(修正版, 2006.3.13)

*Topic*に基づく 統計的言語モデルの最前線

── PLSIからHDPまで ──

山本幹雄
(筑波大学)

持橋大地
(ATR)

URL= http://www.mibel.cs.tsukuba.ac.jp/~myama/pdf/topic2006.pdf

今日の話題

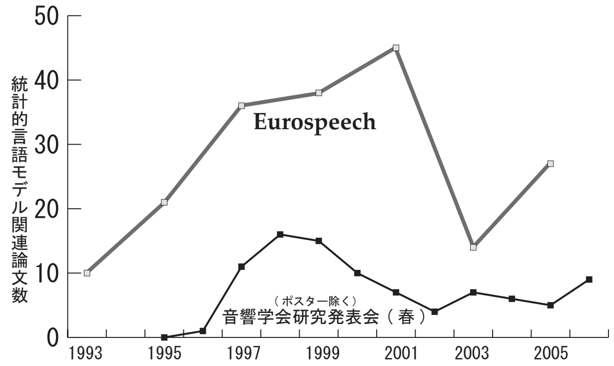
- 従来の統計的言語モデル
  - 目標：真の単語出現確率を一つだけ推定する
    - *ex.* 1999年の言語処理学会チュートリアル [山本 1999]
- トピックに基づく統計的言語モデル
  - 前提：単語出現確率は変動する (変動の要因をまとめてトピックと呼ぶ)
  - 目標：確率の変動を追跡
  - 枠組み：ベイズ統計 (事前分布→事後分布)
    - 単語出現確率 (変動) の事前分布
    - 変動のレベル (単語レベル？文書レベル？)

・局所的モデル
・文の確率

・大域的モデル
・文書の確率

UM
DM
PLSI
LDA
HDP

音声関係学会における 統計的言語モデル関連論文数



概要 1/2

- 単語出現確率の変動
- 統計的言語モデルと*n*-gramモデル
  - Noisy Channel Models
  - 統計的言語モデルとは？
  - パラメータ推定
- トピックに基づく言語モデルの概要
  - 局所的「文」モデルから大域的「文書」モデルへ
  - 枠組み：ベイズ統計学

概要 2/2

- トピックに基づく言語モデルの具体例

	ユニトピック	マルチトピック
パラメトリック	UM (1) (Unigram Mixtures)	PLSI (3) (Probabilisitic LSI)
パラメトリック ベイズ	DM (2) (Dirichlet Mixtures)	LDA (4) (Latent Dirichlet Allocation)
ノンパラメトリック ベイズ	DPM, HDP (5) (Dirichlet Process Mixtures, Hierarchical Dirichlet Process)	

※本チュートリアルではEMアルゴリズムやMCMC等の計算アルゴリズムは原則扱わない。モデルの考え方中心。

(1) (2) ...の順番

単語出現確率の変動

- トピック
- 変動の例
  - 文書の種類, 分野, 時期
  - キャッシュ (繰り返し)
  - トリガー (共起) , 3単語の共起

トピック

- 単語出現確率は（激しく）変動する [Church&Gale 1995]
  - 要因
    - 分野, 話題（Topic）, 時期
    - 文書, 章, 節, 段落
    - 文体, 著者, 想定する読者, 言語, 地方

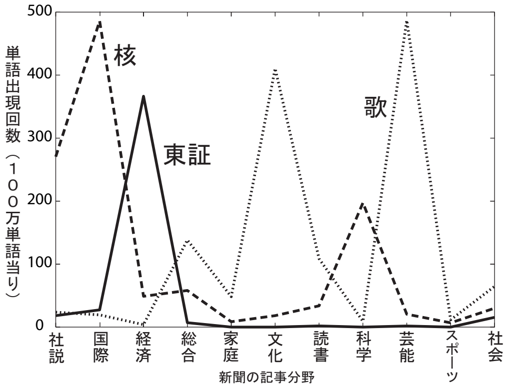
まとめて「トピック」と呼ぶことにする

出現確率の変動: *文書の種類*

‘*said*’の出現回数 [Church&Gale 1995]より引用

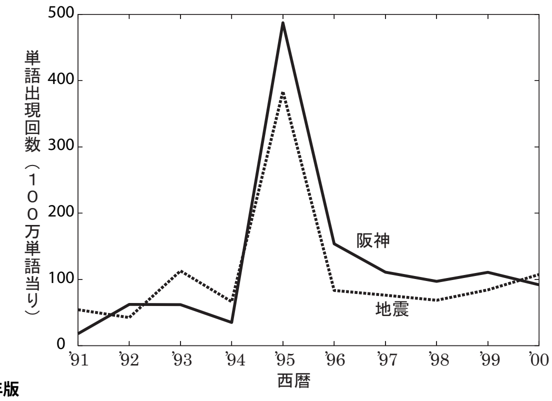
Source	*Frequency* /10^6*words*
Department of Energy Abst.	41
Groliers Encyclopedia	64
Federalist papers	287
Hansard	1072
Harper & Row Books	1632
Brown Corpus	1645
Wall Street Journal	5600
Associated Press 1990	10040

出現確率の変動：*分野*



毎日新聞 CD-ROM 2004年版

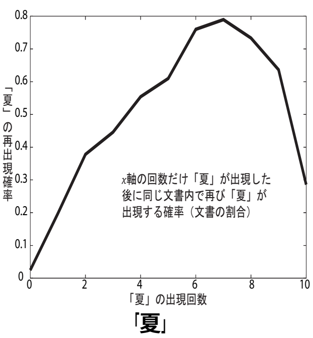
出現確率の変動：*時期*



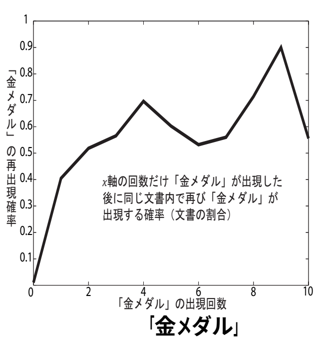
毎日新聞 CD-ROM 1991-2000年版

出現確率の変動： *キャッシュ*

同じ単語が再び出現する確率



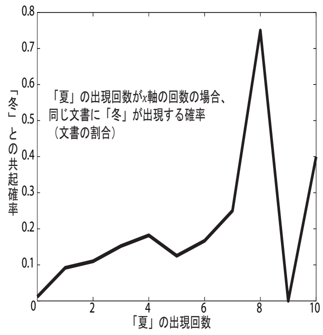
「夏」



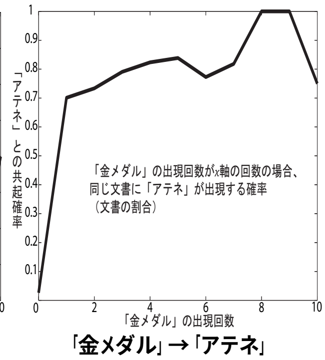
「金メダル」

毎日新聞CD-ROM 2004年版

出現確率の変動：*トリガー* or *共起*



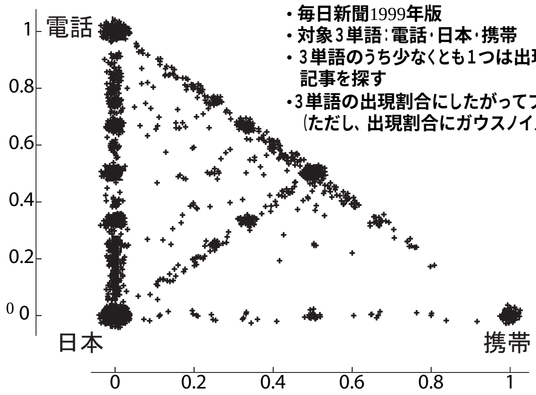
「夏」→「冬」



「金メダル」→「アテネ」

毎日新聞CD-ROM 2004年版

3単語共起: 電話 日本 携帯



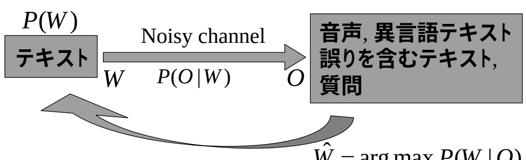
- 毎日新聞1999年版
- 対象3単語:電話・日本・携帯
- 3単語のうち少なくとも1つは出現する記事を探す
- 3単語の出現割合にしたがってプロット（ただし、出現割合にガウスノイズ付加）

統計的言語モデル & n-gramモデル

- Noisy Channel Models
- **統計的言語モデルとは？**
- **パラメータ推定**

統計的言語モデル

- **任意の文・文章** W **の確率を与えるモデル**: $P(W)$
- *Noisy Channel Model* **による音声自然言語処理の基本**
  - 音声認識, 統計的機械翻訳, 情報検索, スペルチェッカ
    [Bahl et al. 1983], [Brown 1993], [Ponte&Croft 1998], [Church&Gale1991]



$$\hat{W} = \arg\max_{W} P(W \mid O) = \arg\max_{W} P(O \mid W) P(W)$$

処理結果として文・文書を出力するシステムに使うことができる。

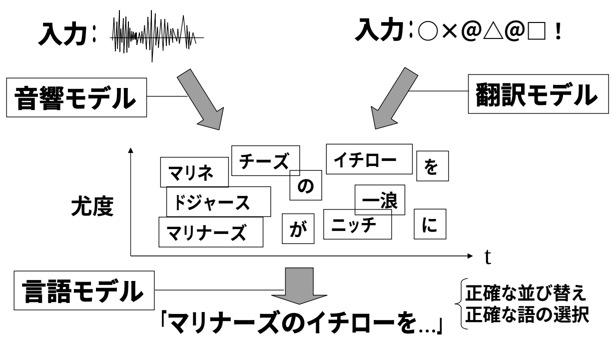
統計的言語モデルの利用

音声認識

入力:

