グラフィカルモデルと情報検索

持橋大地
統計数理研究所
daichi@ism.ac.jp

IFAT 情報アクセスシンポジウム2013
2013-12-06 (金)

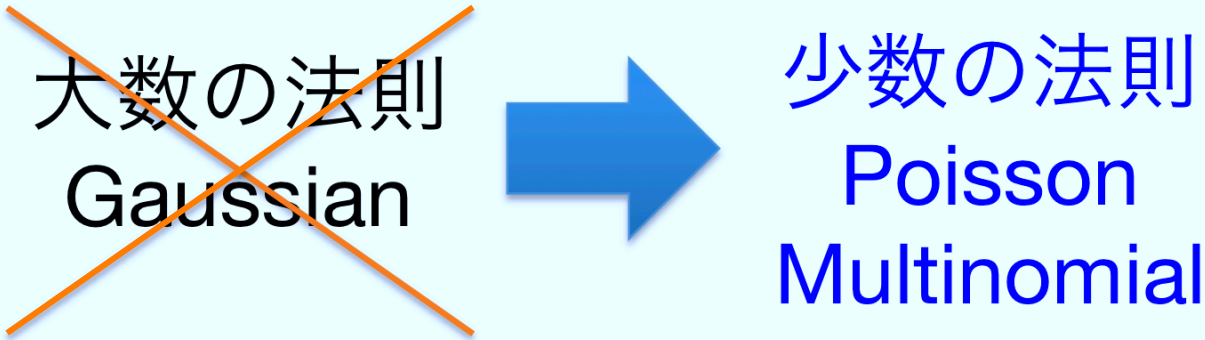
Research Organization of Information and Systems
The Institute of Statistical Mathematics

情報検索: “Relevant Document”

- Query q から文書 D への関連度を測る
- よく使われる Cosine 類似度: $\cos\theta = \frac{\vec{q} \cdot \vec{D}}{|\vec{q}||\vec{D}|}$
  - Cosineにする理由があるのか?
  - どうやって「個人化」するのか?

No More Cosine!

- Cosine類似度＝データが正規分布という仮定 (Papadimitriou 1998)
- 言語データや離散値の頻度データはまったくGaussianではない！
  - そもそも必ず非負値
  - 連続でない、裾がもっと長い



“IR as a Statistical Translation”

- $p(D|q) \propto \underbrace{p(q|D)}\ \underbrace{p(D)}$ に従ったランキング

  文書→クエリの生成　文書の生成

- Berger&Lafferty, 1999

ベイズの定理の復習

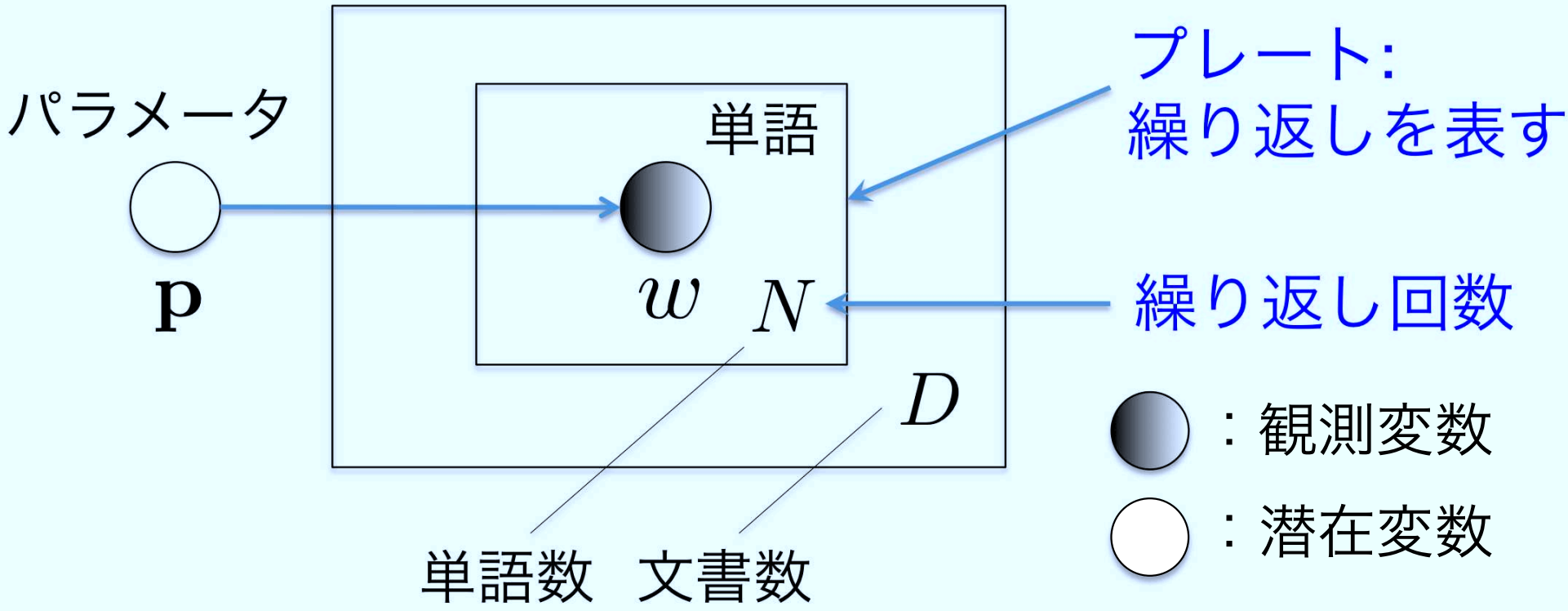
$$p(X|Y) = \frac{p(X,Y)}{p(Y)} \propto p(X,Y)$$

$$= p(Y|X)p(X)$$

- $p(X|Y)$ は、$p(X,Y)$ に比例
- $p(X|Y)$ を、引っくり返した $p(Y|X)p(X)$ で計算できる

文書の生成モデル

- 最も簡単な場合：$\mathbf{p} = (p(w_1), p(w_2), \cdots, p(w_V))$ から文書 $d = w_1 w_2 w_3 \cdots w_N$ が生成
  - Unigram (1-gram) 分布とよぶ



