連続値と統計的自然言語処理

持橋大地
統計数理研究所
daichi@ism.ac.jp

IEICE PRMU研究会, 2012-6-29(金)

自己紹介

- NAIST 情報科学研究科
  松本研

- ATR 音声研／NTT CS研

- 統計数理研究所 (立川市)

LDAの衝撃

- 2002年, NAIST

$\mathbf{w} = w_1 w_2 \cdots w_N$：テキスト

$$p(\mathbf{w}) = \int \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \left(\prod_k \theta_k^{\alpha_k - 1} \right) \prod_n \sum_k p(w_n | \theta_k) \theta_k d\theta$$

  - 自然言語(離散シンボル)なのに $\int$!!!!
  - 一生, 積分は使わないものと思っていた

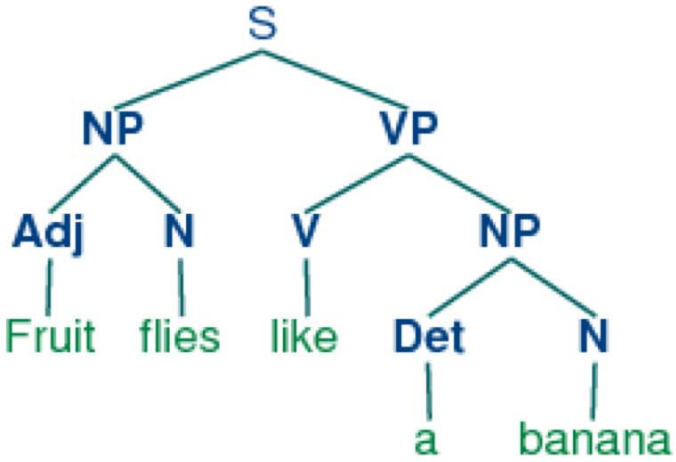
- 上の式の意味は? → これから説明します

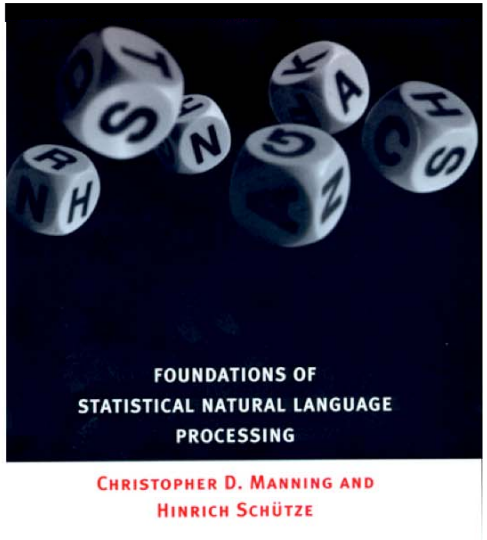
今日の概要

- 言語での連続表現の必要性
- トピックモデルLDAと関連モデル
- 言語のボルツマンマシン― RaP, LBL, HLBL
- ガウス過程, ガウス分布と自然言語
- その他の連続性 (空間など)
- まとめと研究課題

普通の? 自然言語処理

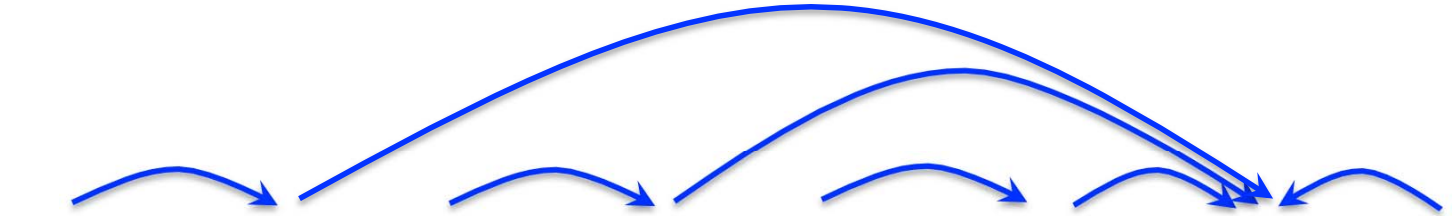
- 構文解析
- 係り受け解析
- 形態素解析
- 意味解析
  - 意味といってもラベルづけがほとんど






```
# S-ID:950104001-012 KNP:96/11/07 MOD:2005/01/27
* 0 1D
九四 きゅうよん * 名詞 数詞 * *
年度 ねんど * 接尾辞 名詞性名詞助数辞 * *
の の * 助詞 接続助詞 * *
* 1 2D
「「 * 特殊 括弧始 * *
減収 げんしゅう * 名詞 サ変名詞 * *
減益 げんえき * 名詞 普通名詞 * *
」」 * 特殊 括弧終 * *
社 しゃ * 名詞 普通名詞 * *
数 すう * 接尾辞 名詞性名詞接尾辞 * *
```

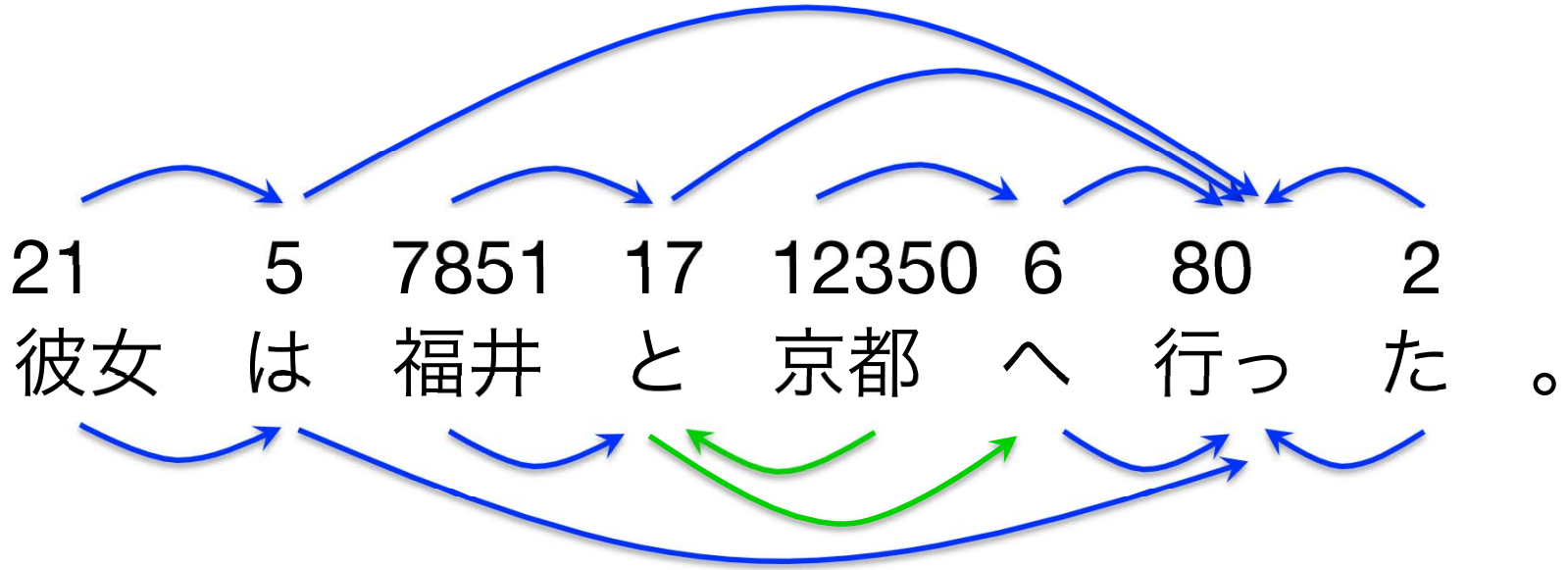
組み合わせ最適化問題?



係り先	2	7	4	7	6	7	0	7
Word	彼女	は	紀子	と	京都	へ	行っ	た
#	1	2	3	4	5	6	7	8

- 分類問題として解く
  - Perceptron, SVM, CRF, ・・・・
- 言語＝組み合わせ最適化で本当に充分?

