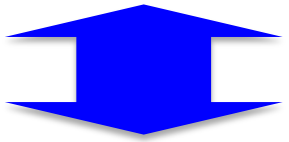
動機: 高次元多項分布の推定問題



- 統計的自然言語処理のあらゆる局面に現れる
  - 構文解析 (PCFG)
  - 文書モデル
  - HMM、n-gram、....
- 出力が単語＝通常、10,000次元以上
  (Google 1T n-gram: 13,588,391次元)

通常の解決: 混合モデル

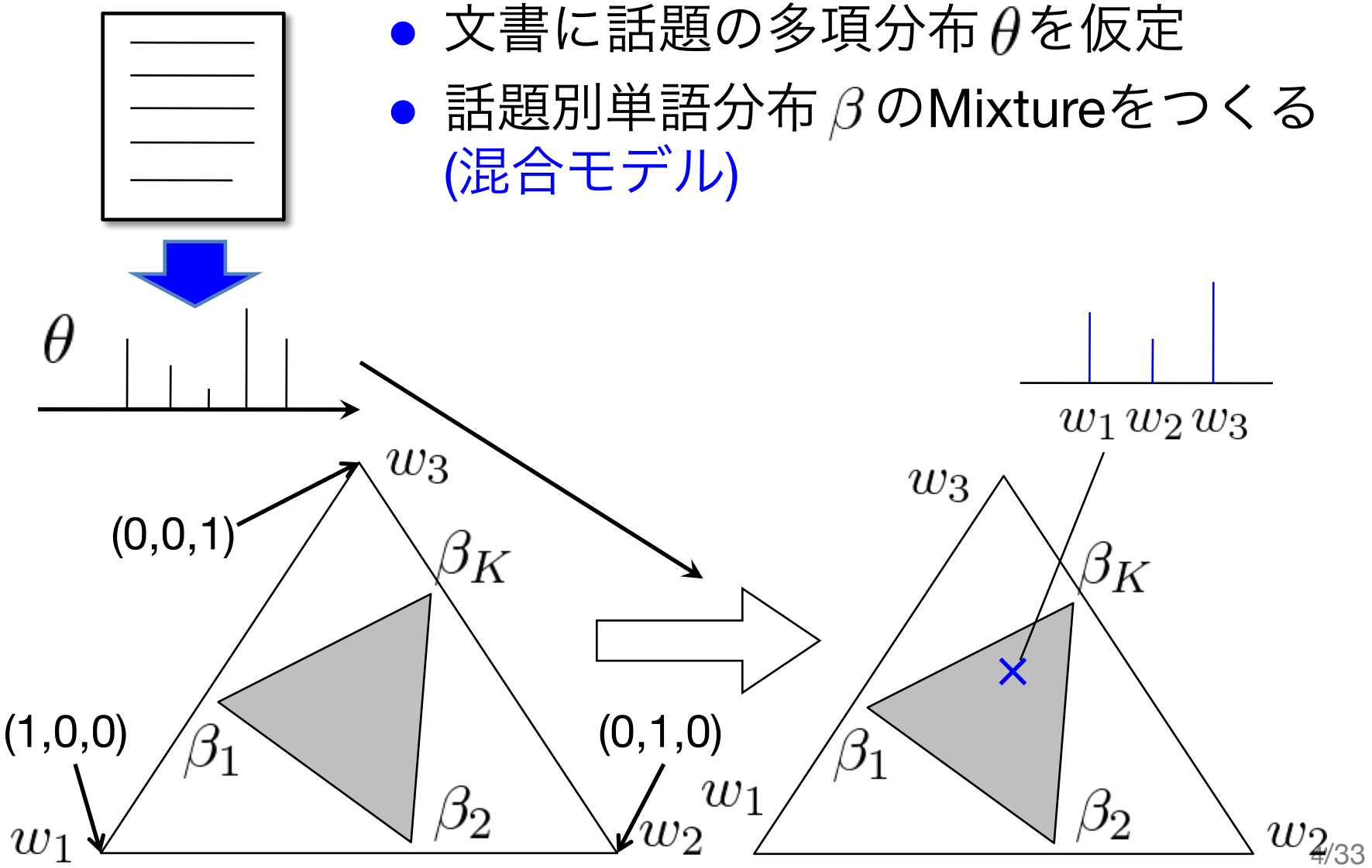
- LDA (Blei+01), DM (Sjölander96, 山本 03)など
- LDAのモデル:
  - 各文書について、トピック分布 $\theta \sim \mathrm{Dir}(\alpha)$ をサンプル
    - For *i*=1.. *N*,
      1. 潜在トピック $z \sim \mathrm{Mult}(\theta)$ をサンプル
      2. 単語 $w \sim p(w|z)$ を生成



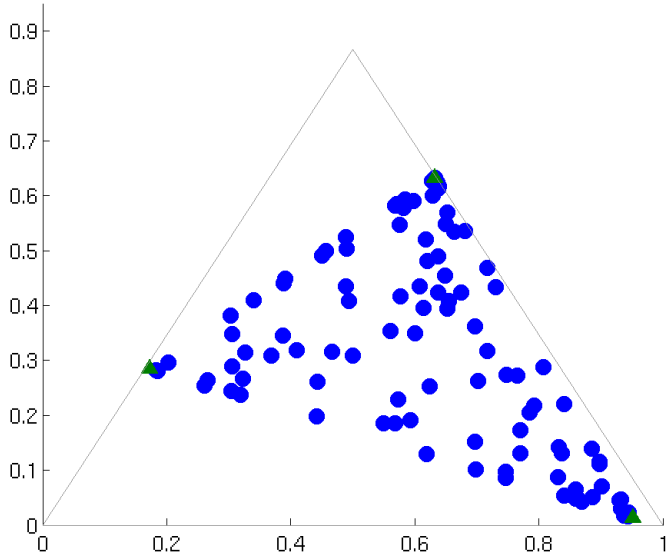
      1. 単語 $w \sim \sum_{k=1}^{K} \theta_k p(w|k)$ を生成.

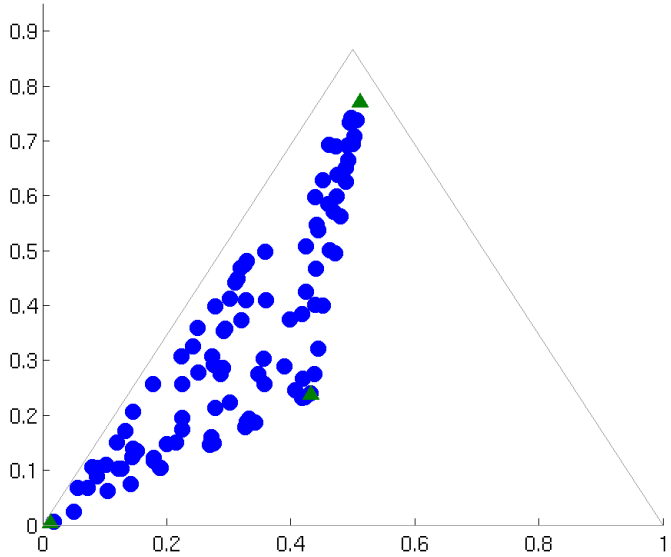
LDAのモデル (NMFでも共通)

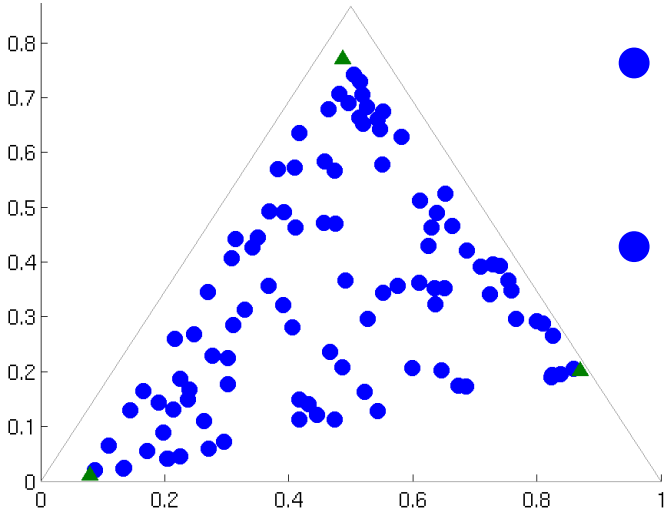
- 文書に話題の多項分布 θ を仮定
- 話題別単語分布 β のMixtureをつくる (混合モデル)