文書の生成モデル: Unigram

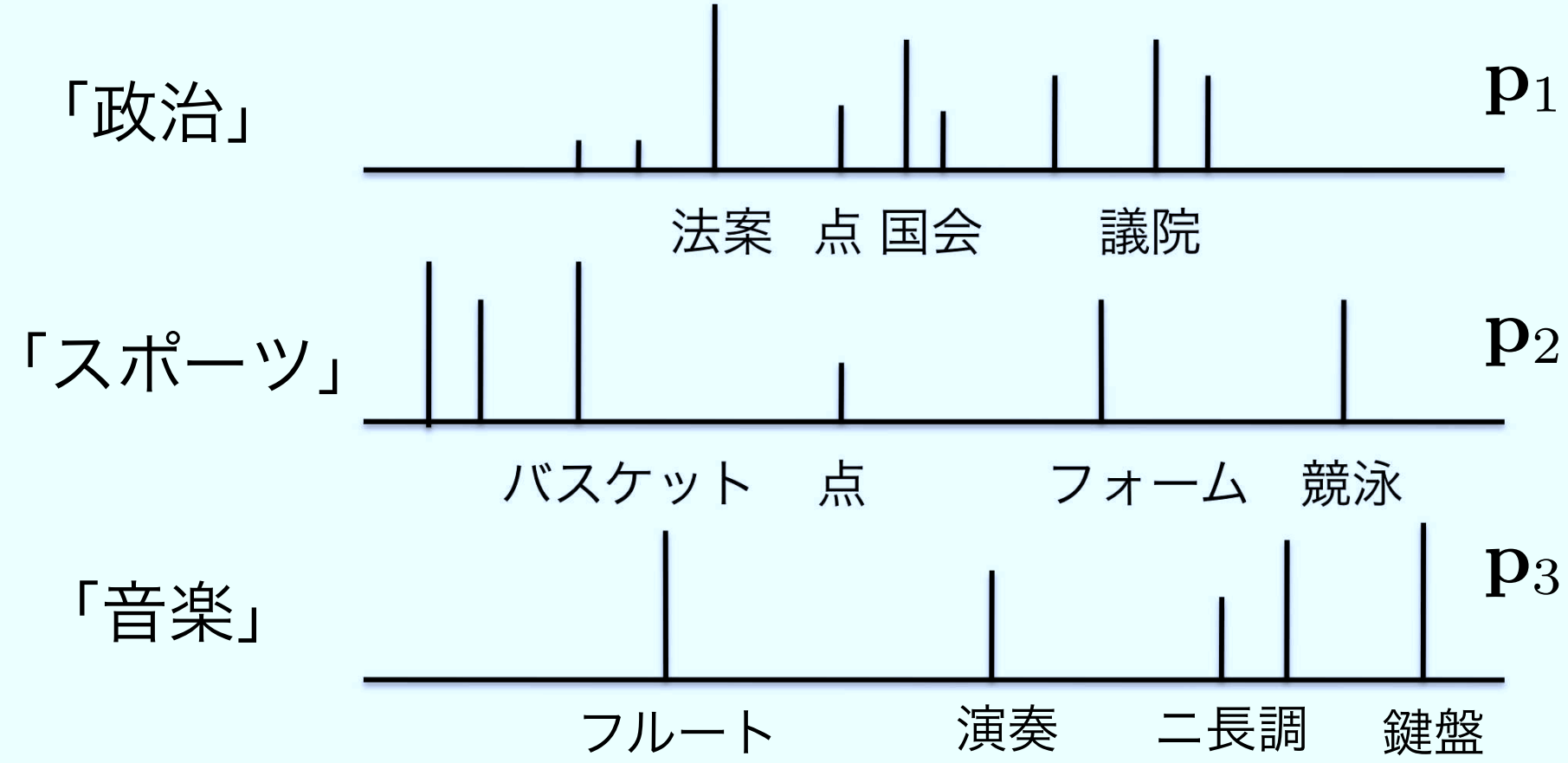
点 あっ おり とか
重要 川崎 従来
求め 国会
アルジェリア 創作
なっ 慎太郎 ない
引き揚げ 者 達成
さ 保育園 大正 結
希望 議会 グローバル
現場 言及 的 中核
週 応募 通商

独自 抱える 場合
村 組 花 準備
として 世界中 いまだ
れ か すさまじい 電子
進ん ターゲット 満員
いる 機体 億 対戦
中期 なく 月 疑い
住宅 性 足利 られる
常識 準 と た 県 目
メンバー 市 は

$$p(d) = p(w_1 w_2 \cdots w_N) = \prod_{i=1}^{N} p(w_i)$$

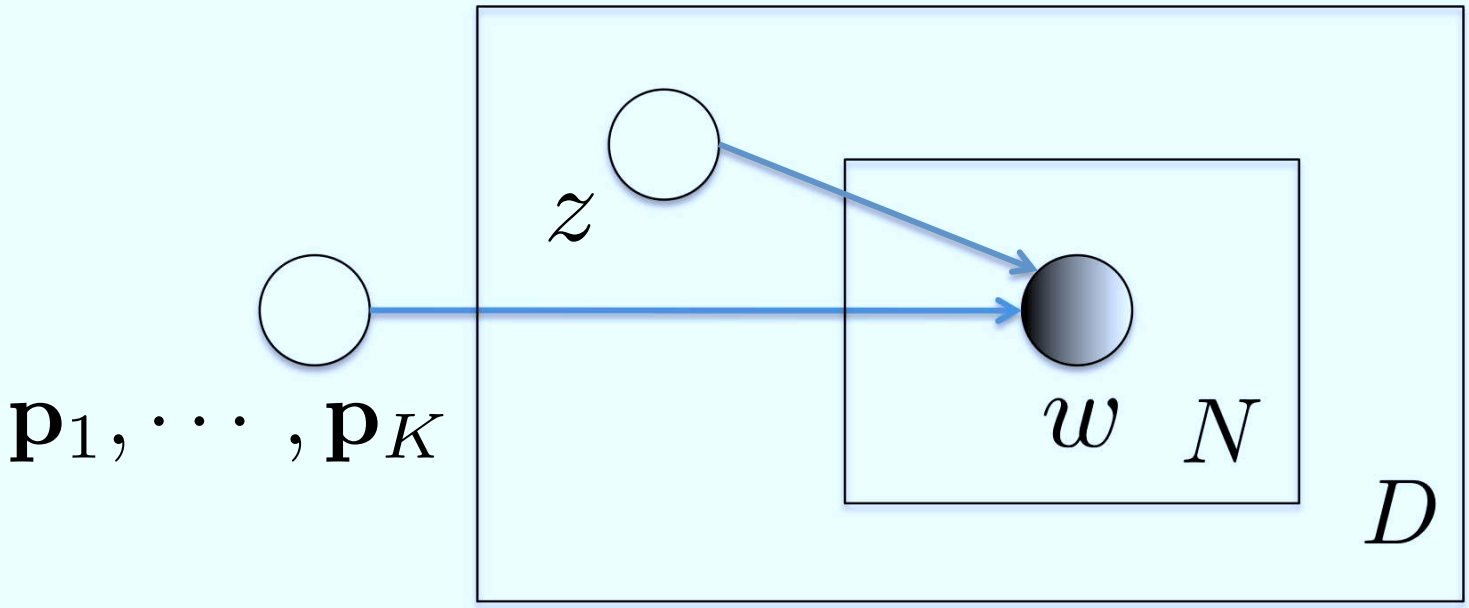
何が足りないか?