計算機から見ると・・・・



- 同じ名詞の違いで、解釈が変わる
- 言葉の意味表現が、本来は不可欠
  - すべての係り関係の組み合わせを覚えるのは不可能 (「音威子府」「すり合わせる」)

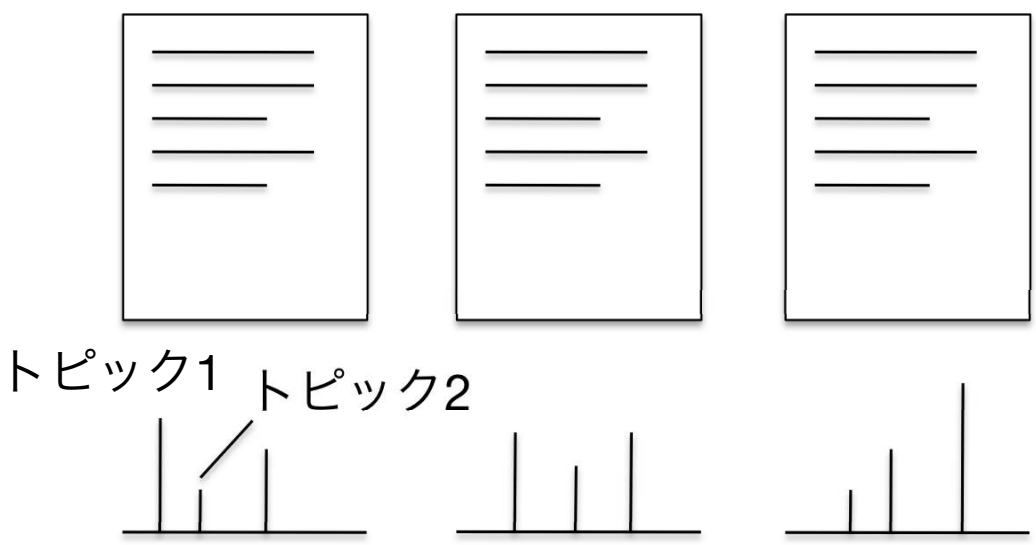
文書処理／情報検索

- テキストを単語の集まり (“国会” “税制” “社会保障” …)とその頻度として単純に表現
  - Bag of Words とよばれる
- 意味が中心課題！！

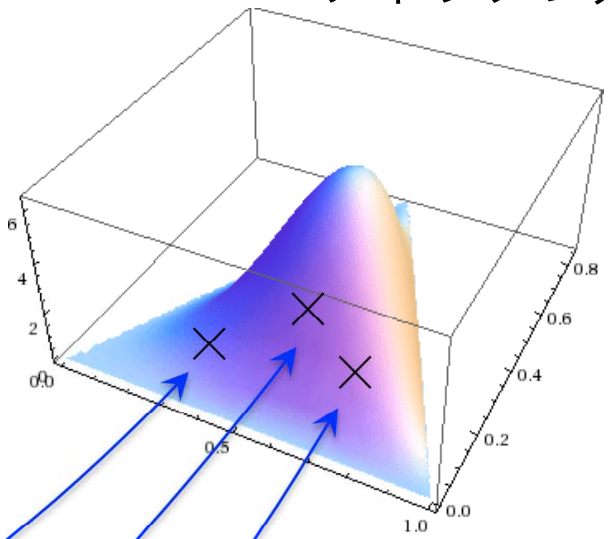
- テキスト分類の研究でよいか??
  - 人が勝手に決めた基準で分類
  - 迷惑メール判定などには確かに役立つが..

Latent Dirichlet Allocation: 潜在意味解析

- テキストに「潜在トピック分布」を割り当てる (Allocate)



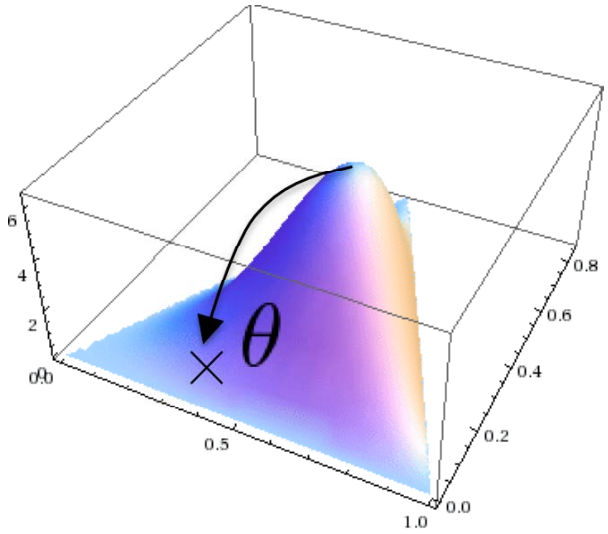
話題空間上のディリクレ分布



データごとの混合モデルになっている！

LDA：生成モデル

- 文書の確率的生成モデル
  - Draw $\theta \sim \mathrm{Dir}(\boldsymbol{\alpha})$
  - For $n = 1, \cdots, N$
    - Draw $z_n \sim \mathrm{Mult}(\theta)$
    - Draw $w_n \sim p(w|z_n)$



- 数式で書くと、

$$p(\mathbf{w}) = \int p(\theta|\boldsymbol{\alpha}) \prod_n \sum_{z_n} p(z_n|\theta) p(w_n|z_n) d\theta$$

$$= \int \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \left(\prod_k \theta_k^{\alpha_k - 1} \right) \prod_n \sum_k \theta_k p(w_n|k) d\theta$$

LDAの解法

- 変分ベイズEMアルゴリズム (オリジナル)
- Gibbs Sampler
- Collapsed 変分ベイズ
- Expectation Propagation
  - 最近は、これらの並列化、オンライン学習化が進んでいる

LDAの解法 (Gibbs)

- 各単語 w の持っている潜在トピック z を下の式に従って次々とサンプル:

$$p(z|w,d) \propto p(w|z)p(z|d)$$

  - 各単語wのトピックzへの割当て回数 n(w,z) がわかれば、
    - $p(w|z) \propto n(w,z) + \beta$
    - $p(z|d) \propto \sum_{w \in d} n(w,z) + \alpha$
  - MCMCで上を更新して繰り返す