統計的機械翻訳

入力:○×@△@□！



言語モデルが与える確率例

- **同じ文字集合を使った文 (文字列) の確率**

P (**成長の可能性はあるという**) ＝ 7.7×10^{-14}

P (**あるという成長の可能性は**) ＝ 9.4×10^{-17}

P (**ある成長というのは可能性**) ＝ 8.7×10^{-19}

P (**可能とは成長性のあるいう**) ＝ 3.6×10^{-21}

P (**のはある能性という可成長**) ＝ 2.3×10^{-33}

P (**ある能いの成う性と可長は**) ＝ 3.9×10^{-51}

モデル: 新聞記事4年分で学習した文字trigramモデル
(backoffスムージング)

n-gramモデル

w = 「**外, は, 闇, だ**」

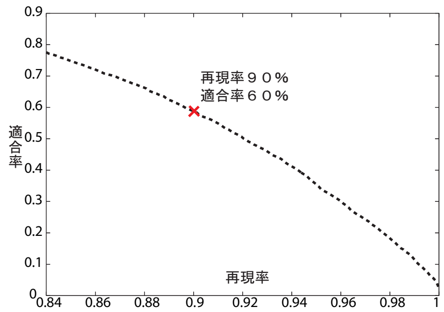
- Trigram**モデル** (文頭記号)
  P(w) = P(**外**| #, #) P(**は**| #, **外**) P(**闇**|**外**,**は**) P(**だ**| **は**,**闇**)
- Bigram**モデル**
  P(w) = P(**外**| #) P(**は**|**外**) P(**闇**|**は**) P(**だ**|**闇**)
- Unigram**モデル**
  P(w) = P(**外**) P(**は**) P(**闇**) P(**だ**) ← **多項分布**

標準的な教科書は[北 1999]

同音異義語スペルチェッカ

- 「方法を実効する」 n-gramモデル

P（〜を実行する）＞P（〜を実効する）？

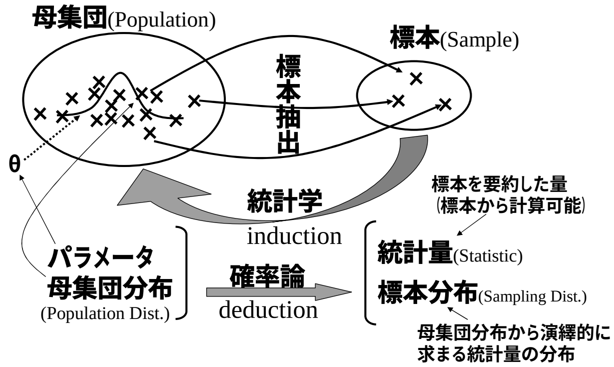


学習：
モデル：Trigramモデル
データ：毎日新聞1999年版
語彙：2万単語
同音異義語：764組1689語

テスト
データ：毎日新聞2000年版
加工：同音異義語に1%の誤りを混入.
測定性能：誤り検出率

[三品他2004]

確率論と統計学



多項分布 (Multinomial distribution)

- 一回の試行（結果はV種類：1〜V）
  - 結果1〜Vが起きる確率： $\theta=\theta_1,\theta_2,...,\theta_V$ ，ただし$\sum_{j=1}^{V}\theta_j=1$
- 多項分布（1）
  - 上記の試行をN回行った場合、どの種類の結果が何回起きたか（ $\boldsymbol{y}=y_1,y_2,...,y_V$ ）に対する確率

$$P_{Mul}(\boldsymbol{y};\boldsymbol{\theta},N)=\frac{N!}{y_1!y_2!\cdots y_V!}\theta_1^{y_1}\theta_2^{y_2}\cdots\theta_V^{y_V}$$

- 多項分布（2） ⇐ 以下、こちらを使う
  - 同様に、結果の列（ $\boldsymbol{d}=w_1,w_2,...,w_N$ ）に対する確率

$$P_{Mul}(\boldsymbol{d};\boldsymbol{\theta},N)=\theta_1^{n(\boldsymbol{d},1)}\theta_2^{n(\boldsymbol{d},2)}\cdots\theta_V^{n(\boldsymbol{d},V)}$$

$n(\boldsymbol{d},v)$：$\boldsymbol{d}$中結果vの数

多項分布＝単語出現回数の基本分布

- V種の結果が生じる確率がθ *で与えられる*試行をN回行った場合、結果の列がdとなる確率

⇩

- V種の単語の出現確率をθとしたとき、長さNの文書が $\boldsymbol{d}=w_1,w_2,...,w_N$ となる確率

$$P_{Mul}(\boldsymbol{d};\boldsymbol{\theta},N)$$

多項分布のパラメータθ＝Unigramモデル

代表的なパラメータ推定法

- モーメント法
  - 方程式： 統計量＝標本分布のモーメント（平均など）
    - 具体例： 標本平均＝母平均
- 最尤推定法
  - データは確率が最大であったから手に入った
    - ⇒ データの確率（尤度）を最大とするパラメータに決める
- ベイズ推定法
  - パラメータの事後分布＝事前分布＋データ
  - 事後確率を用いた推論（例：予測分布）

トピックに基づく言語モデルの概要

- 工学的な期待
  - 文モデルから文書モデルへ
  - 大域的情報による単語予測
- 問題の定式化
- ベイズ統計学

文モデルから文書モデルへ

- 文モデル：n-gramモデル

文書A

「一部 IT企業の不祥事が発覚した」	中確率	中確率
「株安は短期間で回復した」	中確率	高確率 ⇨ 低確率
「キャッチャーとしてメジャーに初挑戦」	高確率	

文書B

- 文書モデル —— 文書全体の整合性をモデル化

大域的情報による単語予測

- wに当てはまる単語を入れよ。

例1
(1) w
(2) の w
(3) マリナーズの w
(4) 大リーグで激しい打率争いをしている
マリナーズの w

例2
(1) 不利益を被るのは個人w
(2) 株式市場に明らかな復調の兆しは見られない。... 持ち合い株の解消売りも継続しており... 不利益を被るのは個人 w

問題の定式化

株式市場に明らかな復調の兆しは見られない。... 持ち合い株の解消売りも継続しており... 不利益を被るのは個人 w

予測

- 文脈情報 $\boldsymbol{d}$ (単語列)
- 単語の出現確率 $P(w|\boldsymbol{d})$ $\left(\sum_w P(w \mid \boldsymbol{d}) = 1\right)$
- 事前の知識

$\boldsymbol{d}$ + 事前の知識 → $P(w|\boldsymbol{d})$

解1：条件付き確率：*big*-gram

- $n$gramのnを大きく → $\boldsymbol{d}$ を直接モデル化 $P(w|\boldsymbol{d})$
  - パラメータ数 = V^n, V:語彙サイズ
  - 例：
    V= 60,000 かつ n=10 (本当は100くらいにしたい)
    $V^n = 60{,}000^{10} = \underline{6.05 \times 10^{47}}$

V=60,000でn=3くらいがいまだ限界

解2：条件付き確率 ：トリガー

[Rosenfeld 1996]

- トリガーモデル (最大エントロピーモデル)
  - $\boldsymbol{d}$を単純化する
    - wに影響を与える単語(トリガー)が$\boldsymbol{d}$中にある／なし

$$P(w \mid \boldsymbol{d}) \propto \exp\left(\sum_i \lambda_i f_i(\boldsymbol{d})\right), \quad f_i(\boldsymbol{d}) = \begin{cases} 1, & if\ 履歴\boldsymbol{d}に素性iがある, \\ 0, & others. \end{cases}$$

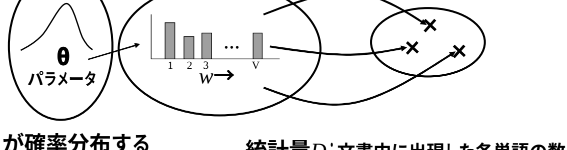
素性iの重み

・推定に必要な計算量膨大
・トリガーは2単語の長距離関係→ 3単語以上の関係は？
(2単語でも膨大な組み合わせ)
・語の多義性 (例：interest)

条件付き確率は融通がきかない
$P(w|a, b) \not\propto P(w|a) + P(w|b)$

解3：ベイズ統計学 1/2

標本：文書標本



θ が確率分布する

ベイズ統計学

統計量D：文書中に出現した各単語の数
標本分布： $P(\boldsymbol{d} \mid \theta)$ ex.多項分布

θ が確率変数だとすると、データ$\boldsymbol{d}$から θ の確率分布を導ける

ベイズの定理 $\underbrace{P(\theta \mid \boldsymbol{d})}_{事後分布} = \dfrac{\overbrace{P(\boldsymbol{d} \mid \theta)}^{尤度}\overbrace{P(\theta)}^{事前分布}}{P(\boldsymbol{d})}$ → $\int P(\boldsymbol{d} \mid \theta)P(\theta)d\theta$

解3：ベイズ統計学　2/2

$$P(\theta \mid d) = \frac{P(d \mid \theta)P(\theta)}{P(d)} \propto P(d \mid \theta)P(\theta)$$

尤度モデル　事前分布（知識）

この2つの組み合わせで
いろいろなモデルができる

P（θ|*d*）の使い方：

（1）期待値

$\hat{\theta}_i = \int \theta_i P(\theta \mid d) d\theta$

（2）最大値（MAP推定）

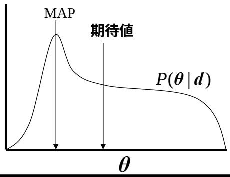
$\hat{\theta} = \max_{\theta} P(\theta \mid d)$