- 文書の「個性」を表現したい
- 「こういう個性の文書を探したい」



Unigram Mixtures (Nigam+ 2000)

- 文書全体の「話題」z_i を付加



- $z_i \in \{0, 1\}$ の場合：文書分類、「ナイーブベイズ」

Unigram Mixtures (2)

- UMからランダムに生成した文書の例

トピック83	トピック93
なる 駒大 た で 五	五輪 浜松 する 発
サッカー の 移転	全国 開く れる
修司 リーグ ドル	と 情報 で O 世界
リード イワノフ	高校生 失明 月
# 山口 連続 に	も は 大会 の
目 に 投手 また	大会 線 田中
の する # と 死	テロ 決勝 返上
新 行わ # 代打	を AP 同 計 の
派 勝負 # の	連鎖 で 現役 規模
今季 試合	は し # もの た 首都

Unigram Mixtures (3)

- どうやって学習するか?→EMアルゴリズム

$$p(d, z) = p(z) \prod_{i=1}^{N} p(w_i|z)$$

  - Eステップ: $p(z|d) \propto p(z, d)$ を計算
  - Mステップ: 対数尤度の期待値

$$\left\langle \log p(d, z) \right\rangle_{p(z|d)}$$

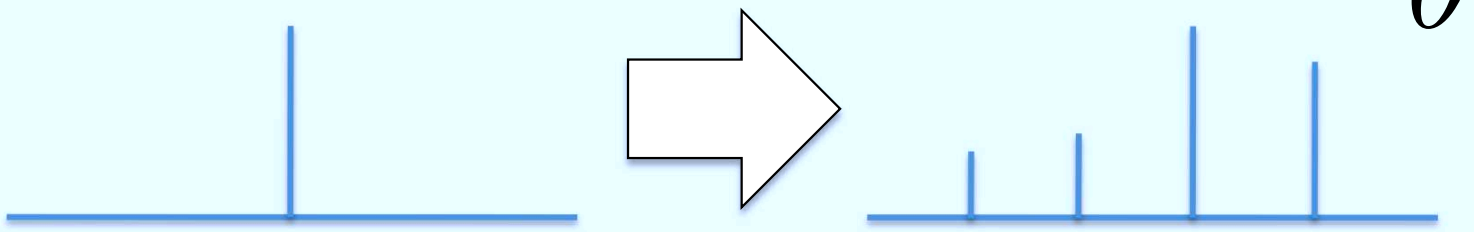
から $p(z)$, $p(w|z)$ を更新

Eステップへ戻る.

LDA:「トピックモデル」

- 動機: Unigram Mixturesでは、文書ごとに1つの話題しか考えていない
  - 「農業政策」「音楽と国際化」「スポーツ振興の予算枠」・・・・

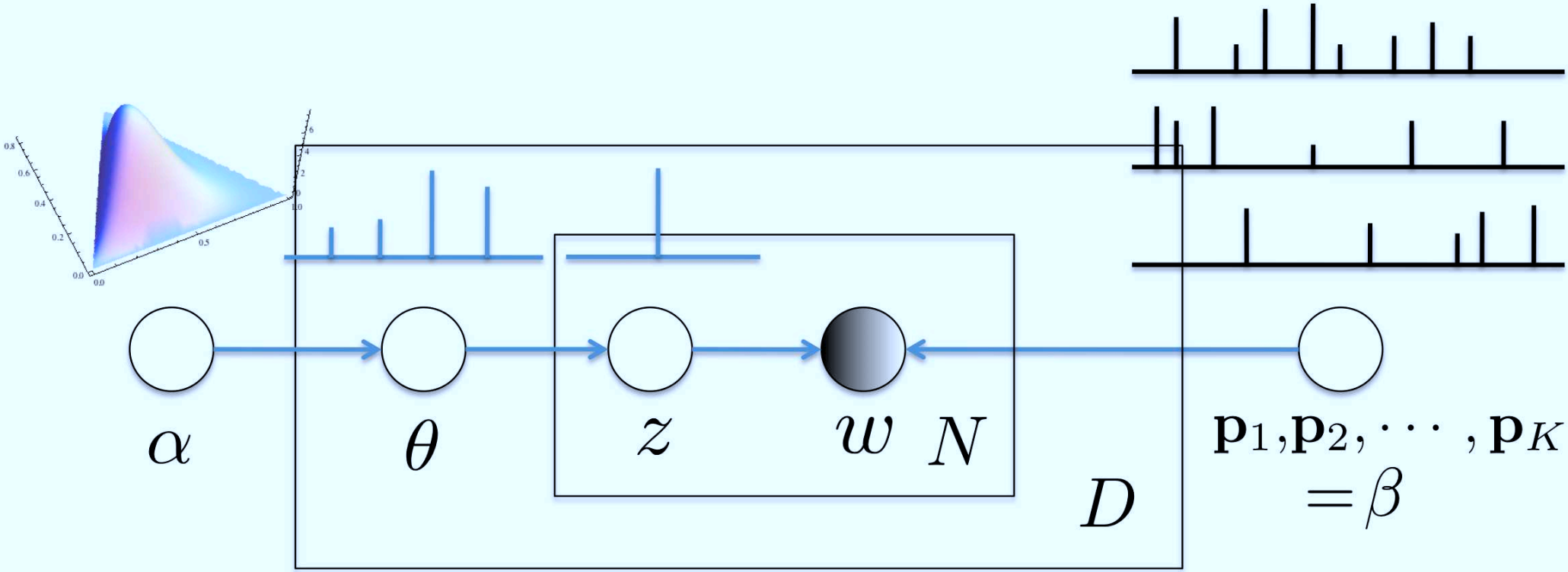
話題の混合を考える



(Blei+ 2001, 2003)

LDAのグラフィカルモデル

- $\alpha \rightarrow \theta \rightarrow z \rightarrow w$ の順で単語を生成
  - θ はディリクレ分布 Dir(α)から生成



LDAによるトピック分析の例

- 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文 (2,887万語) で学習したモデルで分析
    - 青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