LDAの学習結果の例

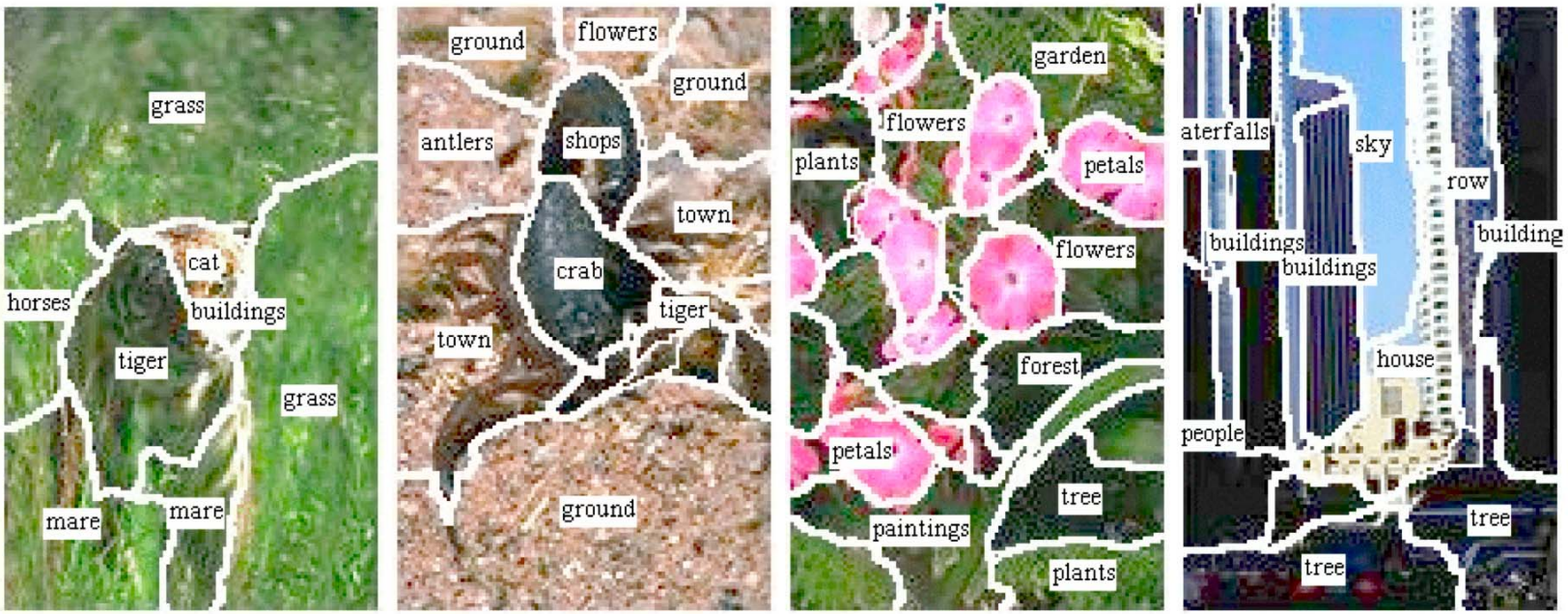
- 川端康成「雪国」の冒頭

国境の長いトンネルを抜けると雪国であった。
夜の底が白くなった。信号所に汽車が止まった。
向側の座席から娘が立って来て、島村の前のガラス
窓を落した。雪の冷気が流れこんだ。…

  - 2000年度毎日新聞記事全部 (2887万語) で学習したモデルで分析
    - 水色のトピックは冬に関係する
    - 緑色のトピックは電車に関係する
    - 黒色は地の文

画像処理への応用

- 古典的な適用: “Matching words and Pictures” (K.Barnard, ICCV 2001/JMLR 2003)



比較的最近の画像への適用

- Topic Random Field (Fei-Fei Li+, ECCV 2010)

Fig. 5. (Best viewed in color). Segmentation results of the MSRC database. From left to right: original image, segmentation result of spatial LDA and TRF.

$$p(\mathbf{z}^d|\boldsymbol{\theta}^d,\sigma) = \frac{1}{A(\boldsymbol{\theta}^d,\sigma)} \exp\left[\sum_n \sum_k z_{nk}^d \log \theta_k^d + \sum_{n\sim m} \sigma I(z_n^d = z_m^d)\right] \quad (1)$$

画像→文の生成

- 単純な最適化モデル (Farhadi, ECCV 2010)

A two girls in the store.

Yellow train on the tracks.

A small herd of animals with a calf in the grass.

A horse being ridden within a fenced area.

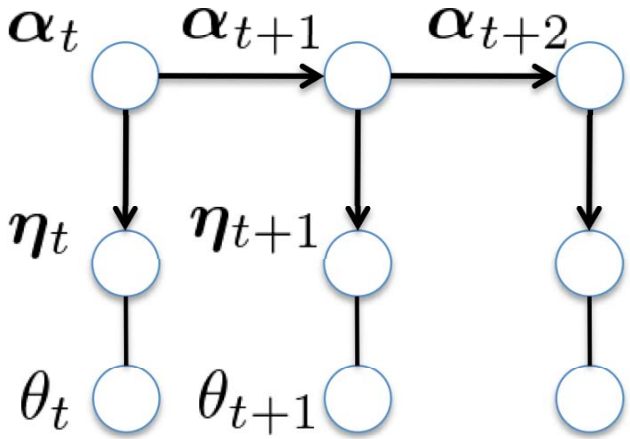
- HMMで生成モデルを真面目に解く (Hal+, EMNLP 2011)

{person,bicycle,ride,street,on}
the person is riding the bicycle on the street.

{person,table,sit,room,in}
three people are sitting at the table in the room.

Logistic Normalとカルマンフィルタ

- Dynamic Topic model (Blei&Lafferty, 2006)



α, ηがGaussianで時間変化

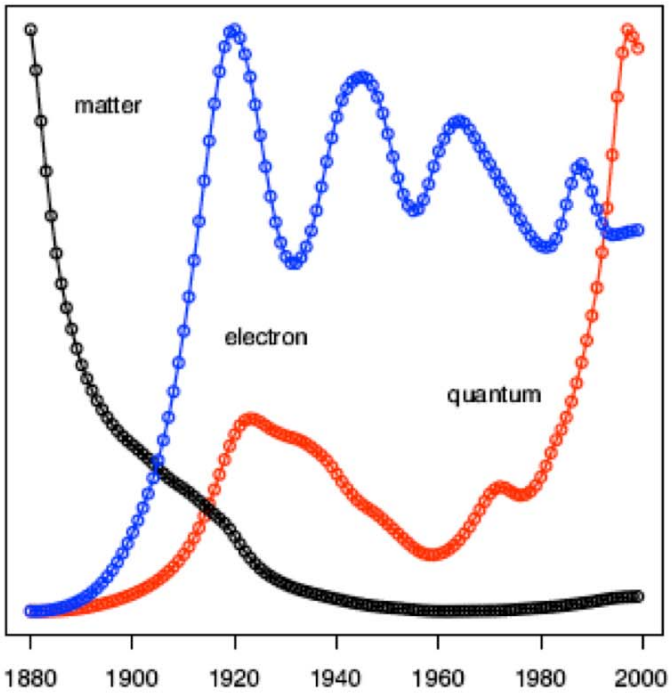
– 文書のトピック分布θは、ηのSoftmax

$$\theta_k = \frac{\exp(\eta_k)}{\sum_k \exp(\eta_k)}$$ (Logistic Normal)