LDAから生成される単語確率分布







- 単語単体の一部しかモデル化できない!!
- 混合しているだけだから

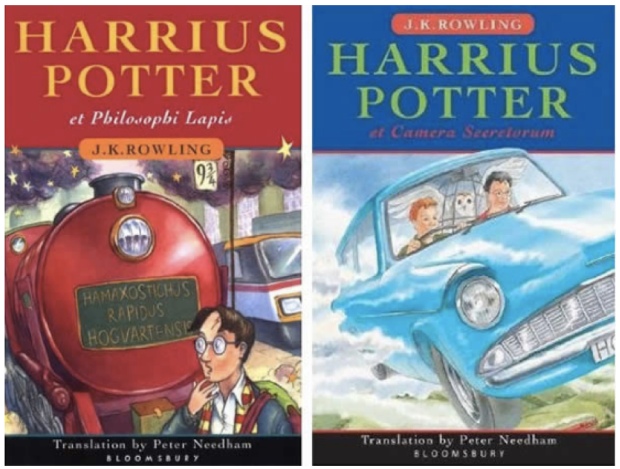
何が問題なのか?

- LDAのGibbsサンプラーの更新式:

$$p(z_{dn}=k|W,Z_{-dn}) \propto \frac{\alpha_k + n(d,k)}{\sum_k \alpha_k + n(d,k)} \cdot \frac{\eta + n(w_{dn},k)}{\sum_w \eta + n(w,k)}$$

  - 各単語は1つのクラスタにしか属さない → 本当?

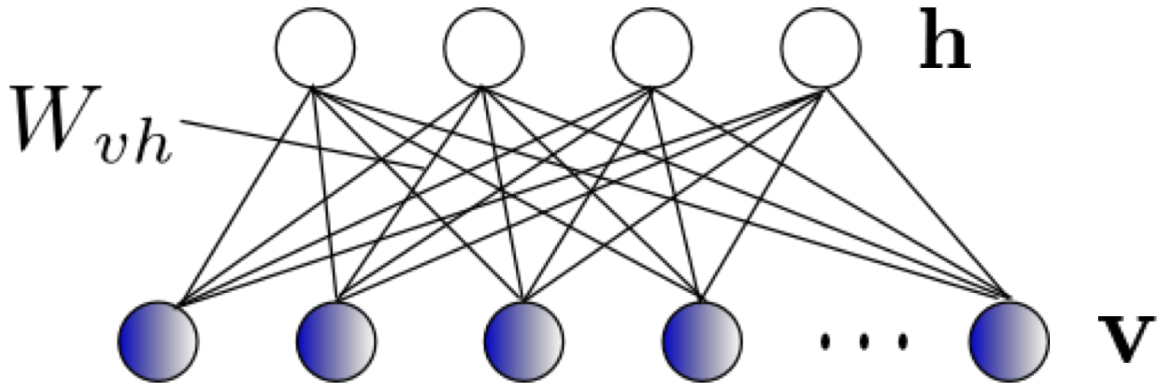
- 文書＝人、単語＝商品と考えてみる
(協調フィルタリング)



さまざまな属性:
小説／本／若者向け／挿絵あり／ラテン語／....

単なるクラスタリングでは表現できない！

Restricted Boltzmann Machines

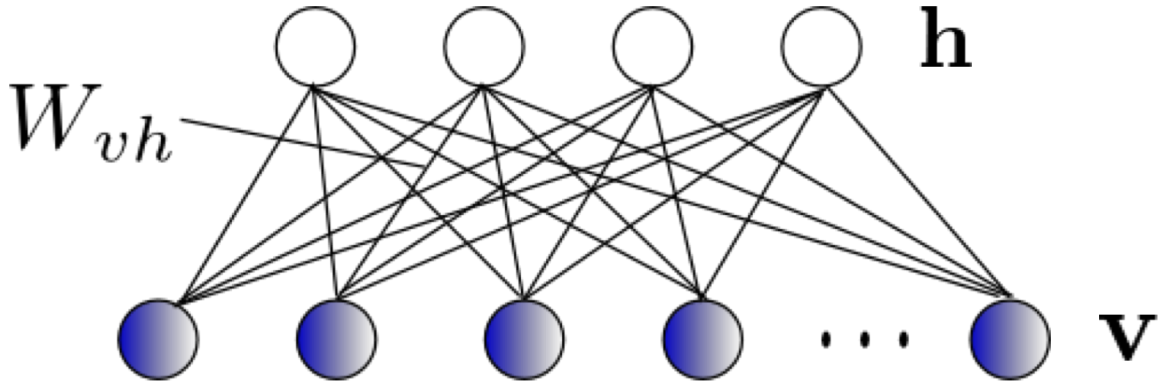


- “Deep Learning”の最も基本的なモデル
  - 出力層 *v* と0/1の潜在層 *h* が重み *W* で結ばれたニューラルネット
- 混合モデルではなく、積モデル (Product of Experts)

Hinton (2002)

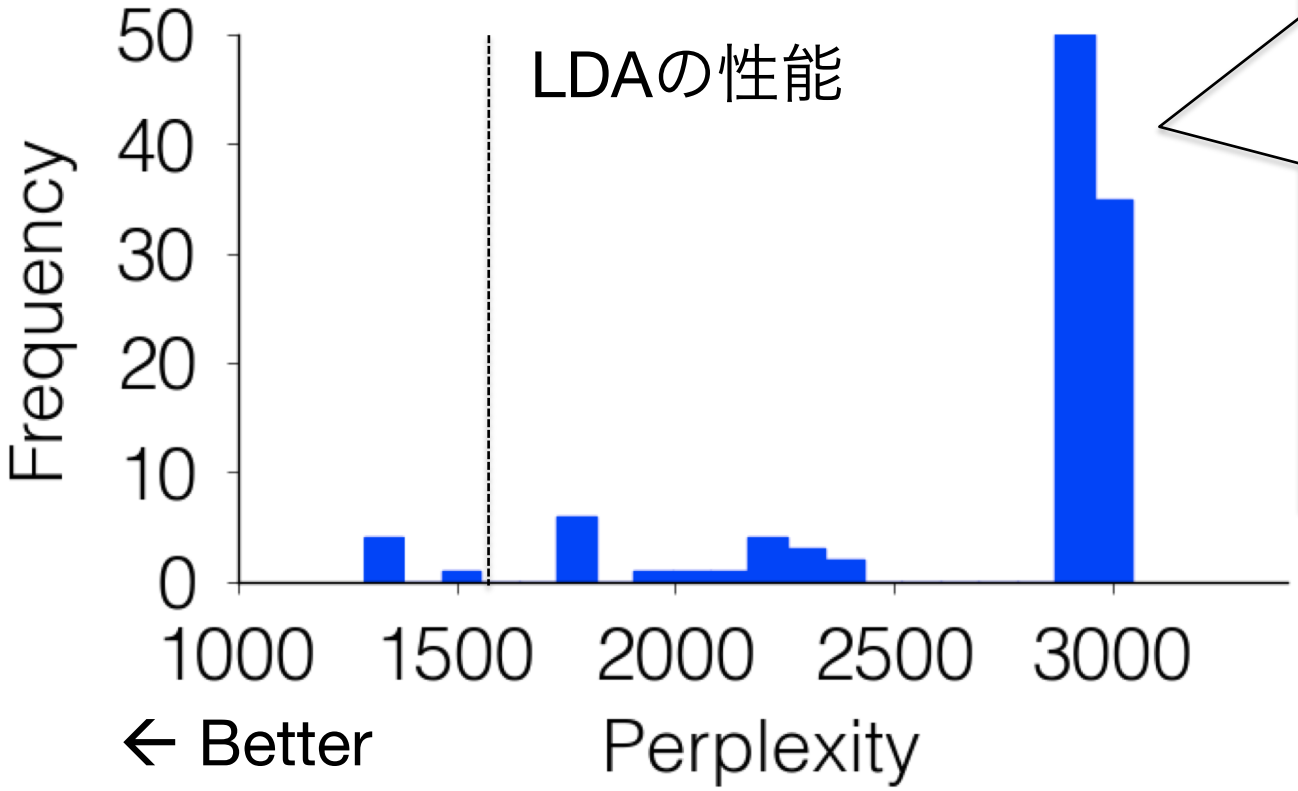
$$p(\mathbf{v},\mathbf{h}) = \frac{\exp(\mathbf{v}^T W \mathbf{h})}{\sum_{\mathbf{v}}\sum_{\mathbf{h}} \exp(\mathbf{v}^T W \mathbf{h})} \propto \prod_i \prod_j e^{W_{ij} v_i h_j}$$