一点に要約

（3）予測分布

$P(d' \mid d) = \int P(d' \mid \theta) P(\theta \mid d) d\theta$

全域を利用



Bayes統計学の問題と対処

- 事前分布の設定問題（誰が事前分布を設定するのか？）
  - 経験ベイズ（Empirical Bayes）[Bradley et al. 2000]
    - データから事前分布の（ハイパー）パラメータを求める
- 積分計算問題（解析的に解けない）
  - モンテカルロ法による近似的積分[Gamerman 1997][伊庭他 2005]
    - MCMC（Markov Chain Monte Carlo）, Gibbs
  - 最適化による積分近似[Jaakkola 2000]
    - 変分ベイズ（変分近似＝変分法＋最適化）
    - EPおよびその拡張

多くの概念が
物理学から

（最近、MCMCのよさが注目され始めている）

経験ベイズ法

- 経験ベイズ法
  - 事前分布 $P(\theta;\tau)$ をデータを用いて決定する
  - データ*D*の尤度

$$\int P(D \mid \theta) P(\theta;\tau) d\theta = \int P(D,\theta;\tau) d\theta = P(D;\tau)$$

τの最尤推定が可能

$P(\theta;\tau)$ を
事前分布とする　経験ベイズ法

$\tau \xrightarrow{P(\theta;\tau)} \theta \xrightarrow{P(D \mid \theta)} D$

θを積分消去

$\tau \xrightarrow[P(D;\tau)]{} D$

Steinのパラドックス　1/2

[久保川 2004]

- 野球の打率の予測[Efron&Morris1975]
  - シーズン初期45打席の打率（最尤推定量）
  - これは最終打率の推定値としては信頼性が低い
  - もっとよい推定法はないか？
- 縮小推定法
  - 方法：他大勢の選手の45打席の打率の平均に近づける
  - 基本となる推定法を必ず改良することを理論的に保証

注意：いいかげんに
近づけてもだめ

- Steinのパラドックス
  - ある選手の打率予測に他の選手のデータを使うほうがよくなるのは、素朴に考えるとおかしくないか？

Steinのパラドックス　2/2

[Efron&Morris1975]

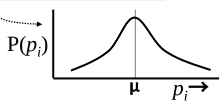
選手	ヒット	初期打率	$\hat{p}_i^S$	最終打率
1	18	.400	.290	.346
2	17	.378	.286	.298
3	16	.356	.281	.276
4	15	.333	.277	.222
5	14	.311	.273	.273
6	14	.311	.273	.270
7	13	.289	.268	.210
⋮	⋮	⋮	⋮	⋮
17	8	.178	.244	.316
18	7	.156	.239	.200

45打席

- 縮小推定量 $\hat{p}_i^S$ は経験ベイズ！

全選手の打率に関する事前分布
があるとする（ex. 正規分布）

データ（初期打率集合）で
事前分布を推定（経験ベイズ）



事前分布と各選手の初期打率を
用いて最終打率をベイズ的に予測

（これが縮小推定量になる）

具体的なトピックモデル

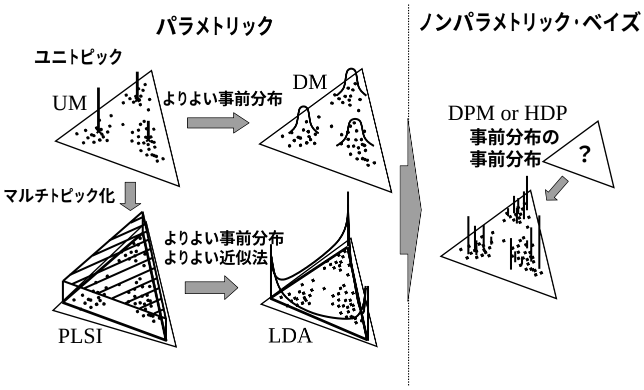
	ユニトピック	マルチトピック
混合モデル	Unigram Mixtures	Probabilistic LSI
ベイズ混合モデル	Dirichlet Mixtures	Latent Dirichlet Allocation

※混合モデルは、ベイズ混合モデルにおける事前分布を
離散分布としたモデルとみなすこともできるので、本稿
では非ベイズモデルもベイズの枠組みで扱う。

ポイント 1/2

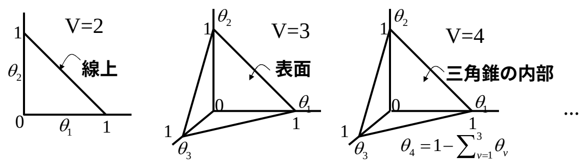
- 5つのモデルにベイズ統計の枠組みを使うと、
  - 事後分布＝事前分布＋尤度
  - 尤度は同じものを使う（多項分布（unigram））
  - 5つのモデルは事前分布が違うだけ
- 前半4つのモデル毎の説明・スライド
  - 事前分布と考え方（文書の生成）
  - 学習時：事前分布の設定→経験ベイズ
  - 利用時：
    - 文書確率
    - 事後分布の期待値（大域的文脈情報による単語予測）

ポイント 2/2



V次元単体：$\Delta(V)$

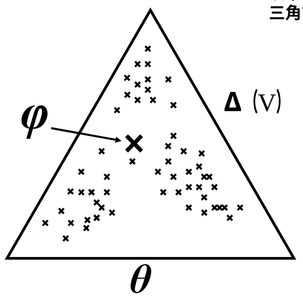
$$\Delta(V) = \{\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_V) \mid \theta_v \ge 0, \sum_{v=1}^{V} \theta_v = 1\}$$



（最後のパラメータθ_V は自由ではないので一般にはV-1次元単体という）

*uni*gramモデルのパラメータ：φ

以下、何次元であろうと単体を三角で表現することとする。

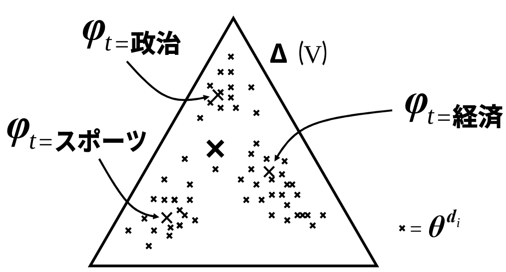


× = θ^{d_i} = 記事d_i毎のθ（相対出現頻度）

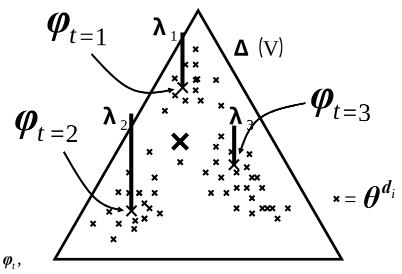
UM: Unigram Mixtures

[Iyer&Ostendorf 1999]
[Nigam et al. 2000]

トピックごとの*uni*gramモデル母数：φ_t



UMの事前分布 $P(\theta;\lambda,\varphi)=\sum_t \lambda_t \delta(\theta,\varphi_t)$



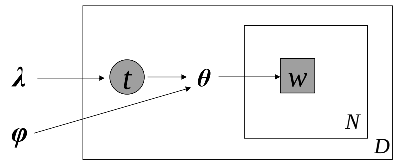
Dirac's delta

$$\delta(\theta,\varphi_t)=\begin{cases}0, & \text{if } \theta=\varphi_t,\\ \infty, & \text{other.}\end{cases}$$

$\int \delta(\theta,\varphi_t)d\theta = 1.$ （本当は各縦棒は無限大の長さ）

UMの概要

- モデルの考え方
  - (1) 確率 λ_t でトピック t が選ばれる
  - (2) トピックtのunigramモデル φ_t を使って単語wの集合である文書 d が生成される（多項分布）



○＝確率変数　□$_M$＝M回繰り返し　x→y ＝ 「yはxに依存」

経験ベイズ：UM

- 事前分布

$$P(\theta;\lambda,\varphi)=\sum_t \lambda_t \delta(\theta,\varphi_t)$$

- 各文書データの確率（尤度）

$n(d,v)$:文書d中の単語vの数

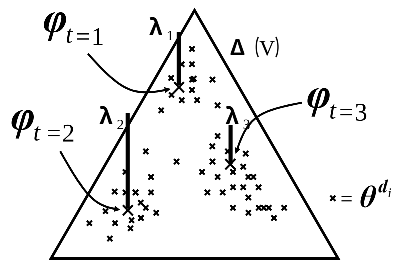
$$P(d;\lambda,\varphi)=\int P_{Mul}(d\mid\theta)P(\theta;\lambda,\varphi)d\theta$$

$$=\int\left(\prod_v \theta_v^{n(d,v)}\right)\left(\sum_t \lambda_t\delta(\theta,\varphi_t)\right)d\theta$$

$$=\sum_t \lambda_t \prod_v \varphi_{t,v}^{n(d,v)}$$

⇒ 最尤推定（EMアルゴリズム）

経験ベイズ：UM　図解



- クラスターを見つけ、クラスター中心を φ_t とする
- クラスターの大きさに応じて λ_t を決める

推論: UM

- 文書確率： $P(d;\lambda,\varphi)=\sum_t \lambda_t \prod_v \varphi_{t,v}^{n(d,v)} \approx \underline{\max_t} \lambda_t \prod_v \varphi_{t,v}^{n(d,v)}$
- 文書データの一部（履歴）hを見た後の θ の予測：