– カルマンフィルタ+変分ベイズでηを追跡

Dynamic topic model の推定例

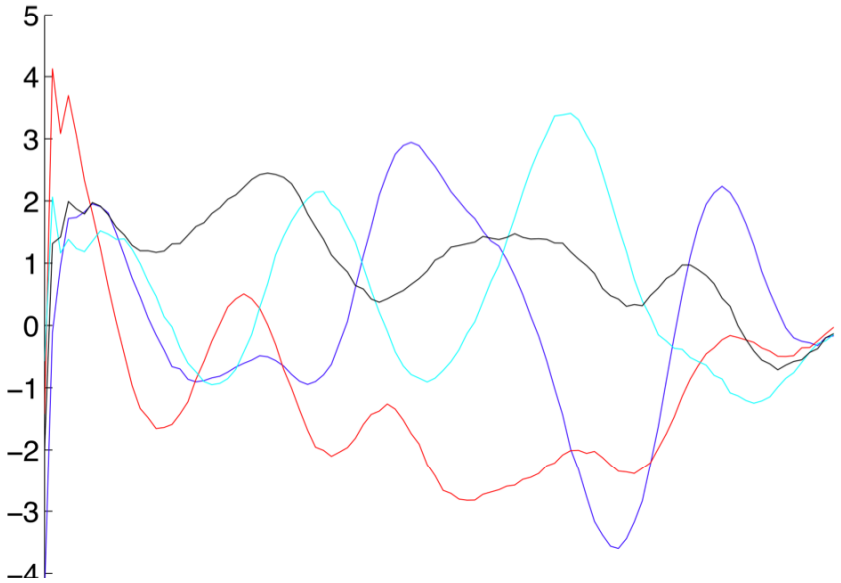
- Scienceコーパスの “Atomic physics” トピックの中身

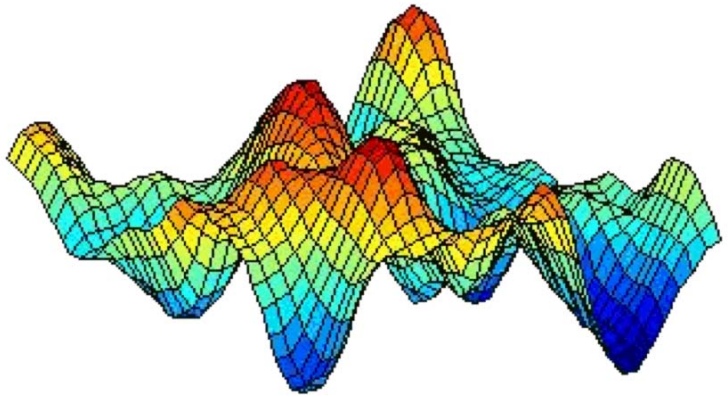


1881 On Matter as a form of Energy
1892 Non-Euclidean Geometry
1900 On Kathode Rays and Some Related Phenomena
1917 ``Keep Your Eye on the Ball''
1920 The Arrangement of Atoms in Some Common Metals
1933 Studies in Nuclear Physics
1943 Aristotle, Newton, Einstein. II
1950 Instrumentation for Radioactivity
1965 Lasers
1975 Particle Physics: Evidence for Magnetic Monopole Obtained
1985 Fermilab Tests its Antiproton Factory
1999 Quantum Computing with Electrons Floating on Liquid Helium

- 同じトピックでも、中身の移動が捉えられる

Gaussian Process とは





- ランダムな関数を生成するprior
  - 座標が似ている→値が似ている
  - 「似ている」の意味は、カーネル関数で定義
- 数学的には、単に無限次元のガウス分布

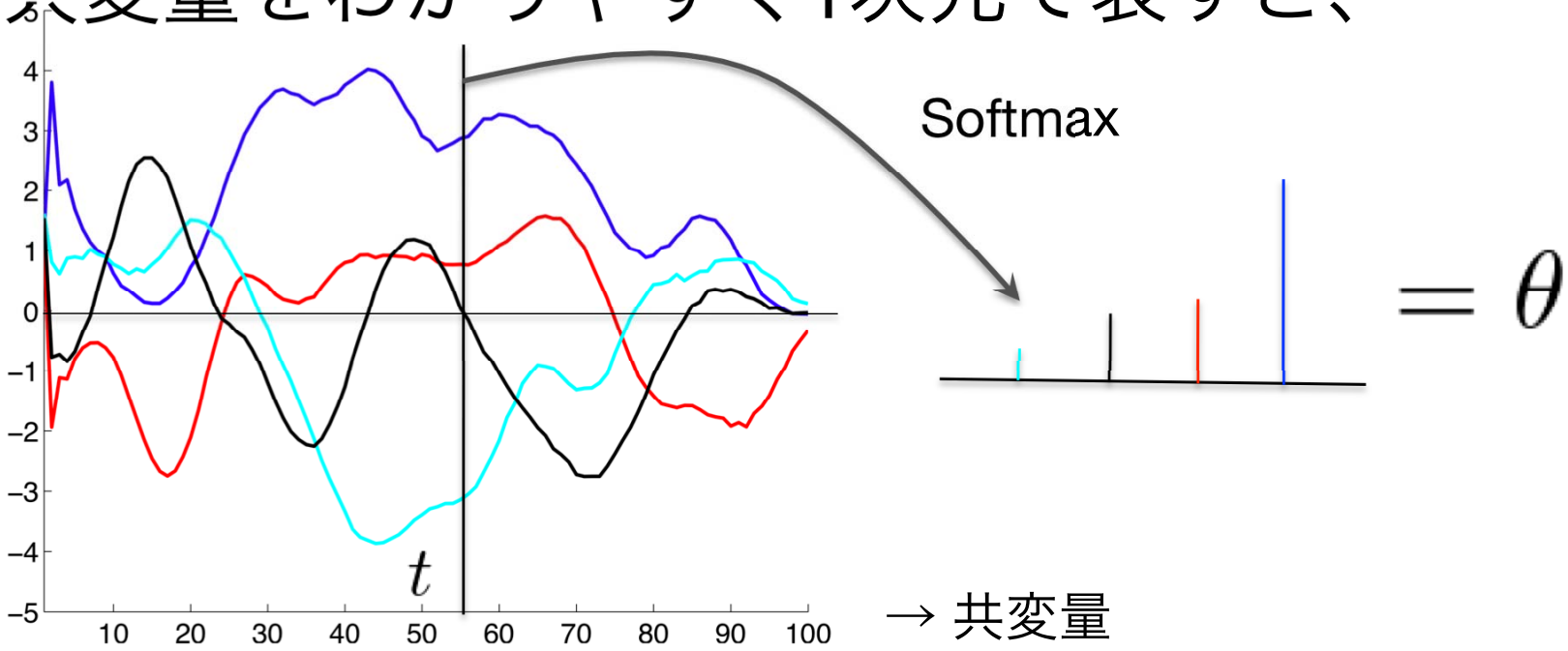
Gaussian Process topic models

- Agovic&Banerjee, UAI 2010
  - “Kernel Topic Models”, AISTATS 2012も同じ
- アイデア：文書の共変量=メタデータ (年, 地域, 時間,著者, …)が似ていれば、トピックも似ている
- 共変量の空間に、K個のGaussian process を発生