情報検索／統計的要約への応用

- 「真に重要な文」だけを取り出したい
  - ムダな単語を削りたい
- 複数の文書の意味的な共通点を要約したい

bqfs (Hal Daume 2006), HierSum (Haghighi&Klein 2009) のアイデアを紹介

Bayesian Query-focused Summarization

Simple Unigram summarization

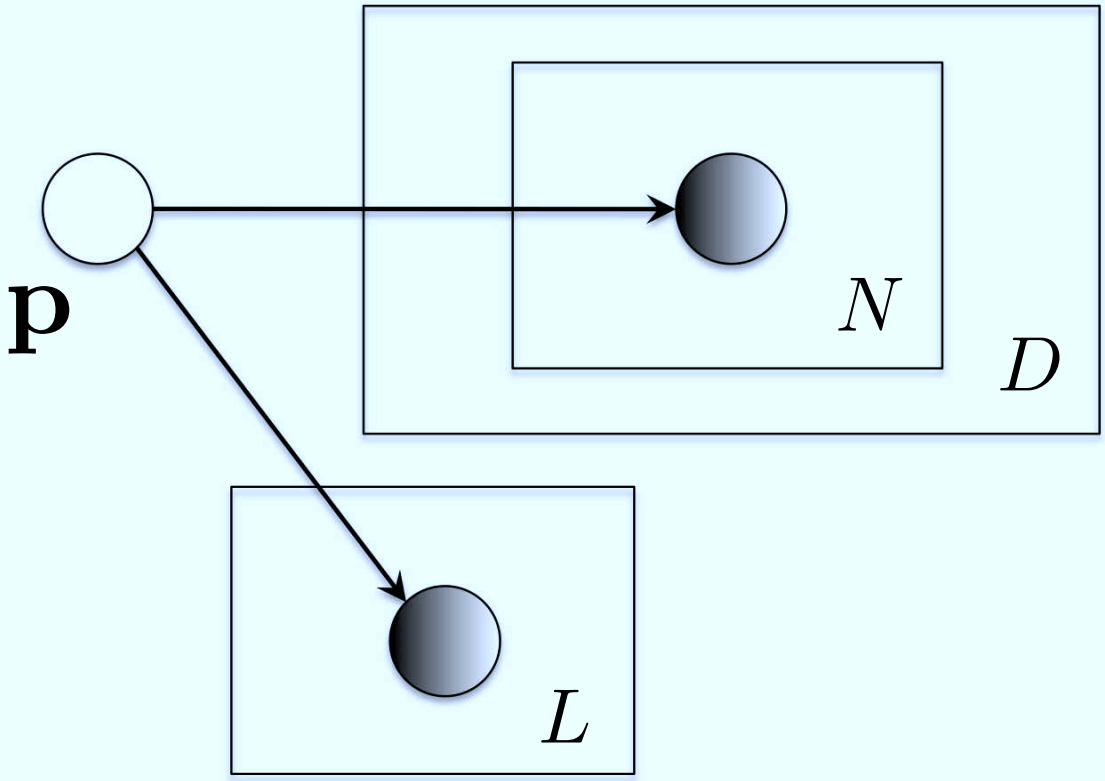
- 要約したい文書集合の単語分布と、要約文の単語分布は似ている（はず）
- Kullback-Leibler ダイバージェンス

$$\mathrm{KL}(D||S) = \sum_w p_D(w) \log \frac{p_D(w)}{p_S(w)}$$

を最小化

  - $p_D = p_S$ のときに最小値0
  - DとSが同じ分布pから生成されるようにする

Simple Unigram summarization (2)



- 文書集合全体と、要約が同じ単語分布pを共有
  - $\mathrm{KL}(D||S) = \sum_w p_D(w) \log \frac{p_D(w)}{p_S(w)}$ を最小化

Simple Unigram summarization (3)

- 要約 S={} として、Sにgreedyに文を追加
  - KLの値が最小になるような文を選択して追加する
  - 結果: DUC 2006の最高値とほぼ同等

System	ROUGE -stop			ROUGE all		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4
SUMBASIC	29.6	5.3	8.6	36.1	7.1	12.3
KLSUM	30.6	6.0	8.9	38.9	8.3	13.7

Global/Local モデル

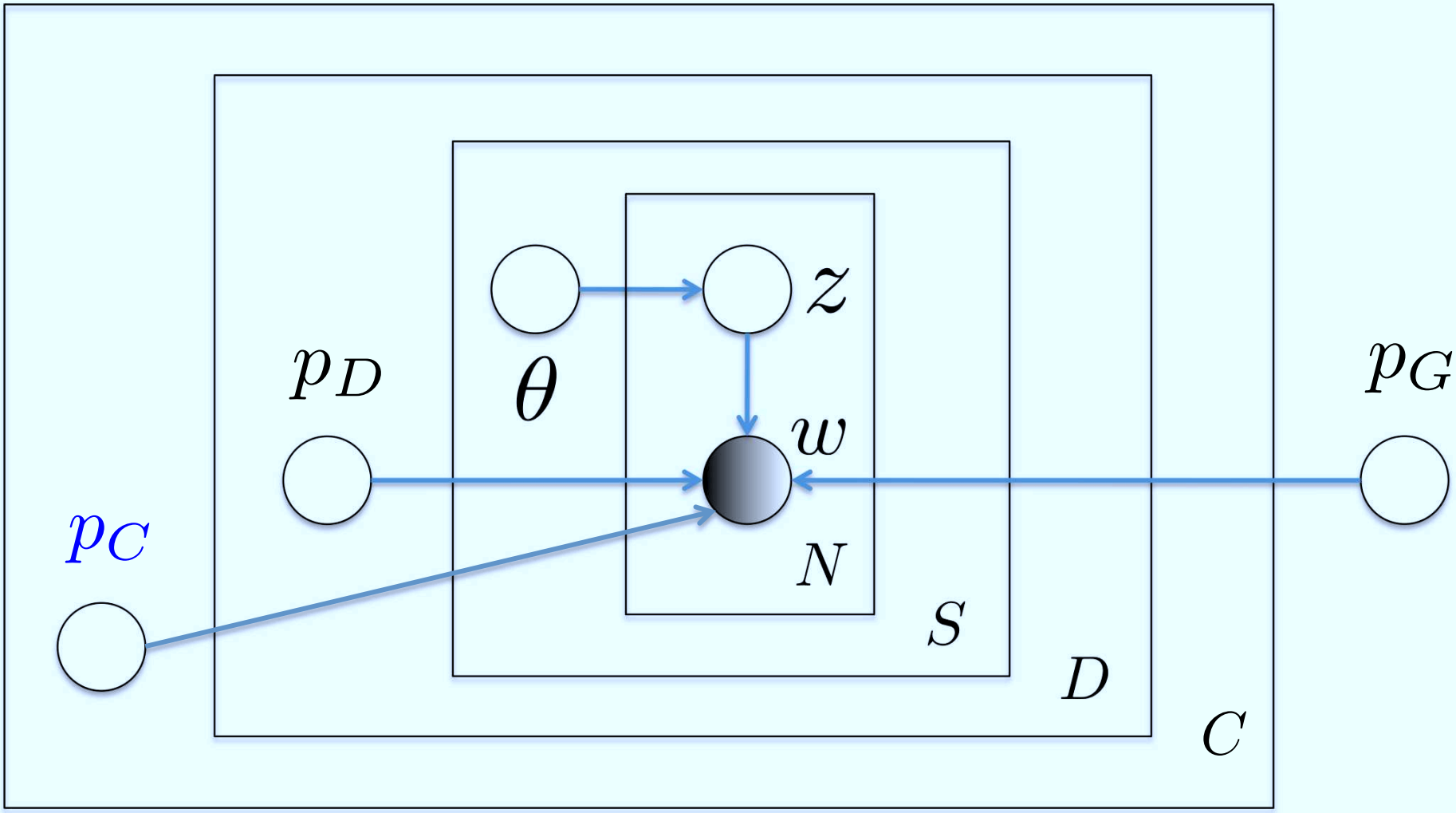
- 文書には、中心の内容と関係ない「背景ノイズ」が沢山入っているのでは?