（事後分布の期待値）

$$\hat{\theta}_v=\int \theta_v P(\theta\mid h)d\theta=\int \theta_v \frac{P(h\mid\theta)P(\theta)}{P(h)}d\theta$$

$$=\frac{\sum_t \lambda_t \varphi_{t,v}\prod_{v'}\varphi_{t,v'}^{n(h,v')}}{\sum_t \lambda_t \prod_{v'}\varphi_{t,v'}^{n(h,v')}}$$

$$=\sum_t w_t \varphi_{t,v}$$

$$\approx \underline{\varphi_{t',v}}$$

$$w_t=\frac{\lambda_t\prod_v \varphi_{t,v}^{n(h,v)}}{\sum_t \lambda_t \prod_v \varphi_{t,v}^{n(h,v)}}$$

tによって何桁も値が違う

⇓

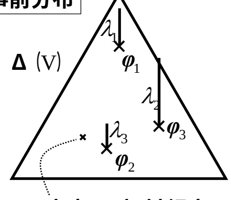
あるt'の$w_{t'}$のみ約1で他のw_tは約0

$$t'=\arg\max_t w_t$$

ユニトピックモデルだから当然

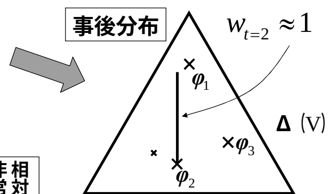
図解：文書確率・単語予測

事前分布



文書dの相対頻度

事後分布



$$P(d;\lambda,\varphi)=\lambda_1 P_{Mul}(d\mid\varphi_1)+\lambda_2 P_{Mul}(d\mid\varphi_2)+\lambda_3 P_{Mul}(d\mid\varphi_3)$$

相対的に非常に小さい

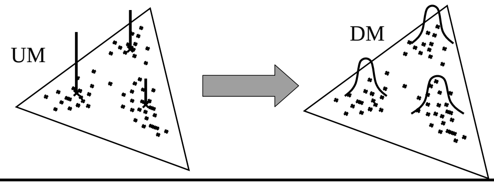
$$\hat{\theta}_v=\sum_t w_t\varphi_{t,v}\approx\varphi_{t=2,v}$$

DM : Dirichlet Mixtures

[貞光他 2005]

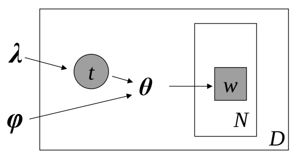
Unigram Mixturesの拡張1

- UMの事前分布
  - φ_t ：各トピック下の単語出現確率は一定と仮定
    - 実際には点ではなく点のまわりにばらつく
  - φ_t 〜ディリクレ分布 ： 混合ディリクレ分布
    Dirichlet Mixtures (DM)



UMとDM

- UM

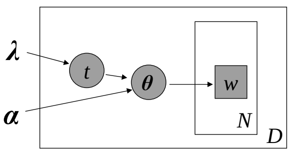


(事前分布＝有限離散分布)

文書レベルの<u>有限</u>混合モデル

$$P(\boldsymbol{d};\boldsymbol{\lambda},\boldsymbol{\varphi}) = \sum_{t=1}^{T}\lambda_t \prod_{v=1}^{V}\varphi_{t,v}^{n(\boldsymbol{d},v)}$$

- DM



(事前分布＝連続分布)

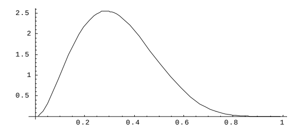
文書レベルの<u>無限</u>混合モデル

$$P(\boldsymbol{d};\boldsymbol{\alpha},\boldsymbol{\lambda}) = \int P_{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})\sum_{t=1}^{T}\lambda_t \prod_{v=1}^{V}\theta_{t,v}^{n(\boldsymbol{d},v)}d\boldsymbol{\theta} = \sum_{t=1}^{T}\lambda_t \int P_{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})\prod_{v=1}^{V}\theta_{t,v}^{n(\boldsymbol{d},v)}d\boldsymbol{\theta}$$

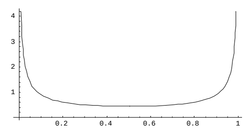
Dirichlet分布

$$P_{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha}) = \frac{\Gamma(\sum_v \alpha_v)}{\prod_v \Gamma(\alpha_v)}\prod_{v=1}^{V}\theta_v^{\alpha_v - 1}$$

2変数 (ベータ分布)：

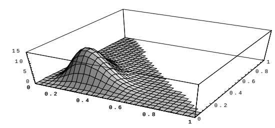


$\alpha_1 = 3, \alpha_2 = 6$

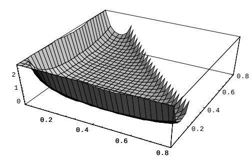


$\alpha_1 = 0.3, \alpha_2 = 0.3$

3変数：



$\alpha_1 = 4, \alpha_2 = 4, \alpha_3 = 8$



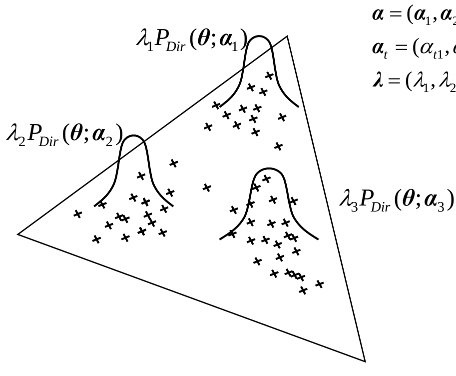
$\alpha_1 = 0.3, \alpha_2 = 0.3, \alpha_3 = 0.3$

DMの事前分布： $P(\boldsymbol{\theta};\boldsymbol{\lambda},\boldsymbol{\alpha}) = \sum_t \lambda_t P_{Dir}(\boldsymbol{\theta};\boldsymbol{\alpha}_t)$

$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_T)$

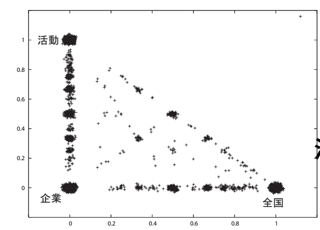
$\boldsymbol{\alpha}_t = (\alpha_{t1}, \alpha_{t2}, \ldots, \alpha_{tV})$

$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_T)$

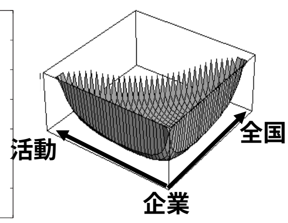


<u>実際に推定すると文書モデルとしては上に凸でなく下に凸の場合が多い</u>

分布例とディリクレ分布



'+'：記事中の相対頻度
(ノイズを加えてある)



ディリクレ分布

経験ベイズ：DM 1/2

各データ（文書）dの確率（尤度）：

$$P(\boldsymbol{d};\boldsymbol{\alpha},\boldsymbol{\lambda}) = \sum_{t=1}^{T}\lambda_t \int P_{Dir}(\boldsymbol{\theta}\mid\boldsymbol{\alpha}_t)P_{Mul}(\boldsymbol{d}\mid\boldsymbol{\theta})d\boldsymbol{\theta}$$
$$= \sum_{t=1}^{T}\lambda_t \int P_{Dir}(\boldsymbol{\theta}\mid\boldsymbol{\alpha}'_t(\boldsymbol{\alpha}_t,\boldsymbol{d}))d\boldsymbol{\theta}$$
$$= \sum_{t=1}^{T}\lambda_t P_{Polya}(\boldsymbol{d};\boldsymbol{\alpha}_t)$$

$$P_{Polya}(\boldsymbol{d};\boldsymbol{\alpha}_t) = \frac{\Gamma(s_t)}{\Gamma(s_t+|\boldsymbol{d}|)}\prod_{v=1}^{V}\frac{\Gamma(\alpha_{tv}+n(\boldsymbol{d},v))}{\Gamma(\alpha_{tv})}$$

$$\left.\begin{array}{l}\text{・Polya分布}\\ \text{・多変量負の超幾何分布}\\ \text{・Dirichlet-Multinomial分布}\end{array}\right\}\text{いろんな名前}$$

$s_t = \sum_v \alpha_{tv}$
$|\boldsymbol{d}| = \sum_v n(\boldsymbol{d},v)$

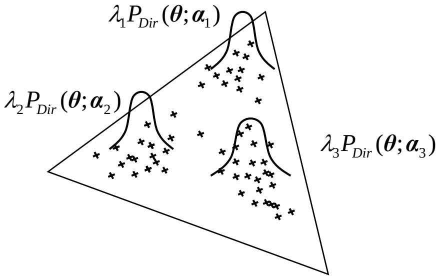
経験ベイズ：DM 2/2

- **データ**（文書集合） $\boldsymbol{D} = \boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_D$ **の尤度の最大化**

$$P(\boldsymbol{D}\mid\boldsymbol{\alpha},\boldsymbol{\lambda}) = \prod_i \left\{\sum_t \lambda_t P_{Polya}(\boldsymbol{d}_i;\boldsymbol{\alpha}_t)\right\}$$

- **アルゴリズム**
  - [Sjolander et al. 1996] → **ニュートン法＋**EM
    - （尤度はPolya分布を使っていない）
    - 収束しない場合がある，計算時間が膨大
  - [貞光他2005][Minka 2003]
    → *leaving-one-out*尤度 [Ney el al. 1997] [Minka 2003]
    ＋ニュートン法の改良[Minka 2003]＋EM
    - 収束する，高速

経験ベイズ：DM 図解



- クラスターを見つけ、クラスターをパラメータ $\boldsymbol{\alpha}_t = (\alpha_{t1}, \alpha_{t2}, \ldots, \alpha_{tV})$ のPolya分布でモデル化
- クラスターの大きさに応じて λ_t を決める