Softmaxを取ってトピック分布を作る

GPTM (KTM): 図解

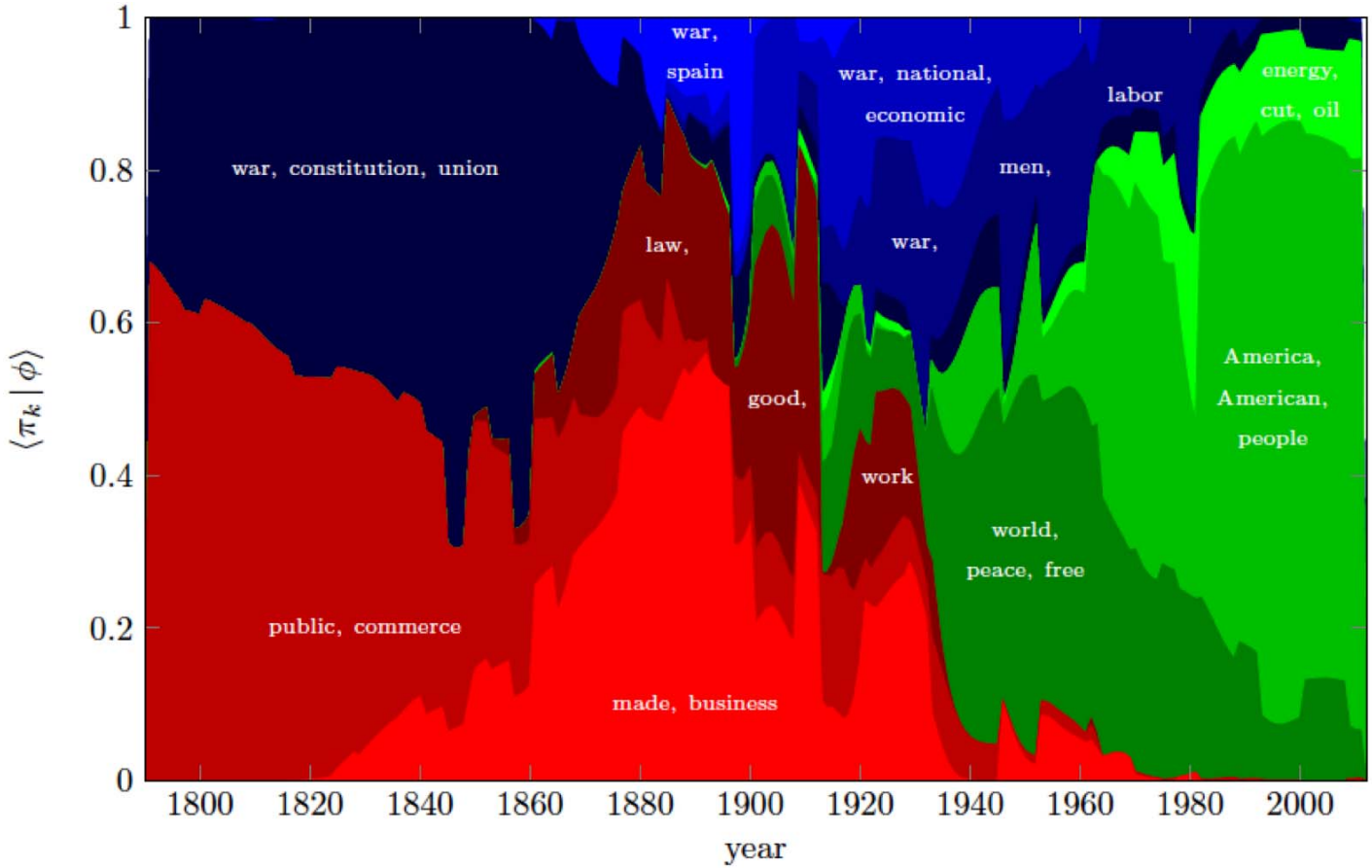
- 共変量をわかりやすく1次元で表すと、



- 点tでのスライスを中心に、さらに正規分布でノイズ→Softmaxで多項分布に

Kernel Topic Model の例

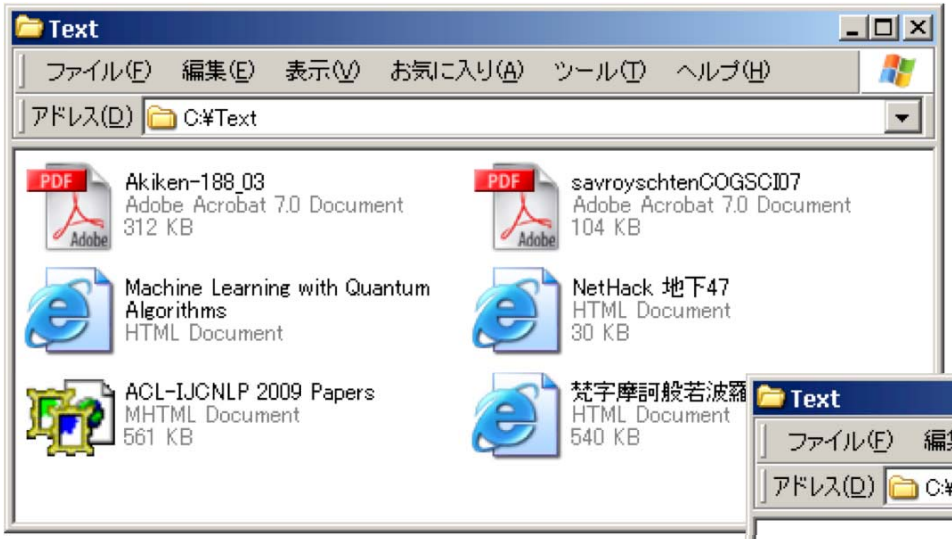
USAの一般教書演説

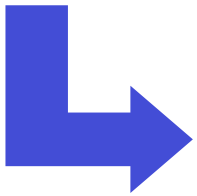


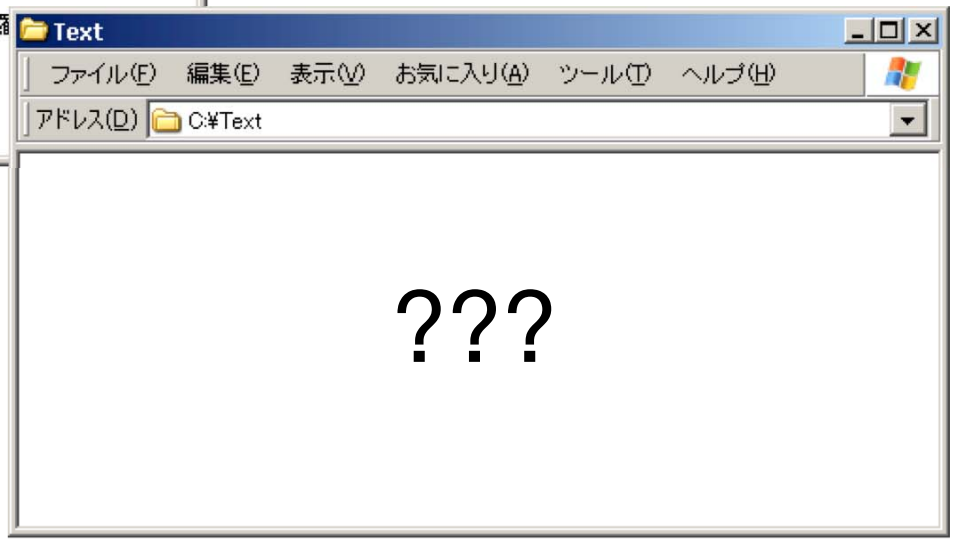
- 時間軸上のGPによる平均トピックの変化
  - 時間軸上のランダムなregression

Latent Topic Image Model (LTIM)

- テキストの中身を直感的にわかりたい

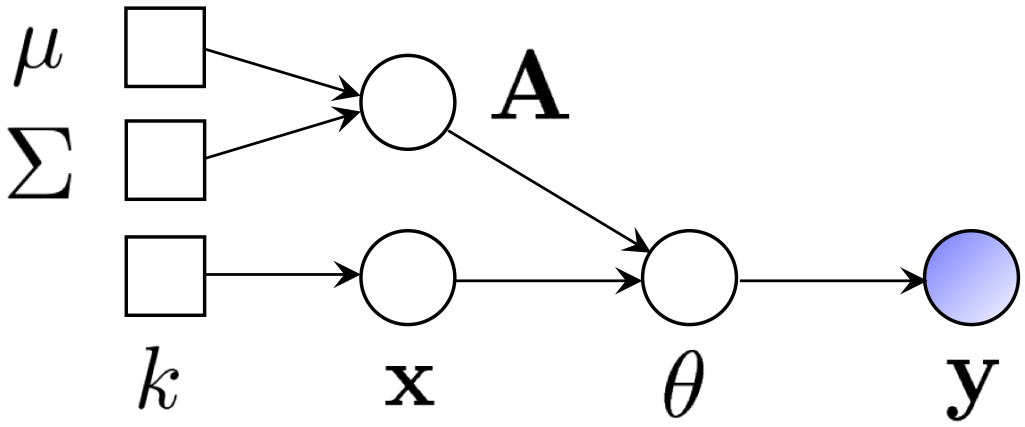






- 内容はまったく異なるが、アイコンはほぼ同じ
- 検索エンジンのWebページでも同様の問題

LTIM (石黒&持橋, IBIS 2010)



- 潜在画像 $\mathbf{x}$ を生成 : $\mathbf{x} \sim \mathrm{GP}(0, k(\mathbf{z}, \mathbf{z}'|\Psi))$
- 単語の “Activation” に変換 :

$$\theta \propto \exp(\mathbf{A}\mathbf{x})$$

- テキストの単語を生成 :

$$y_n \sim \mathrm{Mult}(\theta).$$

z はピクセル

LTIM: Objective

- LTIMの目的関数

$$\log p(\mathbf{A}, \mathbf{X}|\mathbf{Y}, \Psi, \mu, \mathbf{\Sigma})$$
$$\propto \log p(\mathbf{Y}|\mathbf{X}, \mathbf{A}) + \log p(\mathbf{X}|\Psi) + \log p(\mathbf{A}|\mu, \mathbf{\Sigma})$$

- ここで、各項は

$$\log p(\mathbf{Y}|\mathbf{X}, \mathbf{A}) = \sum_d \sum_i \sum_v y_{dnv} \mathbf{a}_v^T \mathbf{x}_d - \sum_d n_d \log \left(\sum_w \exp(\mathbf{a}_w^T \mathbf{x}_d) \right)$$

$$\log p(\mathbf{X}|\Psi) = -\frac{DZ}{2} \log(2\pi) - \frac{D}{2} \log |\mathbf{K}| - \frac{1}{2} \sum_d \mathbf{x}_d^T \mathbf{K}^{-1} \mathbf{x}_d$$

$$\log p(\mathbf{A}|\mu, \mathbf{\Sigma}) = -\frac{VZ}{2} \log(2\pi) - \frac{V}{2} \log(\mathbf{\Sigma})$$
$$- \frac{1}{2} \sum_v \left((\mathbf{a}_v - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{a}_v - \mu) \right)$$