Bungling the Easy Stuff
Hospitals are still overcharging the poor. Obamacare was supposed to fix that too. What went wrong?
One of the patients featured in the TIME cover story I wrote last March--"Bitter Pill: Why Medical Bills Are Killing Us"--was Emilia Gilbert, a school-bus driver. Gilbert was 61 years old in 2008 when she slipped and fell one evening in her backyard in Fairfield, Conn.She was taken to the emergency room at Bridgeport Hospital, where she was treated for some cuts and a broken nose. She left a few hours later with a bill for $9,418, which included $6,538 for CT scans and $239 for a routine blood test. The charges, I found, were based on something called the hospital chargemaster--a list of hugely inflated prices that no one could explain or defend. Medicare--which by law pays …

Global/Localモデル (2)

- モデルを拡張: 各文の単語は、
  (1) 背景分布
  (2) 文書特有の分布
  (3) 要約したい内容分布
  のどれかから生成されている
  - $\theta \sim \mathrm{Dir}(\alpha_1, \alpha_2, \alpha_3)$ でこの3つのどれかを選択
  - “トピック”が3つあるLDAのようなもの

Global/Localモデル (3)



- 潜在変数 z が, どれから単語を生成するかを支配

Global/Localモデル: 実験結果

- さらに要約性能が向上

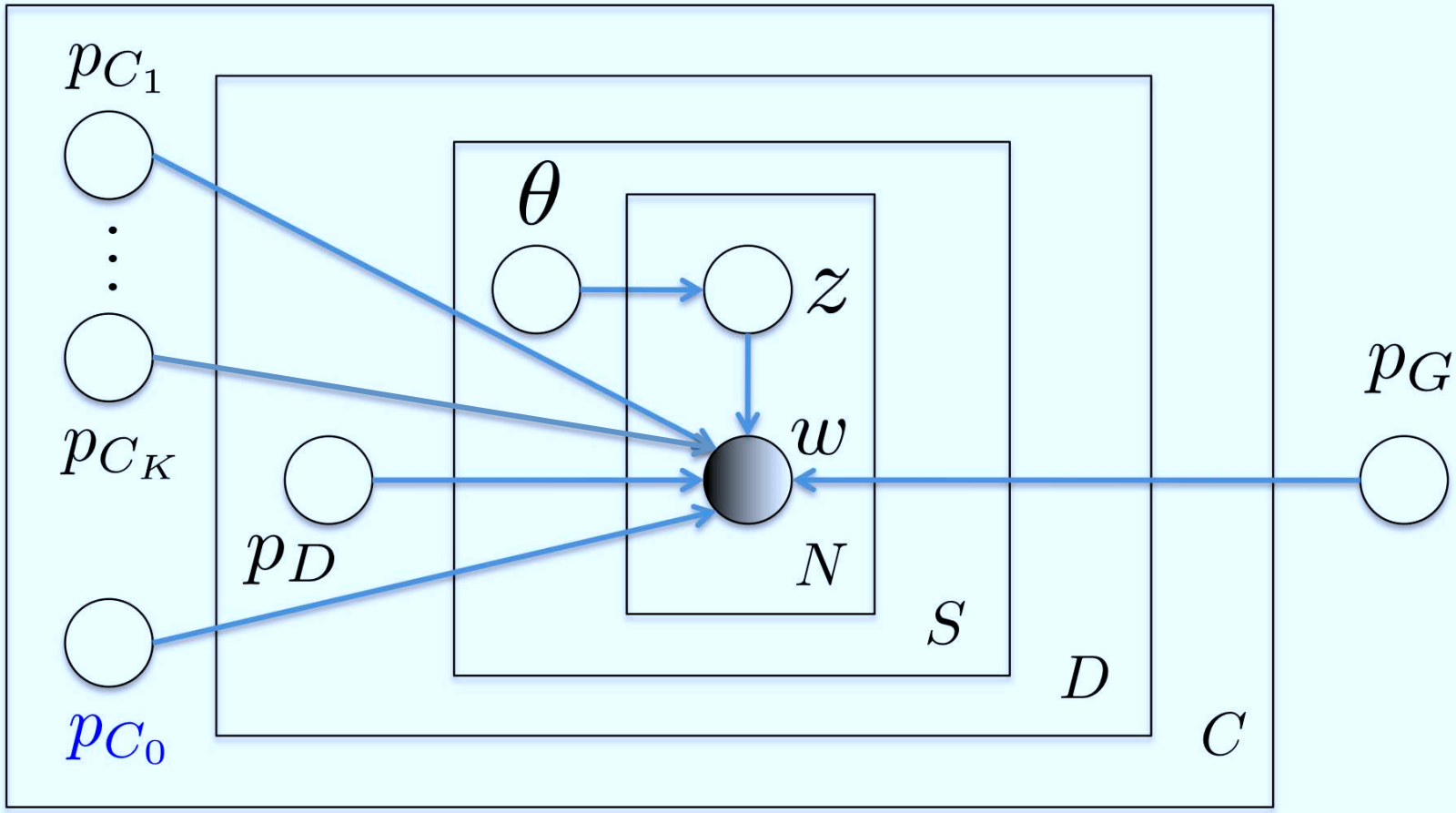
System	ROUGE -stop			ROUGE all		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4
SumBasic	29.6	5.3	8.6	36.1	7.1	12.3
KLSum	30.6	6.0	8.9	38.9	8.3	13.7
TopicSum	**31.7**	**6.3**	9.1	39.2	8.4	13.6

Global/Localモデル＋トピック

- 文書集合の内容 p_C をさらにトピックに分解
  - p_{C_0}：全体的な内容
  - p_{C_1}：トピック1
  - p_{C_2}：トピック2 ...
- コインを投げて、表が出たら p_{C_0} から、裏が出たら $p_{C_1} \cdots p_{C_K}$ のどれかから内容語を生成
  - 前の文のトピックを継続しやすいようなHMM

Global/Localモデル＋トピック

- p_C をさらにトピックに分解



Global/Localモデル+トピック: 結果

- p_{C_0} から要約文を選択すると、最高性能

System	ROUGE -stop			ROUGE all		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4
SUMBASIC	29.6	5.3	8.6	36.1	7.1	12.3
KLSUM	30.6	6.0	8.9	38.9	8.3	13.7
TOPICSUM	**31.7**	**6.3**	9.1	39.2	8.4	13.6
HIERSUM	30.5	**6.4**	9.2	40.1	8.6	**14.3**