推論: DM

- **文書確率**

$$P(\boldsymbol{d};\boldsymbol{\alpha},\boldsymbol{\lambda}) = \sum_t \lambda_t P_{Polya}(\boldsymbol{d};\boldsymbol{\alpha}_t) \approx \underline{\max_t}\, \lambda_t P_{Polya}(\boldsymbol{d};\boldsymbol{\alpha}_t)$$

- **文書データの一部（履歴）hを見た後のθの予測（事後分布の期待値）**

$$\hat{\theta}_v = \int \theta_v P(\boldsymbol{\theta}\mid\boldsymbol{h})d\boldsymbol{\theta} = \int \theta_v \frac{P(\boldsymbol{h}\mid\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int P(\boldsymbol{h}\mid\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta}}d\boldsymbol{\theta}$$

$$= \frac{\sum_t C_t \dfrac{\alpha_{tv}+n(\boldsymbol{h},v)}{s_t+|\boldsymbol{h}|}}{\sum_t C_t}$$

$C_t = \lambda_t P_{Polya}(\boldsymbol{h};\boldsymbol{\alpha}_t)$
$s_t = \sum_v \alpha_{tv}$
$|\boldsymbol{h}| = \sum_v n(\boldsymbol{h},v)$

DMはCacheモデル!

$$\hat{\theta}_v = \frac{\sum_t C_t \dfrac{\alpha_{tv}+n(\boldsymbol{h},v)}{s_t+|\boldsymbol{h}|}}{\sum_t C_t} \approx \frac{\alpha_{t'v}+n(\boldsymbol{h},v)}{s_{t'}+|\boldsymbol{h}|}$$

UMと同じ話

ただし $t' = \arg\max_t C_t$

DMはCacheモデル！

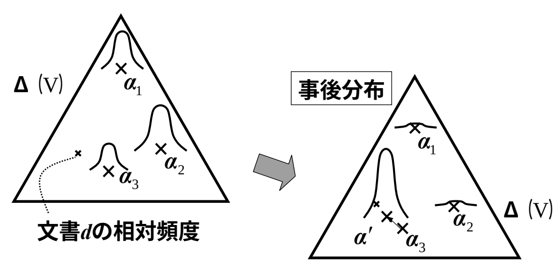
特徴

履歴中に単語vが出ればでるほど（$n(\boldsymbol{h},v)$）、$\hat{\theta}_v$は大！
$\alpha_{t'v}$ はあるトピックt'のもとで、単語vの履歴中の頻度$n(\boldsymbol{h},v)$に従い単語vの確率をどの程度に変更すればよいかを制御するパラメータと解釈できる。
（後半の実験結果を参照。）

弱点

ユニトピック：　単語の共起性についてはトピックの識別により離散的にモデル化しているのみ。h中の他の単語からは直接の影響は受けない。

図解：予測



$$\hat{\theta}_v \approx \frac{\alpha_{t'v}+n(\boldsymbol{h},v)}{s_{t'}+|\boldsymbol{h}|}$$

DMがモデル化したトピック

1999年毎日新聞 98211記事, 20k単語, 10混合

普通のunigram確率から変化の大きかった上位10単語

次元1：遊ゴロ, 逃げ切っ, 逃げ切り, 勝ち進ん, 乱調, ストレート勝ち, 加点, 和喜, 拙攻, 侮れ
次元2：公益社, 喪主, 告別, 斎場, 高生, 心不全, 貞二郎, 肝不全, 数次, じん不全
次元3：現住, 西署, 致死傷, 有印, 焼死体, 銃刀, 不定, 署, 致傷, 軽傷
次元4：デイキャッチ, 日本原子力発電, ジェー, 産科, シー, アミロイド, インフォームド・コンセント, 微量, 内臓
次元5： kyouiku, 本欄, このごろ, 悲しく, 寂しく, お宅, 共働き, 題字, 大笑い, なー
次元6：見識, 指弾, 論じる, イデオロギー, 断じて, 角栄, 変えれ, 曲がり角, 良識, 論調
次元7：日産ディーゼル工業, 続伸, 堅調, フォード・モーター, 日本興行銀行, 銀行株, 全日本空輸, 帝国データバンク, ホールディングス, 克信
次元8： 配役, 滅ぼさ, 描き出す, 好演, 悪党, 定休, 詩情, 役柄, 情感, 筆到
次元9：鉄三, 武法, 裕久, 行彦, 喜朗, 弘治, 民輔, 政審, 要一, 孝弘
次元10：シャナナ, インタファクス通信, 全欧, スカルノプトリ, 新華社通信, 北大西洋, タス通信, 弾道弾, ユーゴスラビア, タルボット

PLSI: Probabilistic LSI

LSI → Latent Semantic Indexing
（特異値分解による次元圧縮）

（or PLSA = Probabilistic Latent Semantic Analysis）

[Hofmann 1999]
[Gildea&Hofmann 1999]

Unigram Mixturesの拡張2

- Unigram Mixtures = **ユニトピックモデル**
  - **文書ごとに確率λ$_t$でトピックを選ぶ**

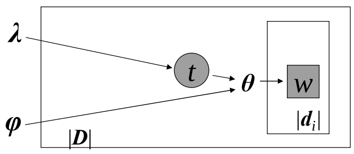
$$\Rightarrow \quad P(\boldsymbol{d};\boldsymbol{\lambda},\boldsymbol{\varphi}) = \sum_{t=1}^{T} \lambda_t \prod_{v=1}^{V} \varphi_{t,v}^{n(\boldsymbol{d},v)}$$

- **マルチトピックモデル →** PLSI, LDA
  - **文書ごとにトピック確率 λ を選び**
  - **単語ごとに λ_t でトピックが選ばれる**

$$\Rightarrow \quad P(\boldsymbol{d};\boldsymbol{\lambda},\boldsymbol{\varphi}) = \prod_{v=1}^{V} \left\{ \sum_{t=1}^{T} \lambda_t \varphi_{t,v} \right\}^{n(\boldsymbol{d},v)}$$

UM と PLSI 1/2

- UM

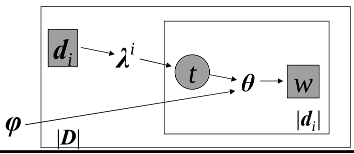


文書レベルの混合モデル

$$P(\boldsymbol{d}_i;\boldsymbol{\lambda},\boldsymbol{\varphi}) = \sum_{t=1}^{T} \lambda_t \prod_{v=1}^{V} \varphi_{t,v}^{n(\boldsymbol{d}_i,v)}$$

入れ替える

- PLSI



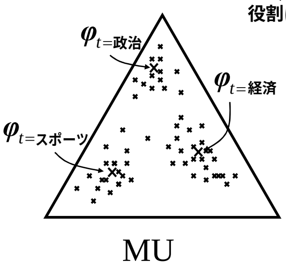
単語レベルの混合モデル

$$P(\boldsymbol{d}_i;\boldsymbol{\lambda}^i,\boldsymbol{\varphi}) = \prod_{v=1}^{V} \left\{ \sum_{t=1}^{T} \lambda_t^i \varphi_{t,v} \right\}^{n(\boldsymbol{d}_i,v)}$$

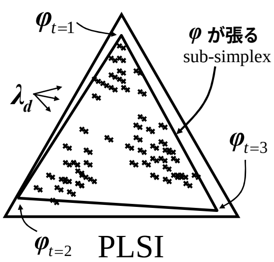
UM と PLSI 2/2

$\boldsymbol{\varphi}_t$ ：クラスター中心　　　　$\boldsymbol{\varphi}_t$ ：sub-simplexのエッジ

役割は異なる



MU



PLSI

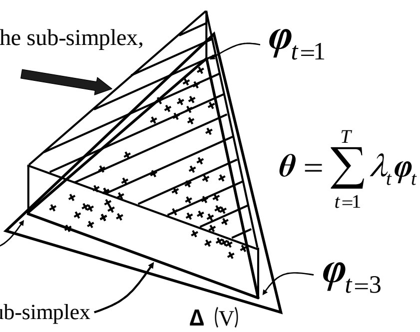
$$\prod_{v=1}^{V} \left\{ \sum_{t=1}^{T} \lambda_t^i \varphi_{t,v} \right\}^{n(\boldsymbol{d}_i,v)}$$

PLSIモデルの事前分布

[Girolami&Kaban 2003]

$$P(\boldsymbol{\theta}) = \begin{cases} c, & \text{if } \boldsymbol{\theta} \text{ is on the sub-simplex,} \\ 0, & \text{others.} \end{cases}$$

$(P(\boldsymbol{\lambda}) = c)$



$$\boldsymbol{\theta} = \sum_{t=1}^{T} \lambda_t \boldsymbol{\varphi}_t$$

経験ベイズ: PLSI 1/2

$$P(\boldsymbol{D};\boldsymbol{\varphi}) = \prod_i \int P(\boldsymbol{d}_i, \boldsymbol{\lambda}; \boldsymbol{\varphi}) d\boldsymbol{\lambda}$$

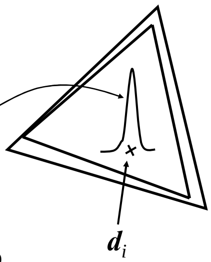
定数 c

$$= \prod_i \int \underline{P_{Mul}(\boldsymbol{d}_i \mid \boldsymbol{\lambda}; \boldsymbol{\varphi}) P(\boldsymbol{\lambda})} d\boldsymbol{\lambda}$$

≈ 積分をmaxで近似

$$\cong \prod_i \underline{c \max_{\boldsymbol{\lambda}} P_{Mul}(\boldsymbol{d}_i \mid \boldsymbol{\lambda}, \boldsymbol{\varphi})}$$

$$= \prod_i c \max_{\boldsymbol{\lambda}} \prod_{v=1}^{V} \left(\sum_t \lambda_t \varphi_{t,v} \right)^{n(\boldsymbol{d}_i, v)}$$