LTIM: Inference

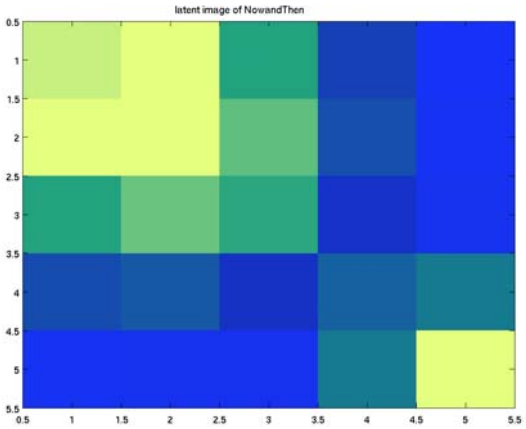
$$\frac{\partial \log p(\mathbf{A}, \mathbf{X}|\mathbf{Y})}{\partial \mathbf{a}_v} \propto \sum_d \left(n_{dv} - n_d \frac{\exp(\mathbf{a}_v^T \mathbf{x}_d)}{\sum_v \exp(\mathbf{a}_v^T \mathbf{x}_d)} \right) \mathbf{x}_d - \Sigma^{-1}(\mathbf{a}_v - \mu) \,,$$

$$\frac{\partial \log p(\mathbf{A}, \mathbf{X}|\mathbf{Y})}{\partial \mathbf{x}_d} \propto \sum_v \left(n_{dv} - n_d \frac{\exp(\mathbf{a}_v^T \mathbf{x}_d)}{\sum_v \exp(\mathbf{a}_v^T \mathbf{x}_d)} \right) \mathbf{a}_v - \Sigma^{-1} \mathbf{x}_d \,.$$

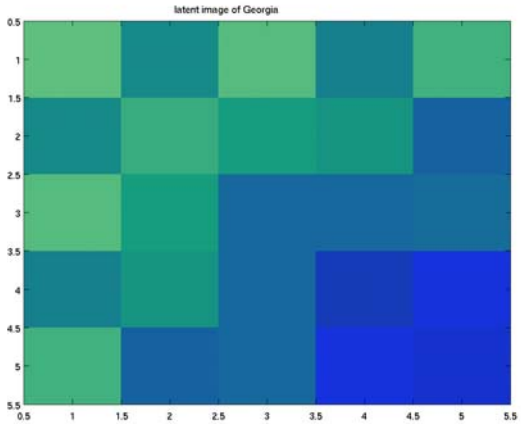
- L-BFGSで高速に最適化可能 (EM不要)
- ハイパーパラメータ $\mu, \mathbf{\Sigma}$ も通常の方法で推定

映画レビューの画像化

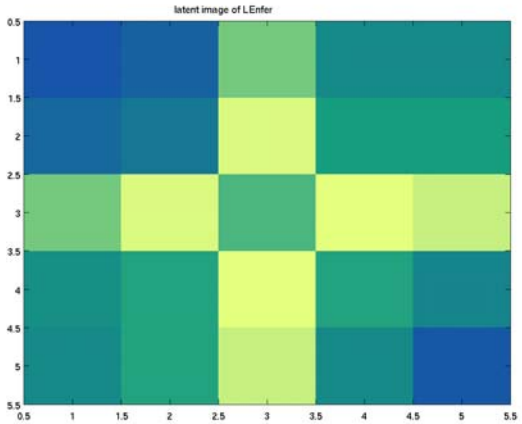
- EachMovieデータセット: 映画のレビュー



Now and Then



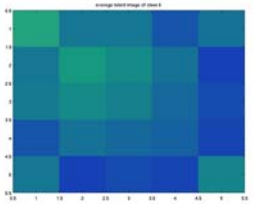
Georgia

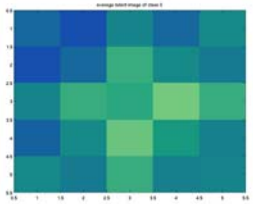


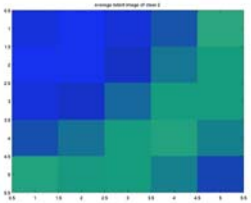
L'Enfer

  - レビュアーからの投票を“単語”とみなしている

- カテゴリ平均画像 [学習には未使用]

Class 6 = 

Class 3 = 

Class 2= 

分類実験

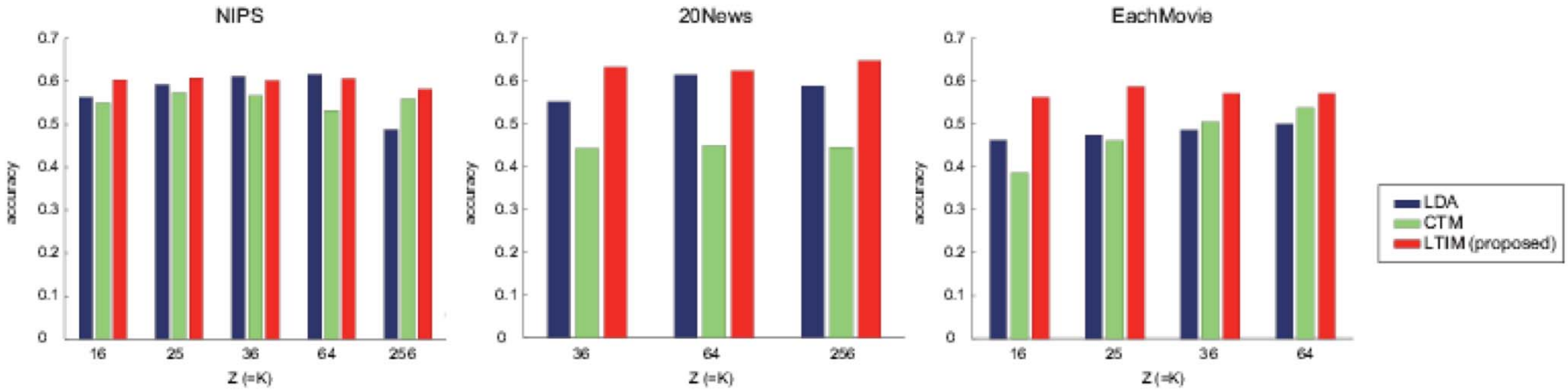
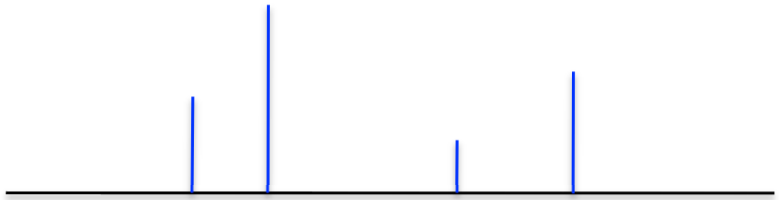


Figure 2: 1-NN classification results for three datasets.

- LTIMで得られた画像を1-NNで分類して精度を比較
- CTM, LDAと比べて常に高い性能(赤)を持つ

LDAは完全か?

- No.