Restricted Boltzmann Machines (2)



- LDAと異なり、意味を分散表現できる
  - 国際経済＝“国際”×“経済”
  - 海外サッカー＝“国際”×“サッカー”
  - 自然言語処理＝ “数学”×“言語学” ....
- しかし、

RBMの最適化の難しさ

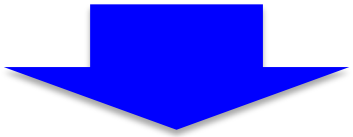


Replicated Softmax (Salakhut-dinov 2009) のNIPSコーパスでの実験結果

- RBMには、学習率、ミニバッチサイズ、モーメント、CD iterations、..などの多数のメタパラメータ
- ほとんどの場合、非常に悪い性能しか出ない

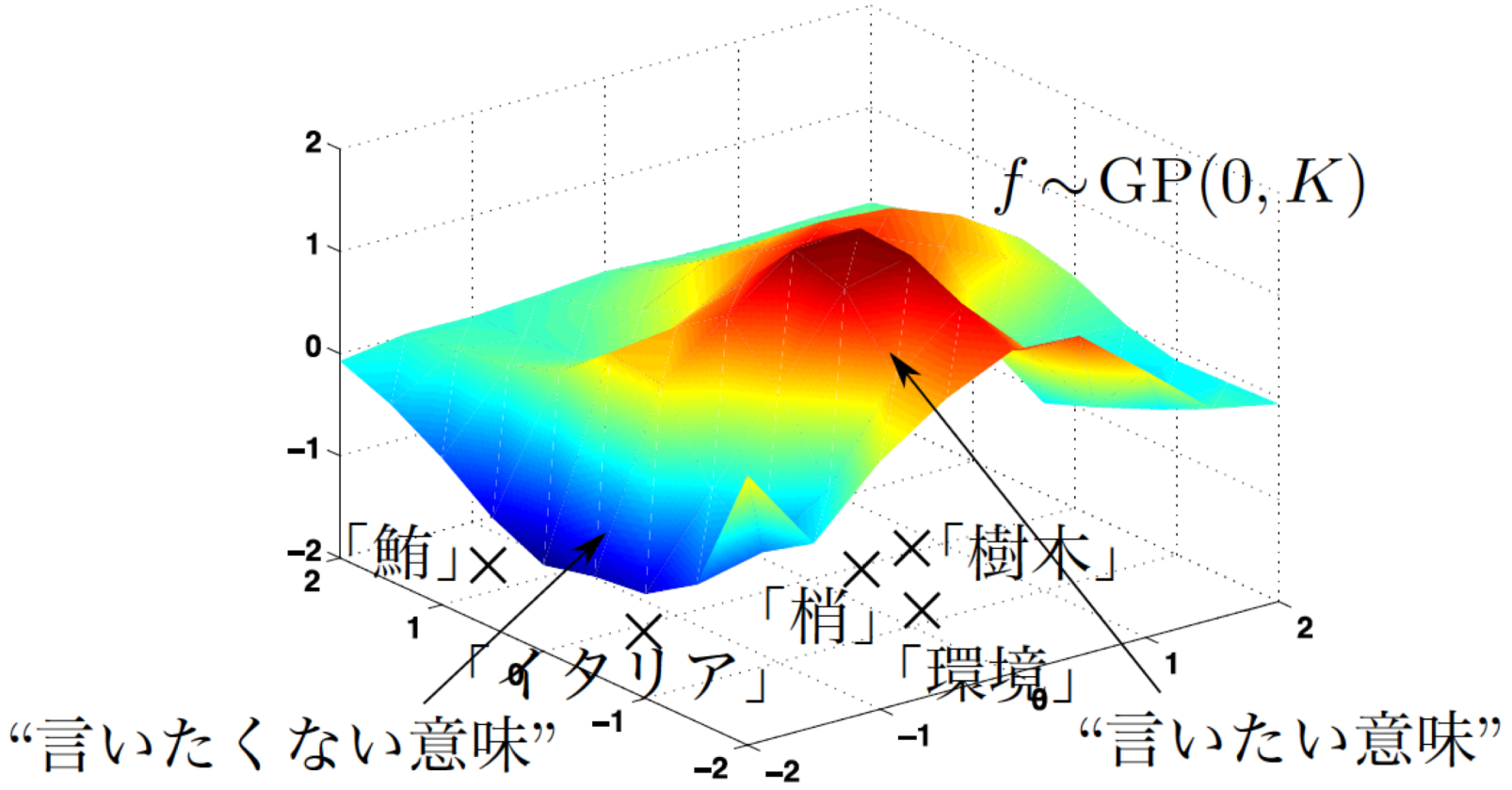
何が問題か?

- RBMは生成モデルがなく、0/1の潜在変数と
  シグモイド関数で強引に正則化している
- RBM, LDAとも、語彙の情報が非常に重要
  - RBM: ニューラルネットの重み W_{vk}
  - LDA: 単語のトピック分布 $p(z|w) \propto p(w|z)p(z)$



- 単語に潜在座標を明示的に与えるモデル.
  - 実は、統計学では Latent space models (Hoff 2002)
    として知られている (社会ネットワーク解析)

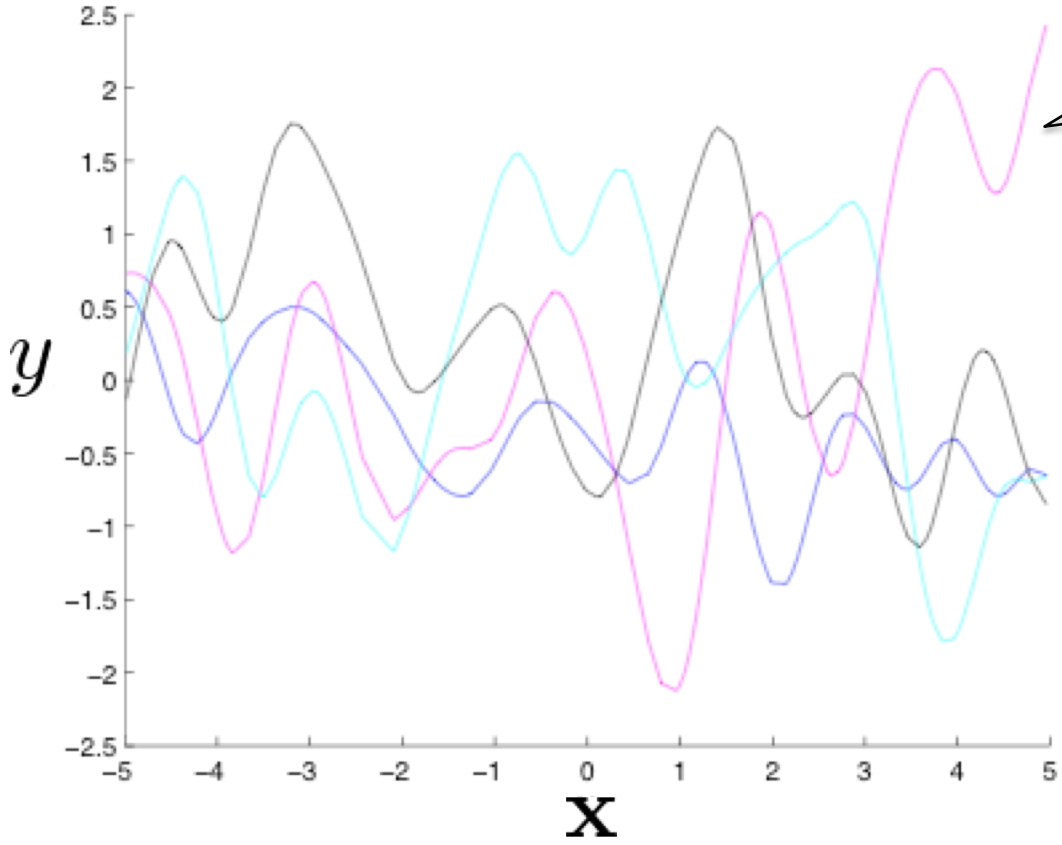
CSTM: Continuous space topic models



- 単語wはd次元の潜在座標 $\phi(w) \sim \mathrm{N}(0, I_d)$ をもつ
- この上に、ガウス過程 $f \sim \mathrm{GP}(0, K)$ を生成