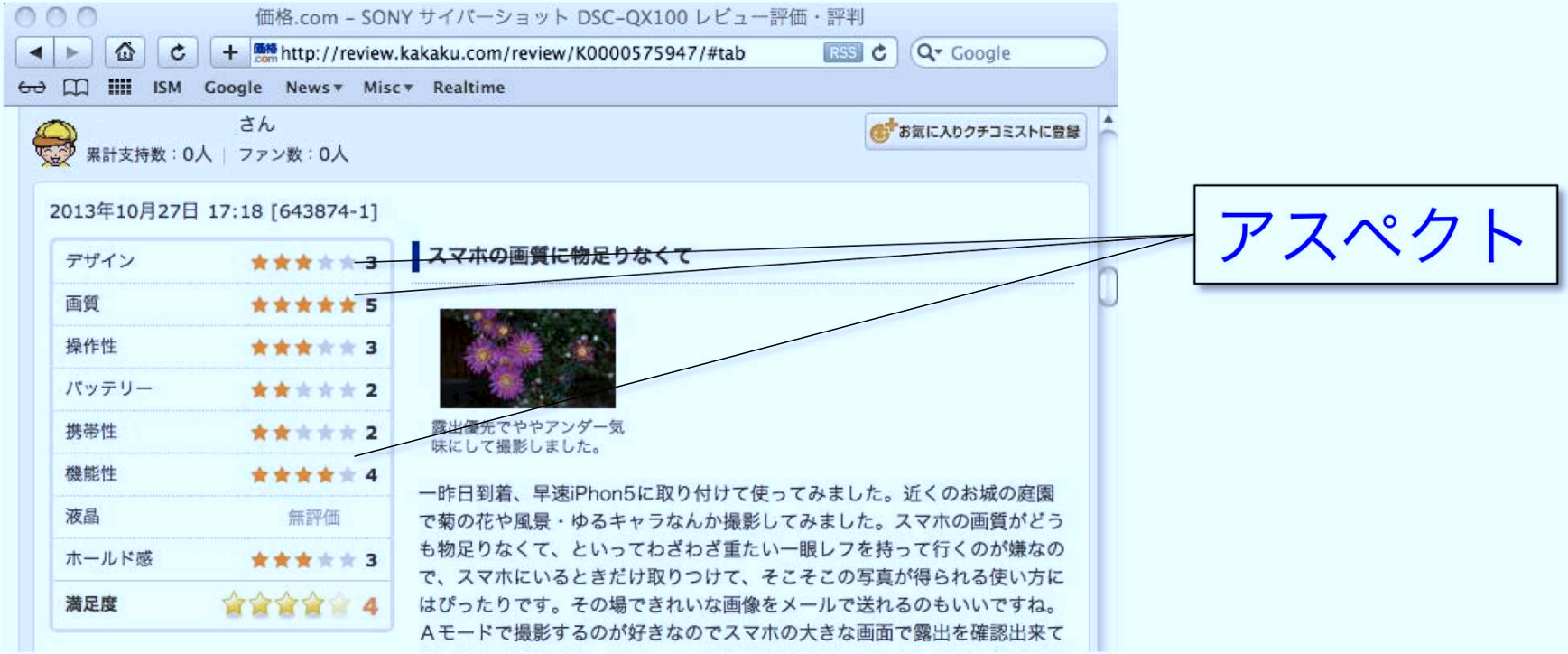
「教師データ」の利用

- これまでの話はほぼ教師なし学習
- 実際には、教師データが一部でも得られることが多い

- しかし、普通の教師あり学習は適用できない！
  - 教師データがすべて付いていることを仮定
  - 個人差やノイズがうまく扱えない
  - (教師なし学習で扱える)中身をまったく見ていない!

Titov&Mcdonald (2008)

- 背景: レビューサイトでのレビュー文には、評価ポイントごとの点がついていることが多い

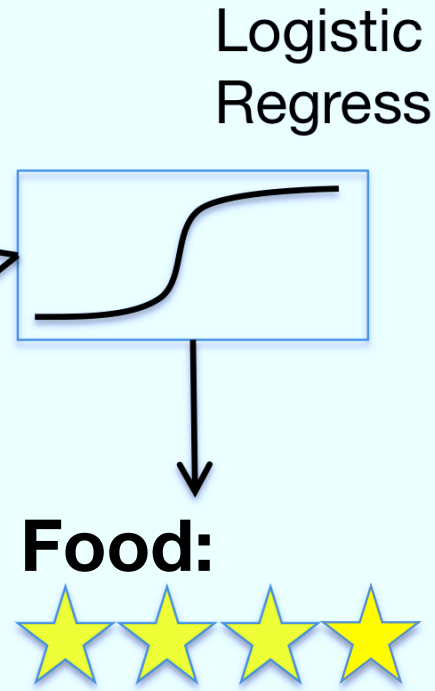


- 問題: どの評価ポイント(アスペクト)がレビュー文のどこに対応しているかわからない!
  - しかし、統計的には相関があるのでわかるはず

MAS (Multi-Aspect Sentiment model)

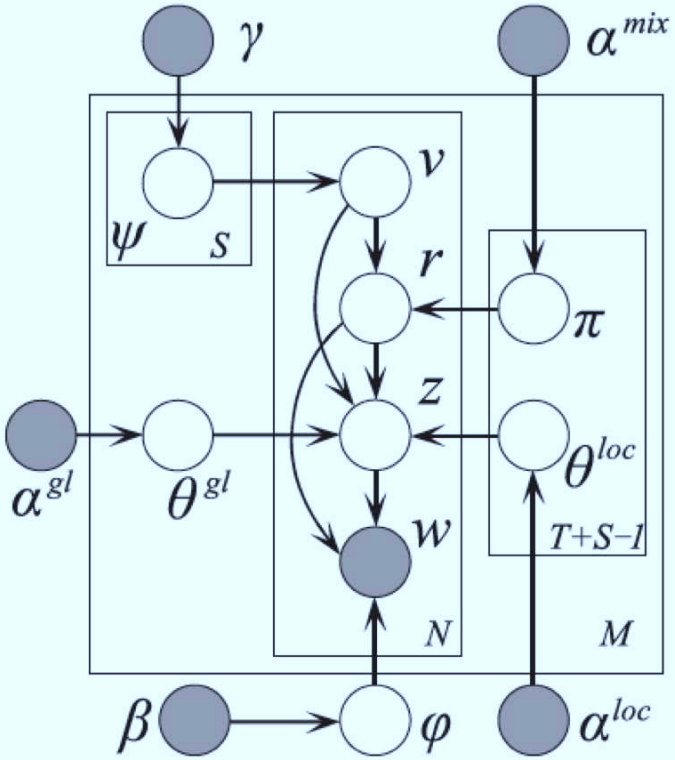
- 解決: アスペクト∈トピックとみて、
  アスペクトに割り当てられた語を使った回帰モデル
  - トピックモデル+ロジスティック回帰