経験ベイズ: PLSI 2/2

$$P(\boldsymbol{D};\boldsymbol{\varphi}) \cong \prod_i c \max_{\boldsymbol{\lambda}} \prod_{v=1}^{V} \left(\sum_t \lambda_t \varphi_{t,v} \right)^{n(\boldsymbol{d}_i, v)}$$

$$= \max_{\boldsymbol{\lambda}^D} \prod_i c \prod_{v=1}^{V} \left(\sum_t \lambda_t^i \varphi_{t,v} \right)^{n(\boldsymbol{d}_i, v)}$$

PLSIの最尤推定

$$\boldsymbol{\lambda}^D = (\boldsymbol{\lambda}^{i=1}, \boldsymbol{\lambda}^{i=2}, ..., \boldsymbol{\lambda}^{i=D})$$

$$\hat{\boldsymbol{\varphi}} = \arg\max_{\boldsymbol{\varphi}} P(\boldsymbol{D};\boldsymbol{\varphi})$$

$$= \arg\max_{\boldsymbol{\varphi}, \boldsymbol{\lambda}^D} \prod_i \prod_{v=1}^{V} \left(\sum_t \lambda_t^i \varphi_{t,v} \right)^{n(\boldsymbol{d}_i, v)}$$

⇨ EMアルゴリズム

弱点: データ数に比例して λ^D (パラメータの一部) が多くなる

⇨ 過適応

推論: PLSI

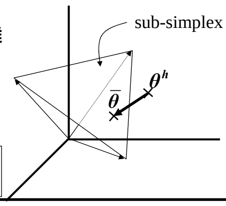
- 文書確率
  - 基本的に計算できない (LDAと同様な計算が必要)
- 文書データの一部 (履歴) $\boldsymbol{h}$を見た後の θ の予測 $\bar{\theta}$

$$\bar{\boldsymbol{\lambda}} = (\lambda_1, \lambda_2, ..., \lambda_T) = \arg\max_{\boldsymbol{\lambda}} \prod_{v=1}^{V} \left(\sum_t \lambda_t \varphi_{t,v} \right)^{n(\boldsymbol{h}, v)}$$

$$\bar{\theta}_v = \sum_t \bar{\lambda}_t \varphi_{t,v}$$

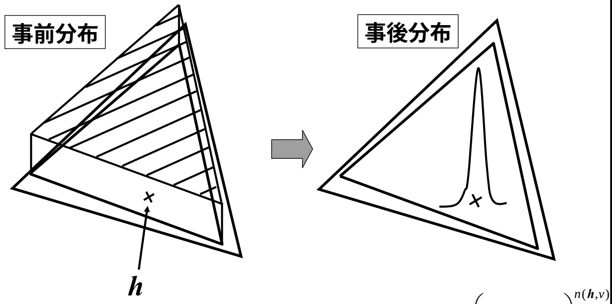
これは θ^h からKL距離最小となる*sub-simplex*の点へマップする $\boldsymbol{\lambda}$を求めていることと等しい

[Hofmann 1999]



弱点 sub-simplex上にも起こりやすい点とそうでない点があることを考慮していない。

図解: 予測



$$\bar{\boldsymbol{\lambda}} = (\lambda_1, \lambda_2, ..., \lambda_T) = \arg\max_{\boldsymbol{\lambda}} \prod_{v=1}^{V} \left(\sum_t \lambda_t \varphi_{t,v} \right)^{n(\boldsymbol{h}, v)}$$

$$\bar{\theta}_v = \sum_t \bar{\lambda}_t \varphi_{t,v}$$

PLSIがモデル化したトピック

1999年毎日新聞 98211記事, 20k単語, 10混合

普通のunigram確率から変化の大きかった上位10単語

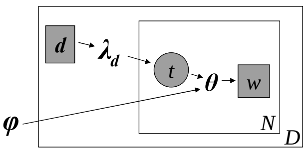
次元1:アメリカンフットボール, 新庄, ラグビー, 準々, 終盤, オールスター, 競技, 天皇杯, 持ち味, 球団
次元2:申し込み, ホール, 遺志, はがき , 消印, 会館, 告別, 喪主, ワッハ, 公益社
次元3:此花, 失跡, 課, 供述, 罪, 不定, 検察, 起訴, 無罪, 同署
次元4:摘出, 体内. 感染, 物質, アルツハイマー, 耐性, 脳波, 精巣, 胚, 卵子
次元5:あんた, 冷たい, ママ, 祖母, ええ, 先日, なあ, 蒙, 通有, あたし
次元6:覆わ, 両側, 運行, シャトル, ヘクタール, ウオーク, 水中, 海中, レジャー, 売店
次元7:エレクトロニクス, 三和, 年度末, 会計, 既存, 顧客, 住友銀行, 住友, 還付, 含み損
次元8:エンターテインメント, 主題歌, スタジオ, 堪能, 美し, 旬, コンセプト, いやす, スティーブン, ステージ
次元9:閣議, 正副, 再選, 選, 自民党, 官房, 党内, 自由党, 民輔, 府連
次元10:米兵, 朝鮮民主主義人民共和国, 平和, 任務, 断固たる, 紛争, ミサイル, 対空, タルボット, モンテネグロ

LDA: Latent Dirichlet Allocation

[Blei et al. 2003]

PLSI と LDA

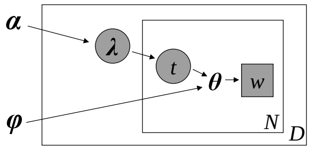
- PLSI



・文書ごとにトピック確率が決まっている
・単語ごとにトピックを選ぶ

$$P(d_i;\lambda_d,\varphi)=\prod_{v=1}^{V}\left\{\sum_{t=1}^{T}\lambda_t^i\varphi_{t,v}\right\}^{n(d_i,v)}$$

- LDA

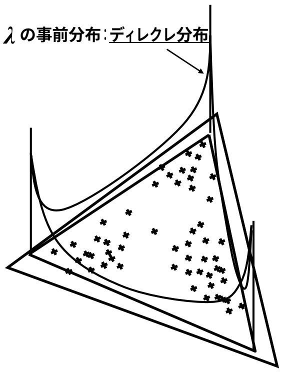


・文書ごとにトピック確率を確率的に決める
・単語ごとにトピックを選ぶ

$$P(d_i;\alpha,\varphi)=\int \underline{P_{Dir}(\lambda;\alpha)}\prod_{v=1}^{V}\left\{\sum_{t=1}^{T}\underline{\lambda_t}\varphi_{t,v}\right\}^{n(d_i,v)}d\lambda$$

LDA: λ の事前分布

λ の事前分布：ディィレクレ分布



文書d_i毎に λ^i を設定しない

ロバストなモデル

$$P_{PLSI}(d_i;\lambda^i,\varphi)=\prod_{v=1}^{V}\left\{\sum_{t=1}^{T}\lambda_t^i\varphi_{t,v}\right\}^{n(d_i,v)}$$

$$P(d_i;\alpha,\varphi)=\int P_{PLSI}(d_i\,|\,\lambda,\varphi)P_{Dir}(\lambda;\alpha)d\lambda$$

EMアルゴリズムも使えない

変分ベイズ

[Blei et al. 2003] [上田 2002] [上田2003]

MCMC

推論: LDA

- 文書確率：変分近似が必要

$$P(d_i;\alpha,\varphi)=\int P_{PLSI}(d_i\,|\,\lambda,\varphi)P_{Dir}(\lambda;\alpha)d\lambda$$ [Blei et al. 2003]

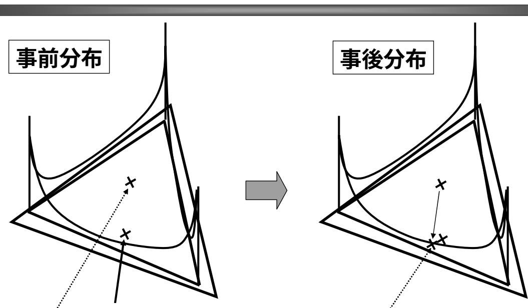
- 文書データの一部 (履歴) hを見た後の θ の予測
  - λ の事後分布の変分分布による近似 $P(\lambda\,|\,h;\alpha,\varphi)\approx P_{Dir}(\lambda;\gamma)$ (変分近似で最適近似となる γ を求める)
  - よって、事後分布の変分近似後に以下のように θ を求める

$$\hat{\theta}_v=\int\theta_v P(\theta\,|\,h)d\theta=\int\left(\sum_t\lambda_t\varphi_{t,v}\right)P_{Dir}(\lambda\,|\,\gamma)d\lambda$$

$$=\sum_t\left(\int\lambda_t P_{Dir}(\lambda\,|\,\gamma)d\lambda\right)\varphi_{t,v}=\underline{\sum_t\frac{\gamma_t}{\sum_{t'}\gamma_{t'}}\varphi_{t,v}}$$

[Blei et al. 2003]
[三品&山本2004]

図解：予測

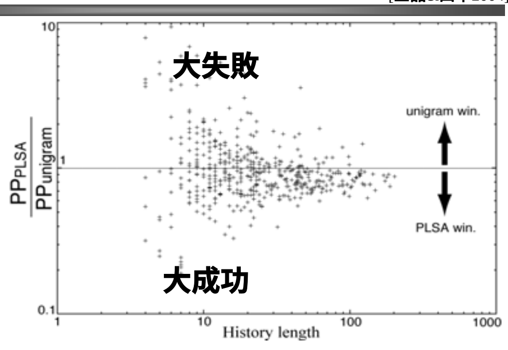


$$\hat{\theta}_v\approx\sum_t\frac{\gamma_t}{\sum_{t'}\gamma_{t'}}\varphi_{t,v}$$