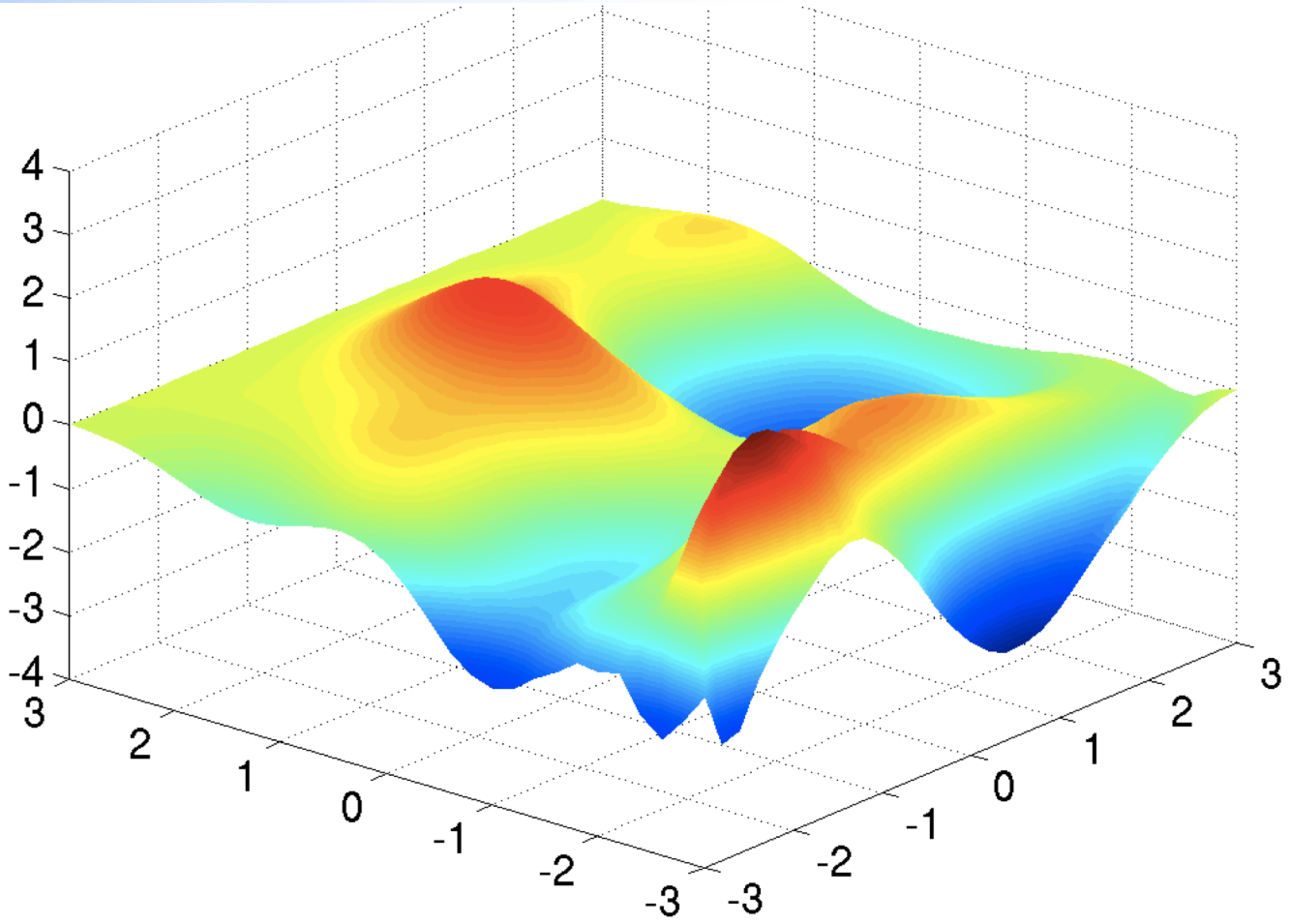
Gaussian process とは

- ガウス過程: $\mathbf{x} \mapsto y$ への回帰関数を生成する確率分布
  - 実際には、無限次元のガウス分布



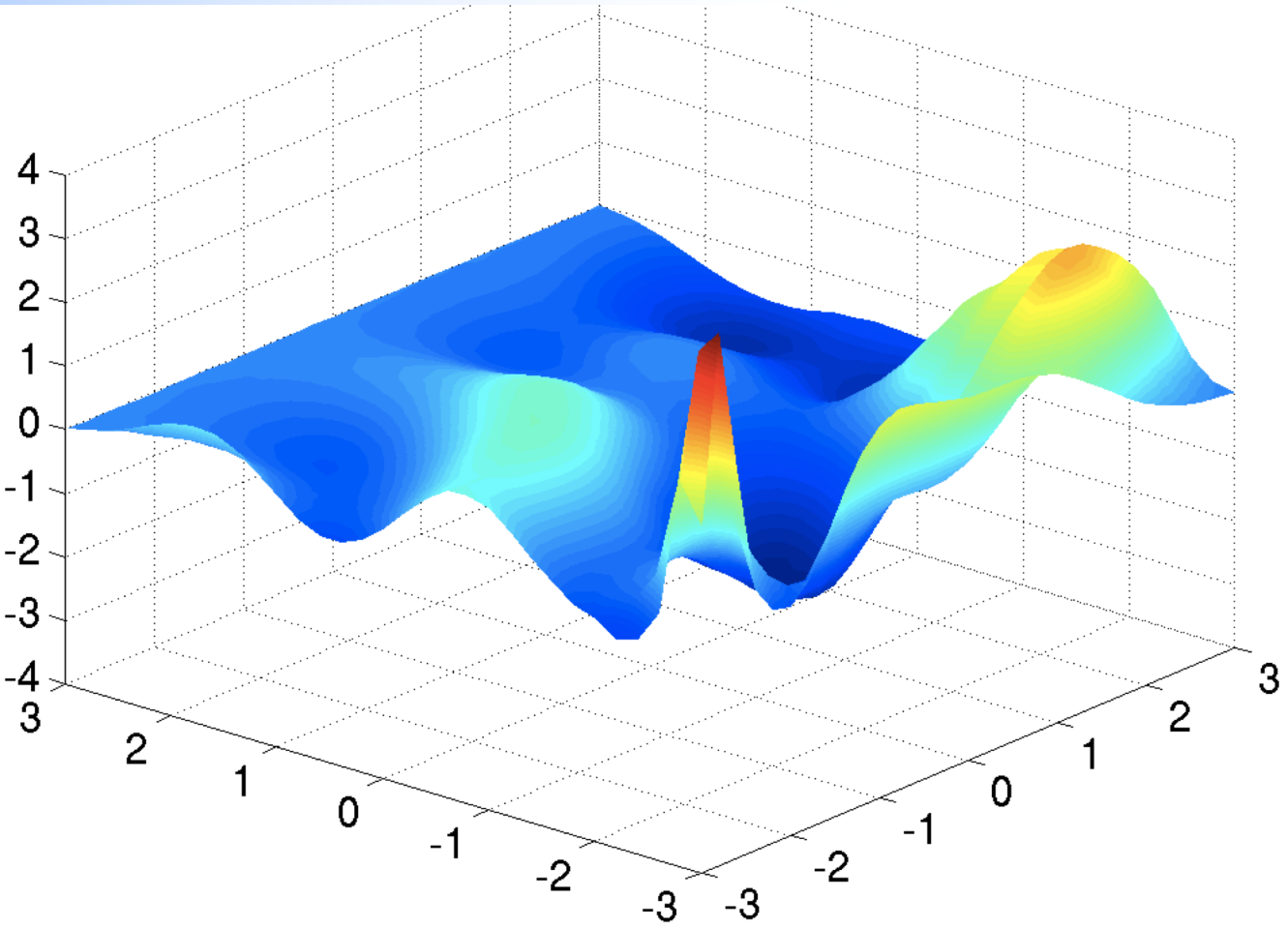
1次元の場合

Gaussian processとは (2)