This hotel has a good location and great service. Lunch is also great, especially with a café style desserts. We can reach any spots from this hotel by walk or a light rails. The most prominent feature of this hotel is its silence; it is a bit far away from the downtown. However, during our stay, we could enjoy fabulous restaurants located in this hotel. …

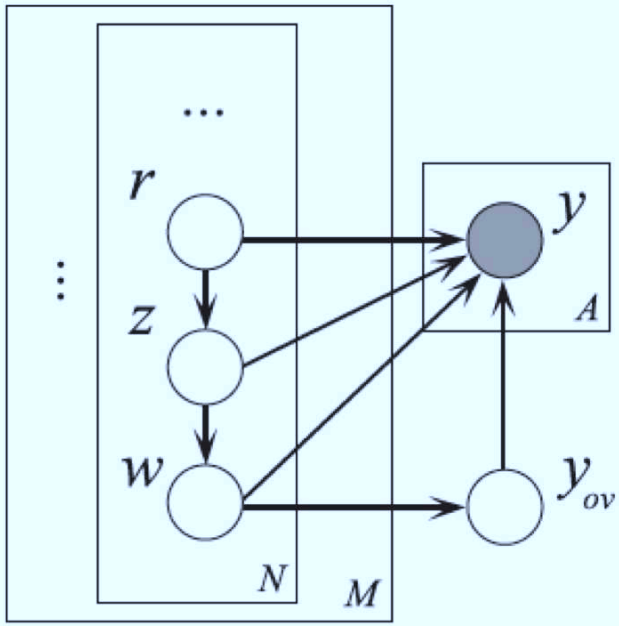


MAS (2)

全体像



回帰モデル部



nグラムfが評価yを生む重み

$$p(y^{(a)}=y|\mathbf{w},\mathbf{r},\mathbf{z}) \propto \exp\left(b_y^{(a)}+\sum_f \lambda_{f,y}+p(a|f)\lambda_{f,y}^{(a)}\right)$$

- トピックをサンプルする際にも、この重みを用いる (同時学習)

Multi-Aspect Sentiment Model

First 3 Local Topics

Service	Location	Rooms
staff	hotel	room
friendly	walk	bathroom
helpful	location	shower
service	station	bed
desk	metro	tv
concierge	walking	small
excellent	away	water
reception	right	clean
pleasant	minute	comfortable

- 10,000 reviews from TripAdvisor.com
  - service, location, rooms aspects
  - Tied first three topics to these aspects ratings
- The first three topics correspond to associated aspects!!

Varying Granularity

"... public transport in London is straightforward, the tube station is about an 8 minute walk ... or you can get a bus for £ 1.50 .... We had a stunning view (from the floor to ceiling window) of the Tower and the Thames."

- Global topic London: *London, tube, £ , Tower, Thames*
  - global topic dist is assigned to the document
- Local topics:
  - Location: *transport, station, walk, bus, minute.*
  - View: *view, window*
  - local topic dist is assigned to current sliding window

Not Associated topics in MAS

Topics not associated to the rated aspects:

Food	Pricing	Getting there	Parking	Spa	Bathroom
breakfast	$	shuttle	parking	pool	shower
free	night	bus	car	tub	water
coffee	parking	taxi	lot	hot	hot
internet	rate	ride	valet	indoor	bathroom
morning	price	train	park	swimming	towels
buffet	paid	hour	garage	outdoor	toilet
day	day	station	free	spa	tub
wine	euros	cab	street	heated	bath
nice	got	took	parked	use	pressure

複雑な依存関係の場合

- 生成モデルの分布が共役でない
  ‥‥Gibbsサンプラーでの学習が難しい
- ネットワークのように、相互依存関係のある
  データの生成モデル?
- 以下は、“Matchbox: Large Scale Online Bayesian Recommendation” (WWW 2009) を簡単に

ユーザーの情報行動

- ユーザーiがある商品jに対して、環境Φの下でレートrをつける
  - r={0,1} なら、買った/買わない、いいね/いいねなしなど
  - rが実数やランクの場合もある
- データ: ペア $\left\langle \mathbf{x}, \mathbf{y}, \Phi, r \right\rangle_{n=1}^{N}$
  - $\mathbf{x} \in \mathbb{R}^n$ ：ユーザーiの特徴ベクトル
  - $\mathbf{y} \in \mathbb{R}^m$ ：商品jの特徴ベクトル
  - $\Phi \in \mathbb{R}^f$ ：環境ベクトル
  - $r \in \mathbb{R}$ ：観測されたレート