PLSIによる言語モデル

[三品&山本2004]

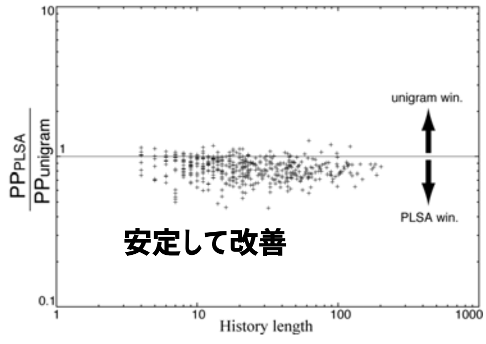
PLSIのPPとunigramモデルのPPの比 (小さいほどPLSIがよい)



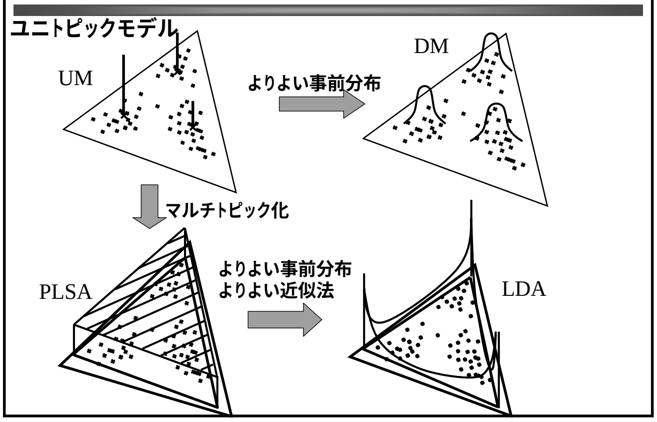
学習データ：毎日新聞1999年版, 20k語彙,100混合
テストデータ：毎日新聞1998年版, 495記事

(擬似) LDAによる言語モデル

[三品&山本2004]



トピックモデルのまとめ



性能比較と応用

- パープレキシティによる比較
- マルチトピック文書
- キャッシュモデル
- 音声認識
- スペルチェッカ

パープレキシティ

- 統計的言語モデルの代表的評価指標

$$PP = P(\boldsymbol{D})^{-\frac{1}{N}}$$

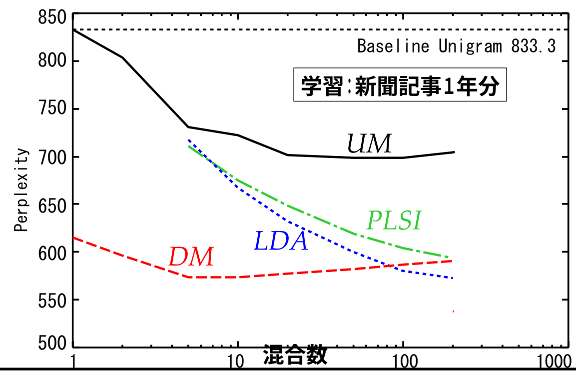
D: テストセット
N: テストセット中の全単語数

- 考え方
  - よい統計的言語モデルは正しい文に高い確率
  - ある正しい文からなる文集合（テストセット）の確率で比較
  - 1単語当りの平均確率の逆数（小さいほど高性能）
- 直感的な意味
  - 次単語候補を絞り込む力
    - モデルがない場合： PP=語彙サイズ
    - モデルが強力だと小さくなっていく

性能比較: パープレキシティ 1

[貞光 2006]

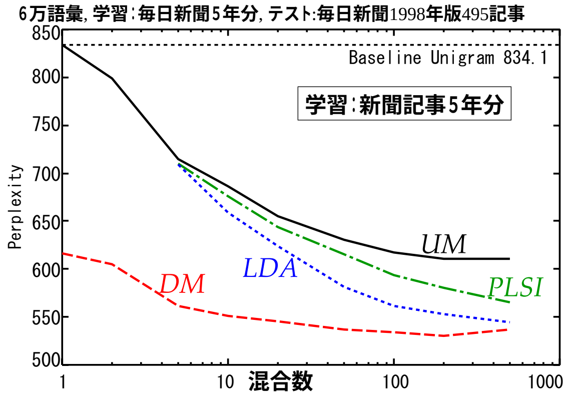
6万語彙, 学習：毎日新聞1年分, テスト:毎日新聞1998年版495記事



性能比較:パープレキシティ 2

[貞光 2006]

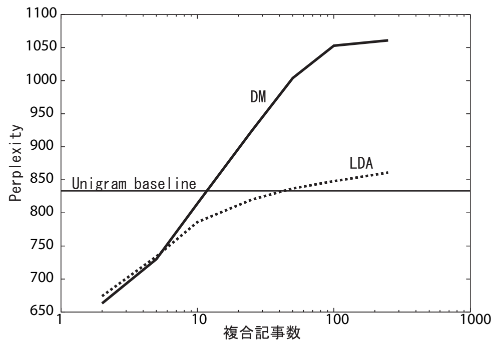
6万語彙, 学習：毎日新聞5年分, テスト:毎日新聞1998年版495記事



マルチトピック文書

[貞光 2006]

- いくつかの記事をまとめて一つの文書とした場合のPP

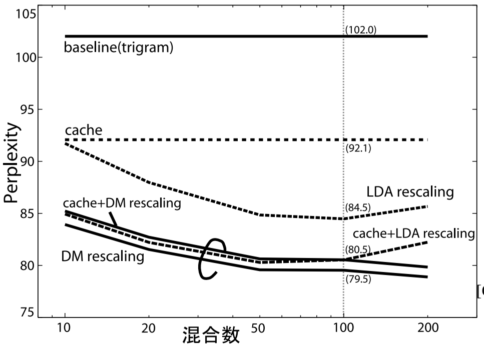


100混合, 6万語彙, 学習：毎日新聞1999年版, テスト:毎日新聞1998年版の記事より

DM+cache　DM　LDA+cache

[中里他 2005]

基本モデル：trigramにunigram-rescaling法でDMまたはLDAを統合



Cacheモデル:
履歴hで$n$gram
モデルを作成
[Kuhn&de Mori 1990]

その他のモデルと
線形補完

rescalingモデル
[Gildea&Hofmann1999]

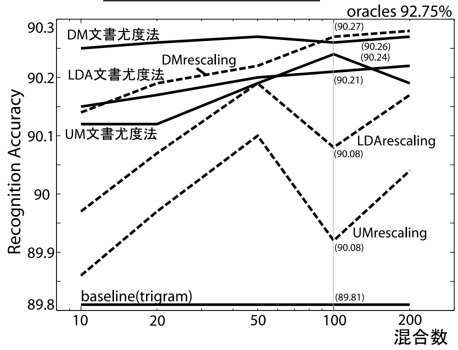
Bigram-cache, 60k語彙, 学習データ：新聞記事5年分, openテストデータ：15記事

記事読み上げ音声認識

記事全体で尤度を最大化する

[中里他 2005]

DMが安定して高性能 ⇨ キャッシュ機能が有効



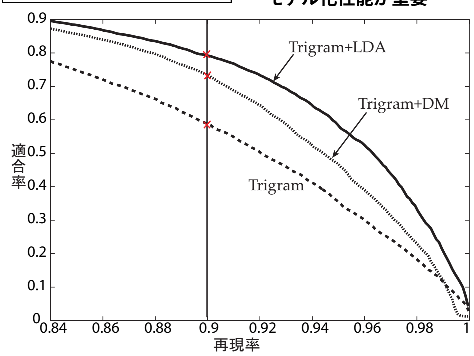
語彙サイズ6万
学習データ
　毎日新聞5年分

テストデータ
　JNAS記事読上音声
　(15記事, 男女各1名)

同音異義語スペルチェッカ

Trigram+LDAが高性能 ⇨ 異なる単語の共起モデル化性能が重要

〜を移す vs. 〜を写す



再現率：90%

適合率：

Trigram 60%

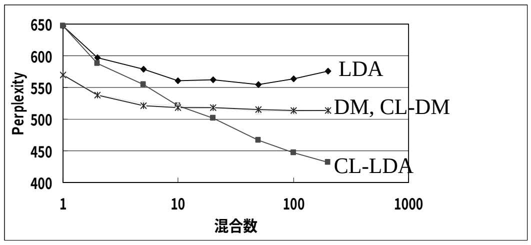
Trigram + DM　75%

Trigram + LDA　80%

言語横断モデル

LDA, DM → 日本語記事に対するPerplexity

CL-DM, CL-LDA → 英語記事で適応後 (事後分布→事前分布) に、日本語記事に対するPerplexityを計算



学習データ： 日英翻訳記事7万記事 (各言語60万文)
モデル： LDAとDMを語彙日英各3万単語 (計6万単語) として作成
テストデータ： 250記事 (オープン)

Dirichlet Process Mixtures: DPM
Hierarchical Dirichlet Process: HDP
—— Nonparametric Bayes ——

[Jordan 2005]
[Teh et al. 2004]

ノンパラメトリック・ベイズ

- トピック数T の決定 (モデル選択の問題)
  - Development test-set, または交差検定
  - AIC, BIC, MDL,...
  - アンサンブルモデル
  - $P(m{=}T \mid D)$を評価

  別の方法
- ノンパラメトリック・ベイズ
  - 確率分布Gの事前分布
    - すなわち、確率分布Gは確率変数！
  - Gの値としての分布のパラメータ数すら決まっていない

Gが離散分布の場合の代表的分布: DP (Dirichlet Process)

DP: Dirichlet Process　1/2

[Ferguson 1973][Teh 2004] [古澄 2005]