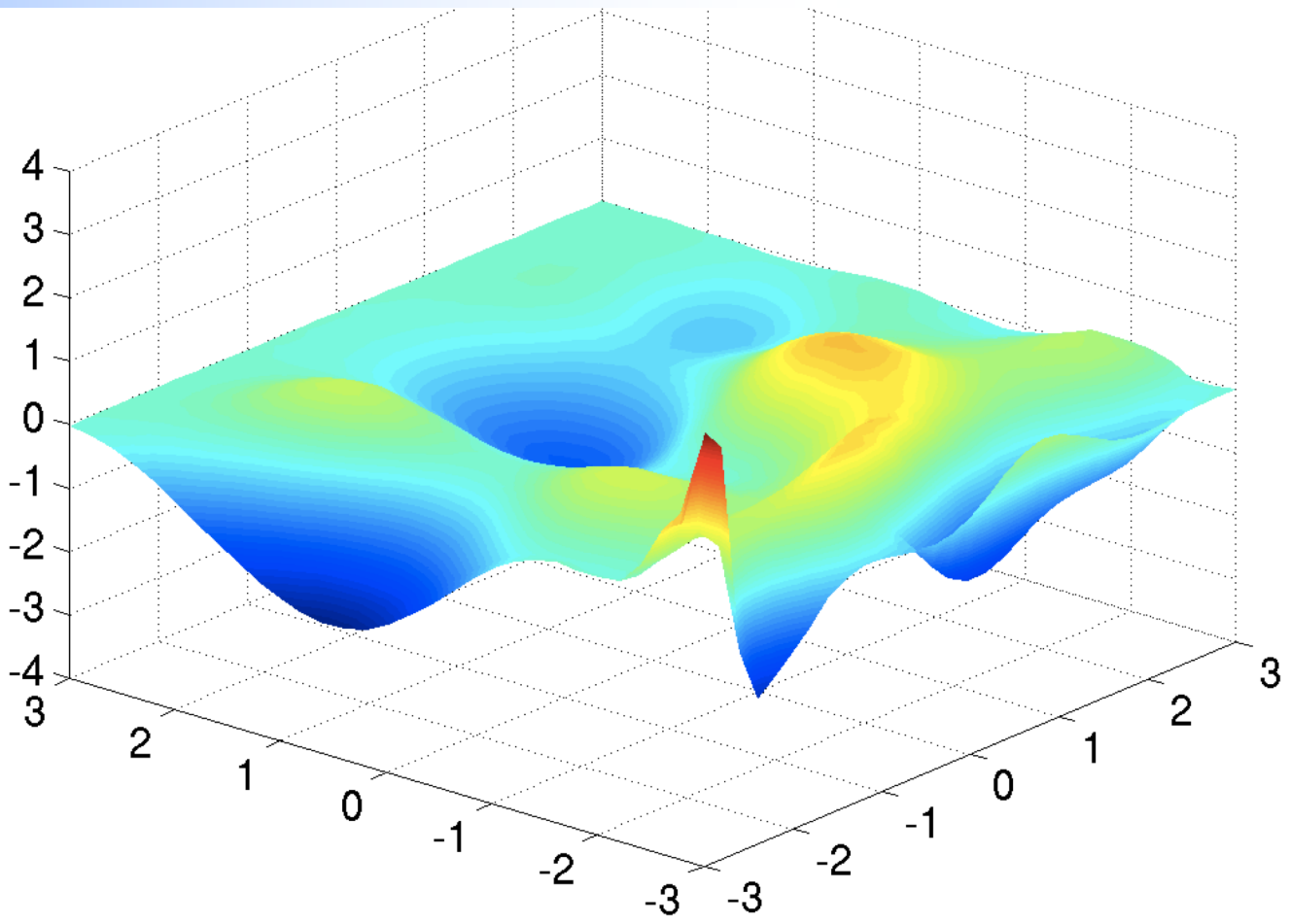
- 2次元の場合

Gaussian processとは (2)



- 2次元の場合

Gaussian processとは (3)



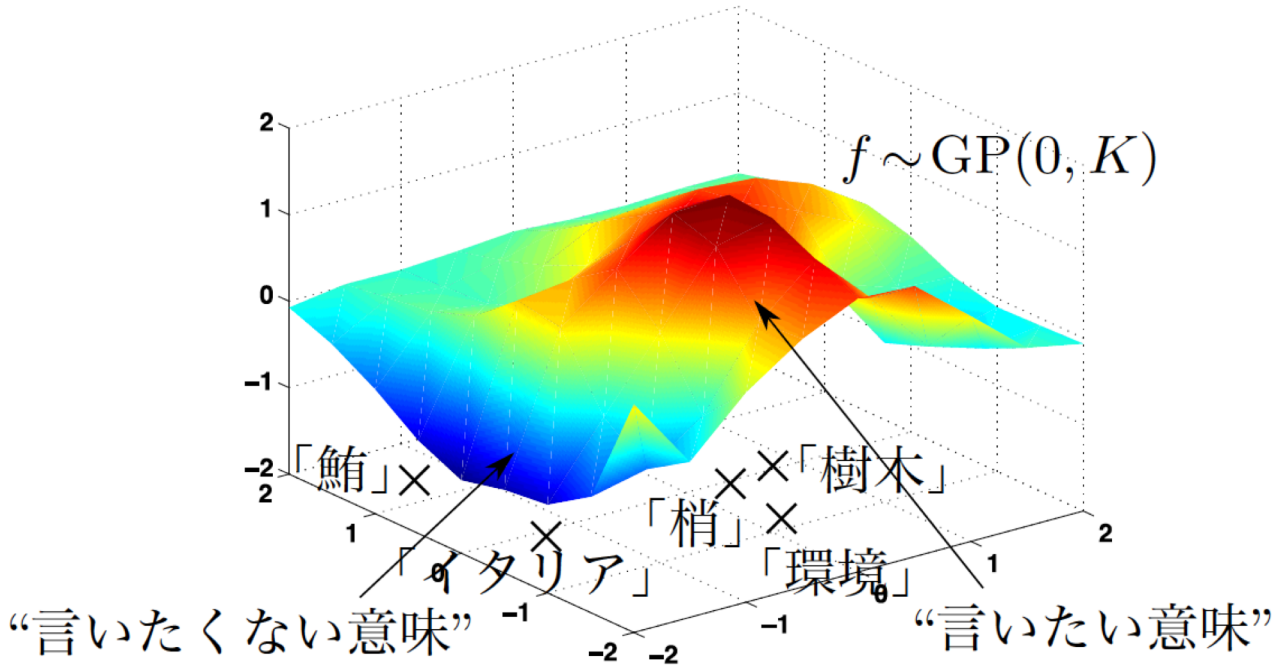
- 2次元の場合

CSTM: 最初のモデル

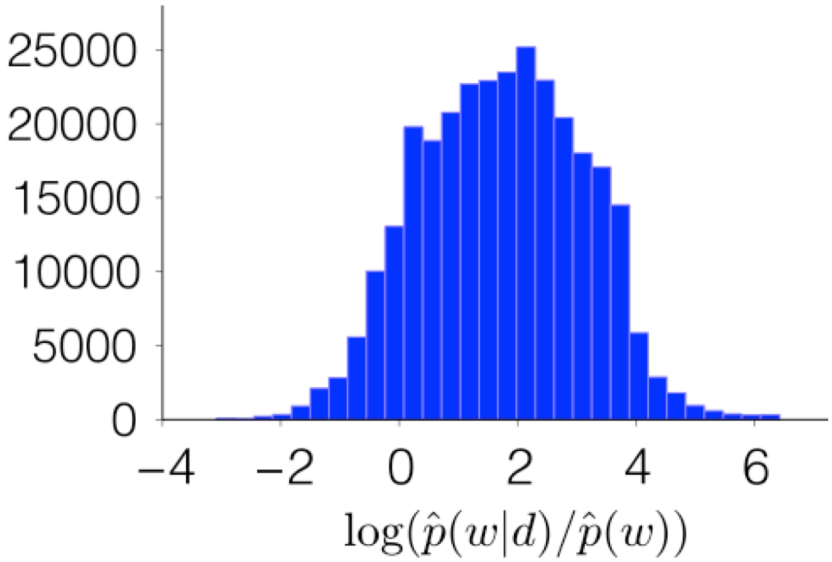
- 単語の平均的な確率(最尤推定)$G_0(w)$を、ガウス過程$f(w)$で Modulate

$$p(w|d) \propto e^{f(w)} G_0(w) = \frac{e^{f(w)} G_0(w)}{\sum_w e^{f(w)} G_0(w)}$$