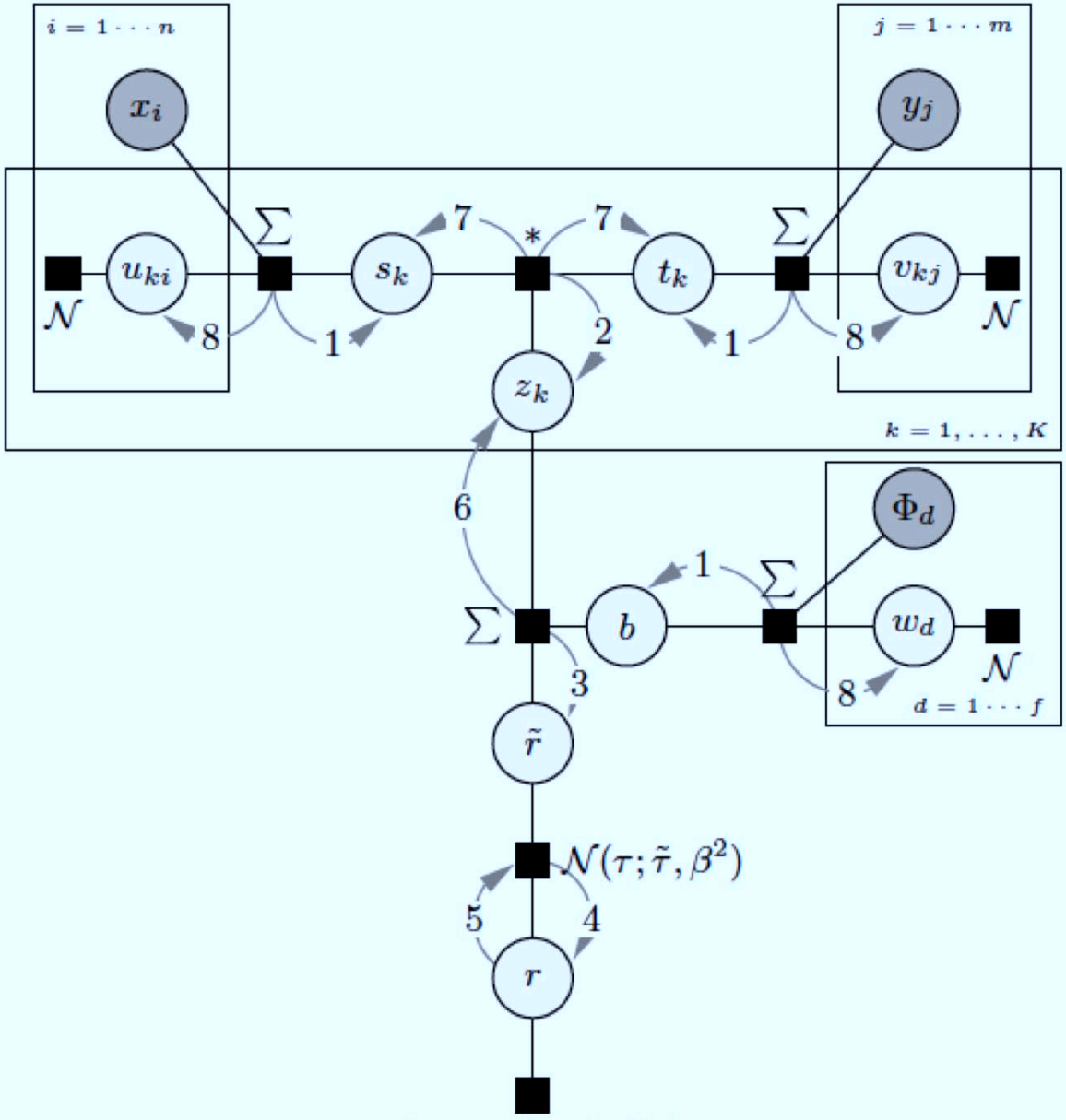
Bilinear Rating Model

- rは圧縮空間での正規分布から生成

$$
\begin{aligned}
r &\sim \mathrm{N}(\mathbf{s}^T\mathbf{t} + b, \beta^2) \\
\mathbf{s} &= U\mathbf{x}, \\
\mathbf{t} &= V\mathbf{y}, \\
b &= \Phi^T\mathbf{w}
\end{aligned}
$$

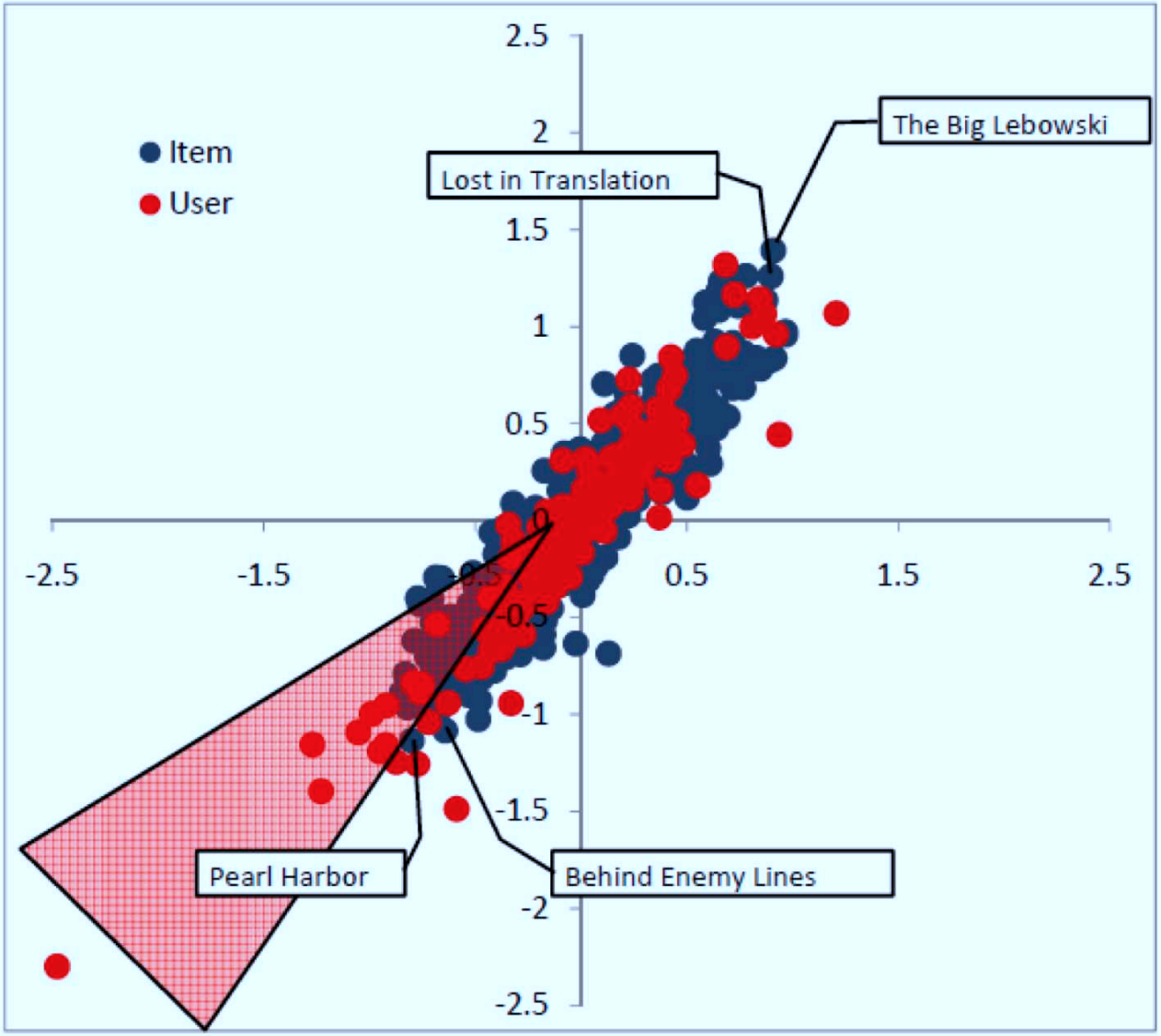
  - 推定すべきパラメータ: *U, V, w*
- 実際には、r自身が観測されているとは限らない
  - 潜在変数rにもとづくProbitモデル

Bilinear Rating Model のグラフィカルモデル



- 推論はEP (Minka 01)
- メッセージパッシング
- Conjugateでなくとも学習可能

2次元へのEmbedding



- Netflixムービーデータ
- 注: Deep Netでは、階層化しないと2次元には埋め込めない

まとめ

- 情報検索におけるグラフィカルモデル
  - モデルの仮定を整理して見るための道具
- 複雑な依存性が扱える
  - 大域的、局所的
  - 個人性とその強さ
  - “教師データ” のモデル化 (× 最適化、SVM)
- ベイズ専用のツールというわけではないが、階層的な情報を上手に扱えるのは、事実上階層ベイズモデルに限られる