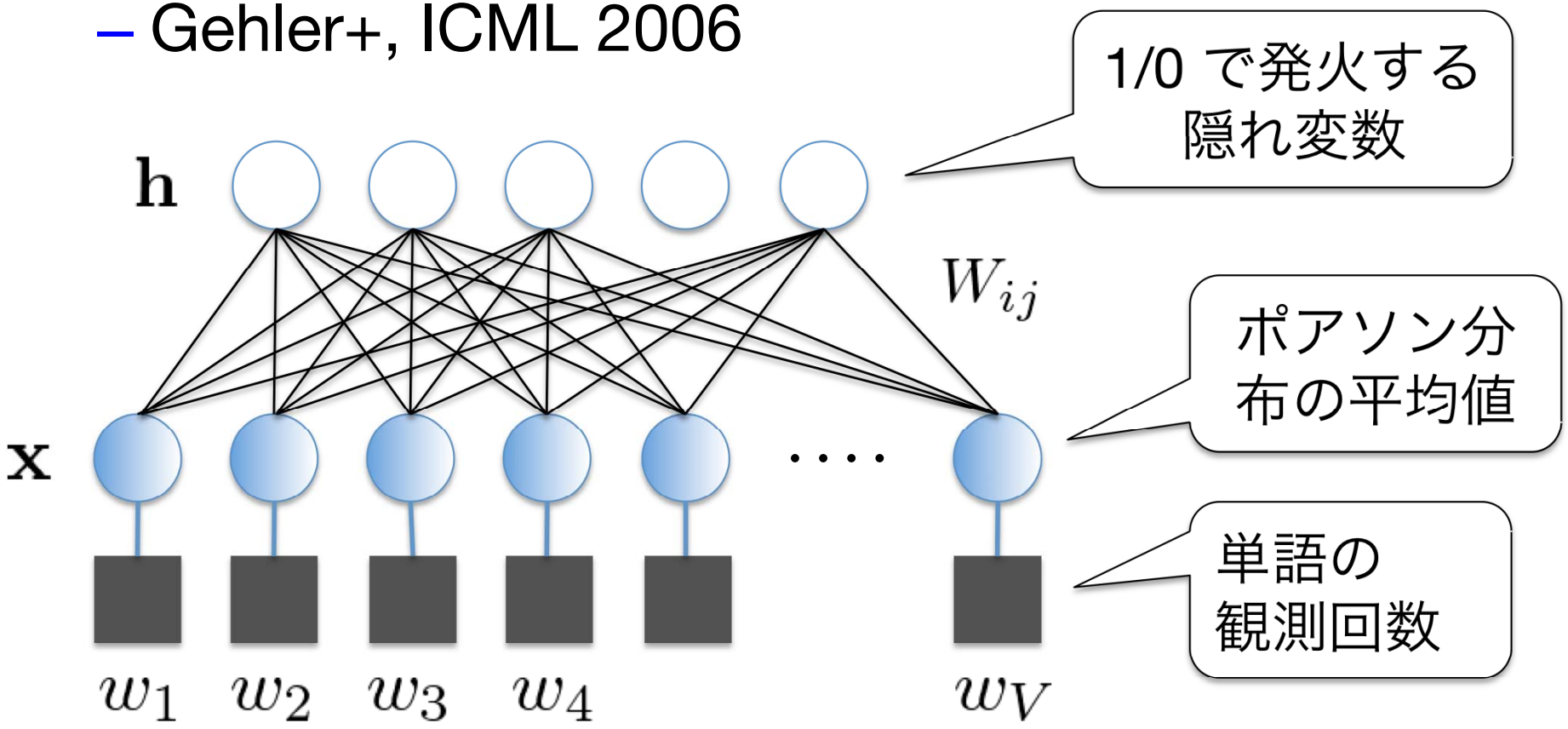
  - 和が1でなければならない (負の相関)
    →IBPによるmultifactorモデル (省略)
- 混合モデルベース

$$p(w) = \sum_k p(w|k)\,\theta_k$$

  - p(w|k)の混合→p(w|k)より鋭い分布は表せない
  - 和ではなく、積で表すモデル (Product model) が必要
    - RaP (ICML 2006), SAGE (ICML 2011) など

テキストの Boltzmann Machine

- RaP (Rate Adapting Poisson) モデル
  - Gehler+, ICML 2006



Restricted Boltzmann Machine (RBM)とよばれるニューラルネット

RaPの確率モデル

- 潜在層と観測層が条件付き確率で結ばれる

$$p(\mathbf{x}|\mathbf{h}) = \prod_i \mathrm{Po}\left(x_i \middle| \log(\lambda_i) + W_{ij} h_j\right)$$

$$p(\mathbf{h}|\mathbf{x}) = \prod_j \mathrm{Bin}\left(h_j \middle| \sigma(\log(\frac{p_j}{1-p_j}) + \sum_i W_{ij} x_i\right)$$

- 学習: xからhをサンプル／hからxをサンプル, をMCMCで繰り返す
  - 特別な形の Poisson regression

RaPの解釈

- 潜在トピック層を周辺化して消去すると,

$$p(\mathbf{x}) \propto \prod_i \lambda_i \frac{e^{x_i}}{x_i!} \cdot \prod_j (1+\exp(\underbrace{\sum_i W_{ij} x_i - \beta_j}_{\text{トピック } j \text{ に関するxの “activation”}}))$$

x の Poisson 事前確率

トピック j の励起度 ≧ 1

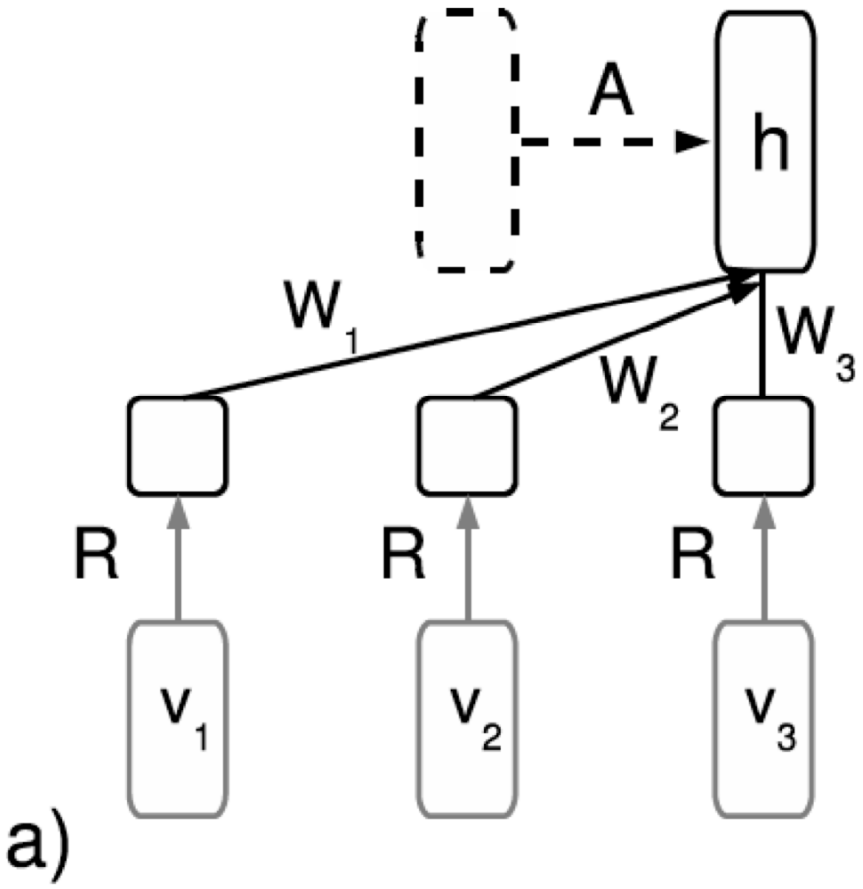
  - ポアソン分布×トピック別の励起度の積

$\beta_j = -\log\left(\frac{p_j}{1-p_j}\right)$ とした

言語モデルへの拡張

- RBMを時系列の言語データに拡張できないか?
- 言語モデル: 文の確率 $p(w_1, w_2, \cdots, w_N)$ を計算
  - $p(w_1, \cdots, w_N) = \prod_{n=1}^{N} p(w_n | w_1 \cdots w_{n-1})$ より、
  - $p(w_n | w_1 \cdots w_{n-1})$ がわかればよい
- Neural probabilistic language model (NPLM) (Bengio 2003)に近い
  - NPLMはn-gramより高性能