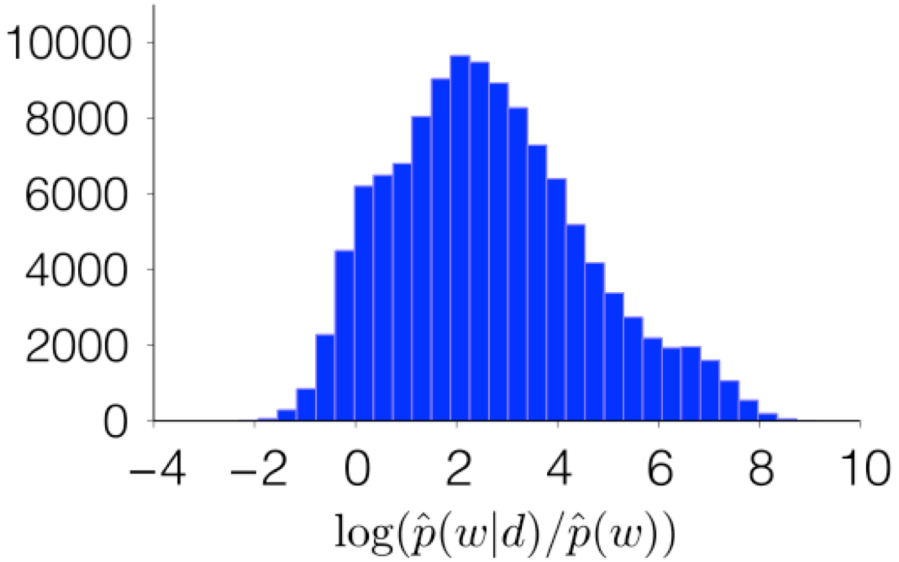
  - $e^{f(w)}$ は、8000倍から0.0001倍くらいの値



Empirical Evidence



Brownコーパス



Cranfield コーパス

- $p(w|d) \propto e^{f(w)} p(w) \Longleftrightarrow f(w) \propto \log\left(\frac{p(w|d)}{p(w)}\right)$ を最尤推定で計算してプロット
  - $\hat{p}(w|d) = n(w,d)/\sum_w n(w,d),\ \hat{p}(w) = n(w)/\sum_w n(w)$
- 確率の比はほぼGaussianで分布している!

Polya分布による拡張

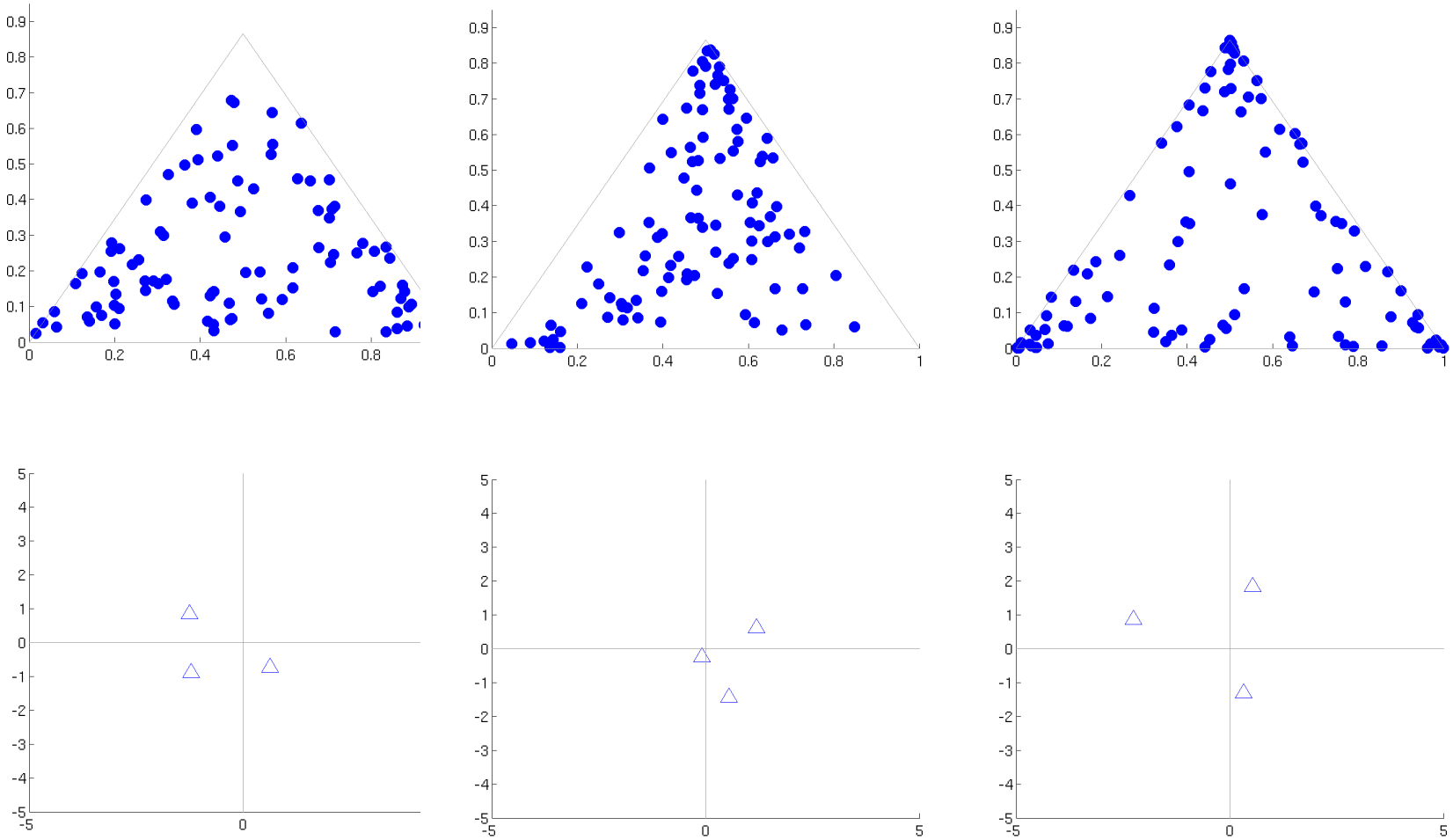
- 言語にはバースト性がある→Polya (DCM)分布

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

  - Draw $\mathbf{p} \sim \mathrm{Dir}(\alpha)$
  - For n=1..N, Draw $w_n \sim \mathbf{p}$.
- $\alpha = (\alpha(w_1), \cdots, \alpha(w_V))$ を文書ごとに下で生成
  - Draw $f \sim \mathrm{GP}(0, K)$
  - Set $\alpha(w) = \alpha_0 G_0(w) e^{f(w)}$
  - Draw $\mathbf{w} \sim \mathrm{DCM}(\alpha)$.

$$\alpha_0 \sim \mathrm{Ga}(a_0, b_0)$$

CSTMから生成される単語確率分布



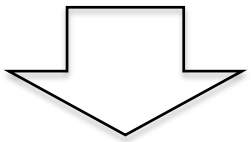
- ほぼ全単語Simplexを網羅 (和が1の制約がない)

学習

- ガウス過程から生成した関数fは文書ごとに無限次元
  → 学習不可能
- DILN (Paisley+ 2012)と同様に、補助変数uを導入
  - 単語座標の行列を $\Phi = (\phi(w_1), \cdots, \phi(w_V))$とする
  - $u \sim \mathrm{N}(0, I_d)$ のとき、$f = \Phi u$ はuを積分消去して

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

  - これは、線形カーネル $k(w_i, w_j) = \phi(w_i)^T \phi(w_j)$ を使ったGPと等価なことを意味する



- $\alpha(w) = \alpha_0 G_0(w) e^{u^T \phi(w)}$として、$u$ と $\phi(w)$の学習問題!

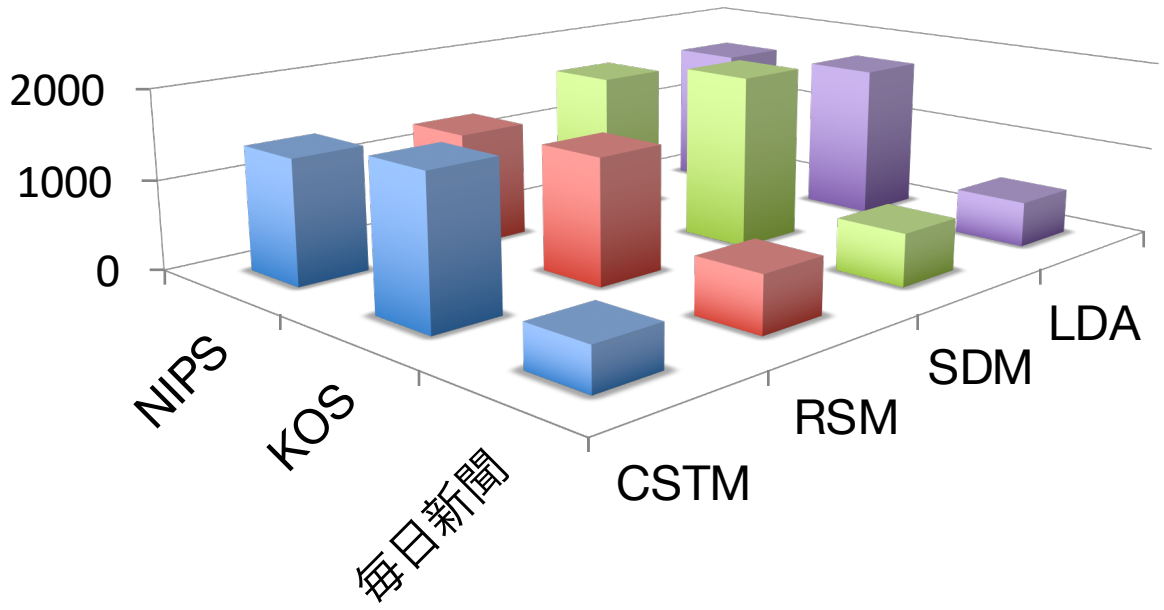
学習 (2)

- 通常のMetropolis-Hastingsで、単語と文書の潜在座標を学習
  - For j = 1 .. J,
    - for i = randperm(1 .. D),
      - Draw u' ~ N(u,σ^2) & MH-accept(u'); Update Z
    - For w = randperm(1 .. W),
      - Draw ϕ'(w) ~ N(ϕ(w),σ^2) & MH-accept(u'); Update Z_1..Z_N
    - z ~ N(0,σ^2); α_0' = α_0・exp(z)
      - If MH-accept(α_0') then α_0=α_0'
  - 実際は、uとϕ(w)の更新をランダムに混合
- 単語間に強い相関があるため、勾配法では局所解

実験

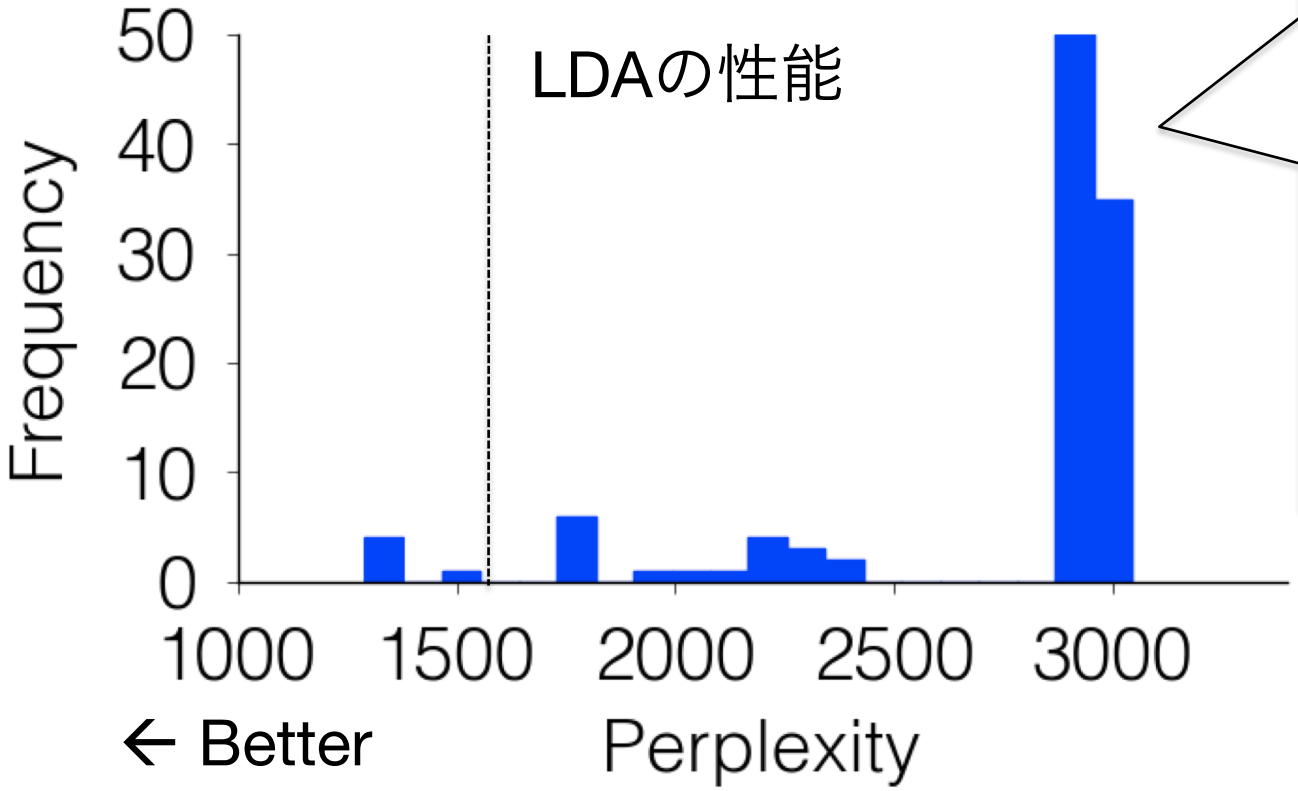
- NIPS, KOS, CSJ話し言葉コーパス、毎日新聞テキストで実験
  - NIPS: 1740文書, V=13649
  - KOS: 3430文書, V=6906
  - CSJ : 3302文書, V=14993
  - 毎日新聞: 10000文書, V=16496 (2000年度からランダムに選択)
- 各文書のランダムな80%の単語でモデルを計算、残りの20%の単語を予測

実験結果 (予測パープレキシティ)