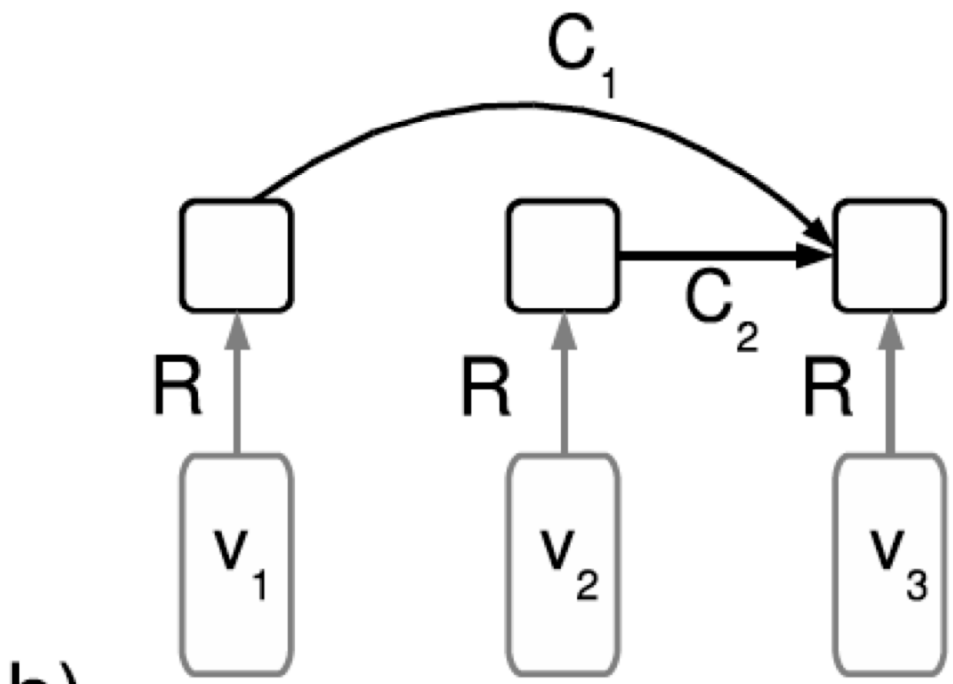
単純な拡張 (Mnih+ 2007)



- 各文脈に隠れ層hあり
- 単語v_iの連続表現 $v_i^T R$ とhを重み行列 W_i で内積
  →全体のエネルギー

$$E(w_1, \cdots, w_n, h) = -\sum_{i=1}^{n} (v_i^T R) W_i h + (\text{正則化項}).$$

LBL (Log-Bilinear Language model)



b)

(Mnih&Hinton, 2007)

- 隠れ層hを消去
- 予測語w_nと文脈w_iの連続表現を、位置依存のC_iで内積

$$E(w_1, \cdots, w_n, h)$$
$$= -\left(\sum_{i=1}^{n-1} v_i^T R C_i\right) R^T v_n$$
$$= -\sum_{i=1}^{n-1} \vec{w}_i^T C_i \vec{w}_n$$

  - これに正則化項

Word embeddingの例 (Mirowski+10)

Table 8. Examples of 10 closest neighbors in the latent word embedding space on the Reuters dataset, using an LBLN architecture with 500 hidden nodes, $|Z_W| = 100$ dimensions for the word representation and $|Z_X| = 5$ dimensions for the POS features representation. The notion of distance between any two latent word vectors was defined as the cosine similarity. Although word representations were initialized randomly and WordNet::Similarity was not enforced, functionally and semantically (e.g. both synonymic and antonymic) close words tended to cluster.

debt	aa	decrease	met	slow
financing	aaa	drop	introduced	moderate
funding	bbb	decline	rejected	lower
debts	aa-minus	rise	sought	steady
loans	b-minus	increase	supported	slowing
borrowing	a-1	fall	called	double
short-term	bb-minus	jump	charged	higher
indebtedness	a-3	surge	joined	break
long-term	bbb-minus	reduction	adopted	weaker
principal	a-plus	limit	made	stable
capital	a-minus	slump	sent	narrow

LBL＞n-gram

Table 2. Perplexity scores for the models trained on the 14M word training set. The mixture test score is the perplexity obtained by averaging the model's predictions with those of the Kneser-Ney 5-gram model. The log-bilinear models use 100-dimensional feature vectors.

Model type	Context size	Model test score	Mixture test score
Log-bilinear	5	117.0	97.3
Log-bilinear	10	107.8	92.1
Back-off KN3	2	129.8	
Back-off KN5	4	123.2	
Back-off KN6	5	123.5	
Back-off KN9	8	124.6	

- LBLはKneser-Ney n-gramよりかなり高性能

LBL/NPLMの最近の話

- Hierarchical LBL (HLBL)
  - (Mnih&Hinton, NIPS 2008)
  - 語彙を階層クラスタリングして計算量削減
- LBLの学習高速化 (Mnih&Teh, ICML2012)
  - Contrastive estimationで勾配を計算
- 音声認識への適用 (Mirowski+ 2010)

Table 7. Speech recognition results on TV broadcast transcripts, using the same training set and test set as in Table 6, but with the true sentence to be predicted included among the n-best candidates.

Method	Accuracy
Back-off KN 4-gram	86.9 %
LBLN+POS+init	**94 %**
"Oracle"	100 %

そのほかの連続性と言語

- Geographic topic model (Eisenstein+, EMNLP 2010)
  - トピックの地域毎のvariantを生成
  - Logistic Normal ベース (+VB-EM)
- 時間的モデル
  - 年などでは多数の研究
  - 時刻ベースの研究はあまりない (vMF? GP?)

Geographic topic modelの例

	"basketball"	"popular music"	"daily life"	"emoticons"	"chit chat"
	PISTONS KOBE LAKERS game DUKE NBA CAVS STUCKEY JETS KNICKS	album music beats artist video #LAKERS ITUNES tour produced vol	tonight shop weekend getting going chilling ready discount waiting iam	:) haha :d :(;) :p xd :/ hahaha hahah	lol smh jk yea wyd coo ima wassup somethin jp
Boston 	CELTICS victory BOSTON CHARLOTTE	playing daughter PEARL alive war comp	BOSTON	;p gna loveee	*ese* exam suttin sippin
N. California 	THUNDER KINGS GIANTS pimp trees clap	SIMON dl mountain seee	6am OAKLAND	*pues* hella koo SAN fckn	hella flirt hut iono OAKLAND
New York 	NETS KNICKS	BRONX	iam cab	oww	wasssup nm
Los Angeles 	#KOBE #LAKERS AUSTIN	#LAKERS load HOLLYWOOD imm MICKEY TUPAC	omw tacos hr HOLLYWOOD	af *papi* raining th bomb coo HOLLYWOOD	wyd coo af *nada* tacos messin fasho bomb
Lake Erie 	CAVS CLEVELAND OHIO BUCKS od COLUMBUS	premiere prod joint TORONTO onto designer CANADA village burr	stink CHIPOTLE tipsy	;d blvd BIEBER hve OHIO	foul WIZ salty excuses lames officer lastnight

研究課題

- 現在の連続モデルは、ユニグラム(bag of words)のモデル
  - 構文解析や係り受け解析、統計的翻訳などとの同時モデルはない
  - 素性として使うだけでは、性能はあまり上がらない
- 動作、動画とのモデルは面白い
  - 教師データは大量にある (TV番組, 映画等)
  - キャプションやコメントは内容と一対一対応していない＆ノイズがあるため、統計モデル必須

まとめ

- 言語は表面は離散だが、
  - トピックモデル
  - Gaussian process, ガウス分布
  - Boltzmann Machine

  などを使って連続化して考えることが可能
- 自然言語処理の人は一般に上には詳しくない
  → 研究チャンス
- 雑音の多い学習データなので、頑健な統計的手法が必須

- ご清聴ありがとうございました。