	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

RBMの最適化の難しさ

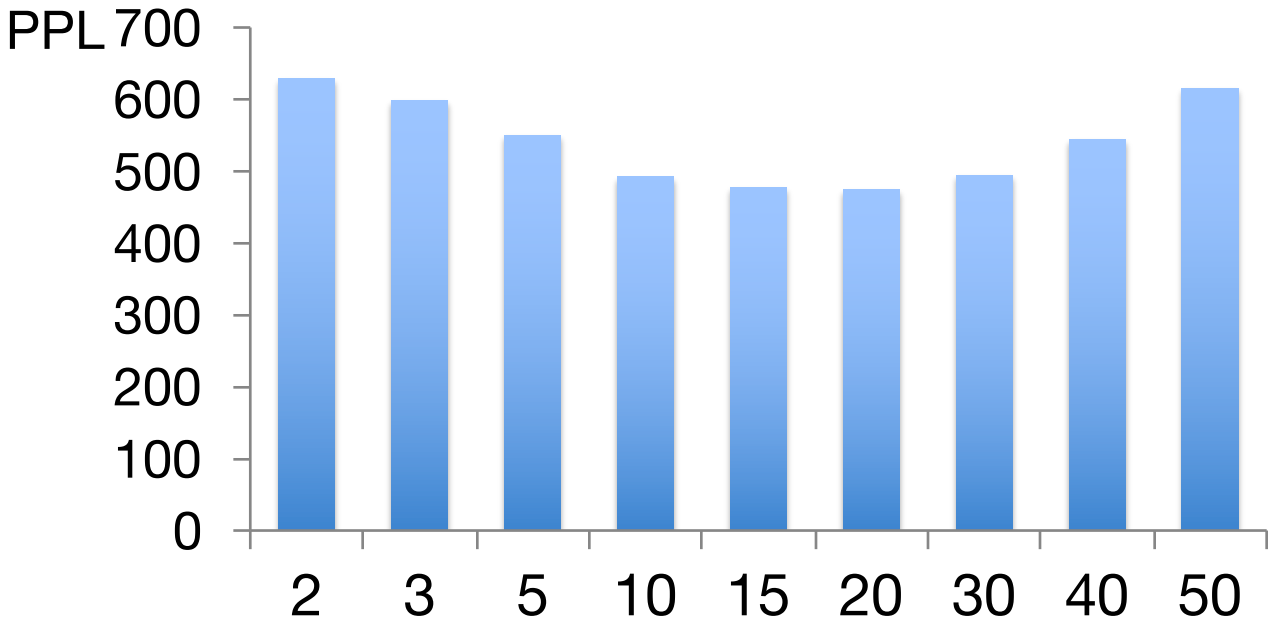


Replicated Softmax (Salakhut-dinov 2009) のNIPSコーパスでの実験結果

- RBMには、学習率、ミニバッチサイズ、モーメント、CD iterations、..などの多数のメタパラメータ
- ほとんどの場合、非常に悪い性能しか出ない

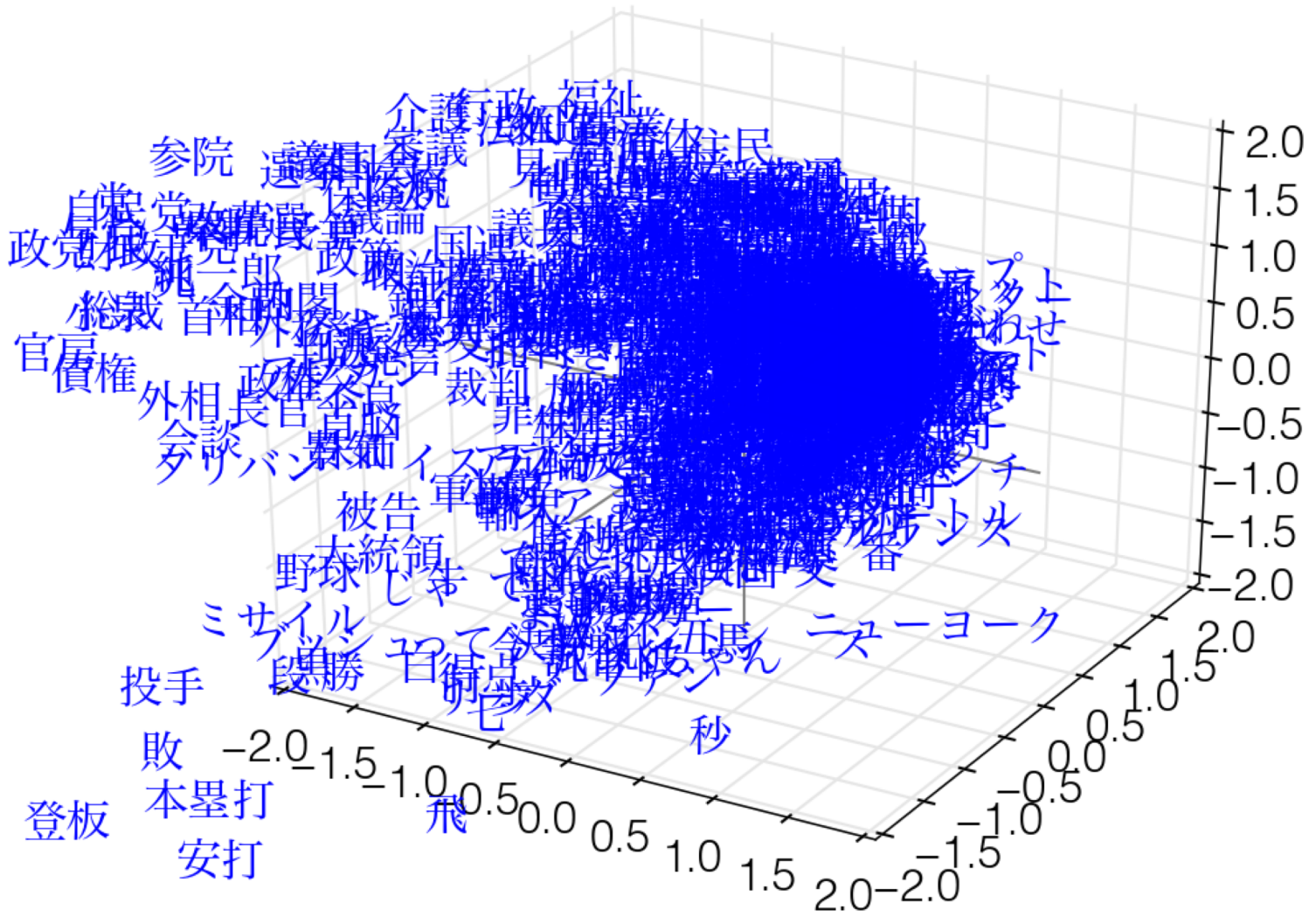
CSTMの次元選択

- 毎日新聞データでの性能と潜在次元数



  - 文書の潜在次元が連続なため、小さい値で高性能
  - 次元選択を行う簡単な方法はない (Beta FA?)

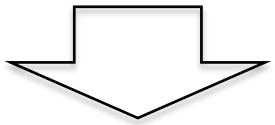
毎日新聞テキスト (2000年度)



- 出現に偏りの大きい語ほど原点から遠くに位置する

拡張: 文書共変量

- ラベルやキーワードなどの文書共変量を利用する dependent トピックモデルの研究は無数にあるが..
  - Supervised topic model (Blei+2007), Labeled LDA (Ramage+ 2009), DHNRM (Chen+ 2012), ...
- すべて、共変量をトピック混合比 θ に反映させるのみ



- 単語と共変量の関係がとらえられない!
  - 書き言葉と話し言葉の違い
  - 女性の書いた文書と男性の文書の違い
  - 夜に書いたメールに出やすい単語 (内容とは直交して)

拡張: 文書共変量 (2)

- 文書dの共変量を $c(d) = (c_1, \cdots, c_m)$ とおいて、基本モデルを拡張

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

  - MCMCの中で、共変量→単語の重み行列を同時学習

- 実験: CSJ話し言葉コーパス
  - 女性の講演と男性の講演が存在 (1381/1921)
  - 上のモデルにより、内容の影響を除去

単語の「女性度」の上位・下位語

上位語		下位語	
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

拡張: 語彙共変量

- 単語は単なるIDではなく、さまざまな素性が存在
  - 文字種 (ひらがな／カタカナ／漢字..)
  - 活用形
  - 起源 (roman/saxon words、漢／呉／唐)
  - 丁寧語、口語、謙譲語、...
- 通常、文書に共通した特徴があるはず
- 語彙共変量を $c(w) = (c_1, \cdots, c_n)$ とおくと、次式でモデルを拡張

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

  - 文書ごとの ζ_d を同時に学習

拡張: 語彙共変量 (注意)

- 両辺の対数をとれば、

$$\log \alpha(w) = \log \alpha_0 + \log G_0(w) + \phi(w)^T u_d + c(w)^T \zeta_d$$

  - これは対数線形モデル (一般化線形モデル)
  - 「切片」$\log \alpha_0, \log G_0(w)$ と「説明変数」$\phi(w)$による回帰からの残差を説明

拡張: 語彙共変量 (2)

- 毎日新聞テキストで、単語の文字種を素性

η(カタカナ)

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

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 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 ...

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 ...

普通のトピックモデルではとらえられない情報がモデル化できる!

- 著者の個性のモデリング
- 協調フィルタリングへの応用

まとめと展望

- 潜在空間上のガウス過程を考えることで、潜在変数が連続なRBMの生成モデル
  - MCMCによって容易に最適化できる
  - 通常のトピックモデルを常に超える性能
- 閉じた「ニューラルネット」ではなく、確率モデルとの連繋
  - 基底測度の Exponential Tilting
- 課題: 階層モデルへの適用
  - Normalized Random Measureの話だが、正規化定数を求めることが困難
  - これをバイパスできるか?