定義： $G \sim DP(\alpha, G_0)$

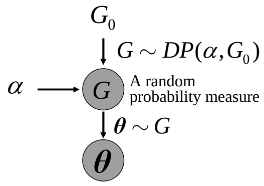
全事象, ボレル集合族

確率分布G_0が定義された可測空間(Ω, B)を考える。αは正の実数とする。このときランダムな確率分布Gに対して，Ωの任意の分割$A_1,\ldots,A_m$（つまり$A_i \in B, A_i \cap A_j = \phi, \bigcup_{i=1}^{m} A_i = \Omega$）を考えたとき，$(G(A_1),\ldots,G(A_m))$がパラメータ$(\alpha G_0(A_1), .., \alpha G_0(A_m))$を持つディリクレ分布にしたがうとき，$G$をディリクレ過程であるという。

可測空間？？
→ ルベーグ積分（ちょっと難しい）

もっと分かりやすい言い換えが必要！

DP: Dirichlet Process　2/2



$E[G(\theta)] = G_0(\theta)$

G_0はGの平均を表す

$$Var[G(\theta)] = \frac{(1-G_0(\theta))G_0(\theta)}{\alpha+1}$$

αはGの分散を決める

[古澄 2005]

もっと分かりやすい言い換えが必要！ [Teh 2004]

（Q1）Gはどのような分布？
（Q2）Gからのサンプルは？
（Q3）Gを使ってトピックモデルは作れるか？

Q1：Gはどのような分布？

[Sethuraman 1994]

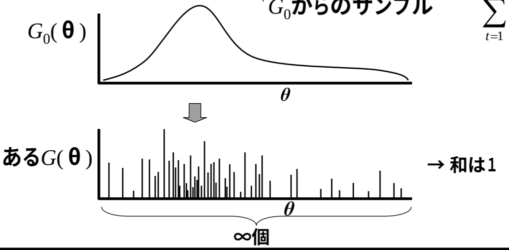
• Gは以下の**離散分布**になる　θが確率変数

$$G = \sum_{t=1}^{\infty} \lambda_t \delta(\theta, \varphi_t)$$

φ_tだけに集中した分布（φ_t以外では確率0）

G_0からのサンプル

$$\sum_{t=1}^{\infty} \lambda_t = 1$$



→ 和は1

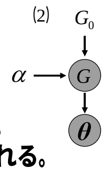
Q2：Gからのサンプルは？ 1/2

[Blackwell&MacQueen 1973]
[古澄 2005][Escobar 1994]

[Gからのサンプル（事前分布）]

$\theta_1,\ldots,\theta_{i-1}$の中で$\varphi_j$と同じ$\theta_k$が現れた回数

$$P(\theta_i = \varphi_j \mid \theta_1, \theta_2, \ldots, \theta_{i-1}) \propto \begin{cases} n_j, & (\varphi_j \text{がすでに出現}) \quad (1) \\ \alpha. & (\text{新しい}\ \varphi_j) \quad (2) \end{cases}$$



(1)の場合、以前出現したφ_jが再び選ばれる。
(2)の場合、新しいφ_jがG_0からサンプリングされる。

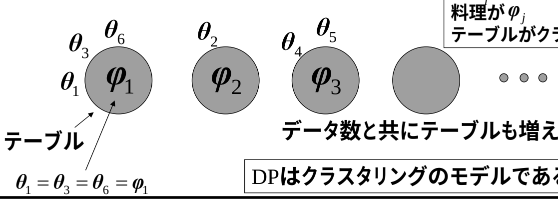
Q2：Gからのサンプルは？ 2/2

CRP: Chinese Restaurant Process [Aldous 1985]

$$P(\theta_i = \varphi_j \mid \theta_1, \theta_2, \ldots, \theta_{i-1}) \propto \begin{cases} n_j, (1) \\ \alpha. \ (2) \end{cases}$$

客θ_iが中華料理屋に来たとき、
(1) すでに客がいるテーブルに座って、同じ料理を食べる、
(2) 新しいテーブルに座って、新しい料理を注文。

客θ_iが個々のデータ
料理がφ_j
テーブルがクラスター

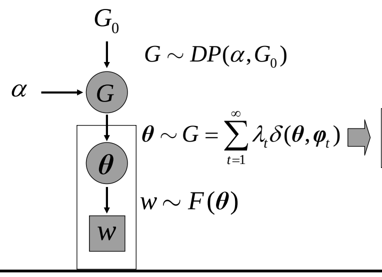


データ数と共にテーブルも増えてくる！

DPはクラスタリングのモデルである

Q3：Gを使ってトピックモデルは作れるか？ 1/3

• DPM: Dirichlet Process Mixtures
– Gからサンプルされたパラメータθで決まる分布F(θ)に、実際のデータwが従うモデル



可算無限の混合モデルと解釈できる

有限混合モデルを拡張できる？

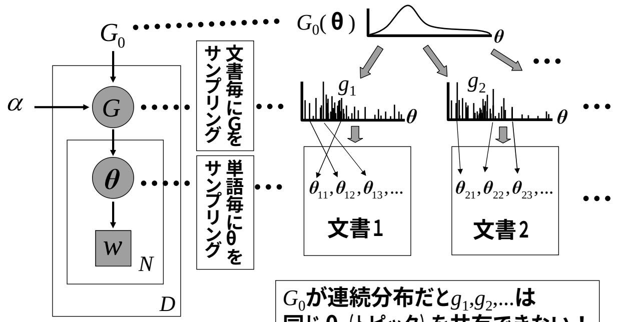
Q3：Gを使ってトピックモデルは作れるか？ 2/3

- DPM: Dirichlet Process Mixtures
  - 無限混合モデルへの拡張
    - 指数分布族分布の混合モデル
      - Infinite Gaussian Mixture Model [Rasmussen 2000]
    - Unigram Mixtures
      - 多項分布の無限混合はうまくいかないらしい？
    - Dirichlet Mixtures
      - Polya分布の無限混合モデル [持橋&菊井 2006]
    - PLSI&LDA？

      次ページ

Q3：Gを使ってトピックモデルは作れるか？ 3/3

- DPMではLDAをモデル化できない

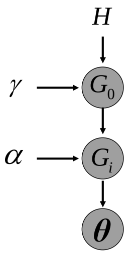


G_0が連続分布だと$g_1,g_2,\ldots$は同じθ（トピック）を共有できない！

Hierarchical DP

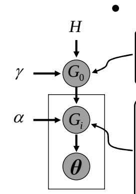
[Teh et al. 2004]

- 文書間の φ_t に関連がない → 関連を与えたい
  - G_0が連続分布だとGの値（特に φ_t）が同じになる可能性はゼロ
  - G_0が離散分布だとよい
    - DPは離散分布の事前分布！
  - $G_0 \sim DP(\gamma, H)$とすればよい
    - Hは連続分布でもよい



CRF: Chinese Restaurant Franchise

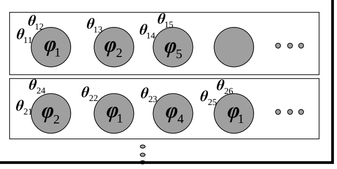
- 中華料理屋フランチャイズ（基本メニューは同じ）



- フランチャイズ料理開発部のおすすめ料理は全店での人気で決まるが、稀に新しい料理を開発する。
- 客が既にいるテーブルに座って同じ料理を食べる。
  - たくさん客がいるテーブルの料理は人気がある
- 新しいテーブルで、店のおすすめ料理を食べる。
  - 店のおすすめ料理は他のテーブルの人気で決まるが、稀に新しいおすすめ料理をフランチャイズ開発部に依頼

料理→トピック
レストラン→文書
客→単語（出現確率）

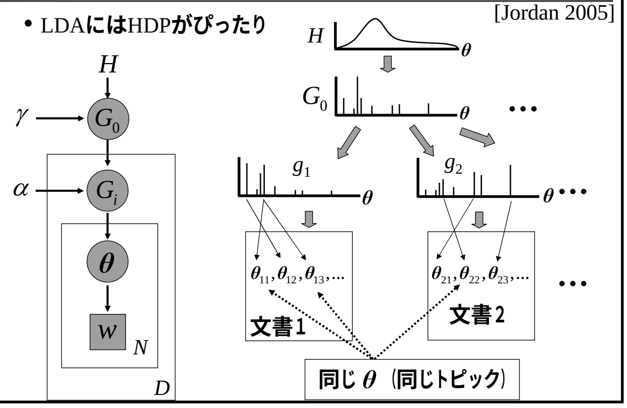
HDPによるLDA



HDPの適用例：LDA

[Jordan 2005]

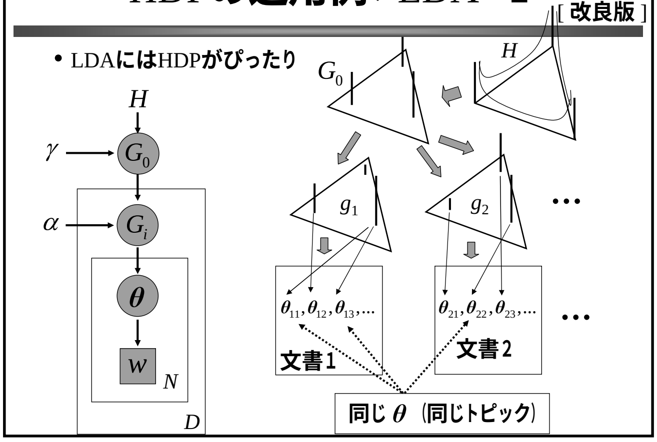
- LDAにはHDPがぴったり



HDPの適用例：LDA　2

[改良版]

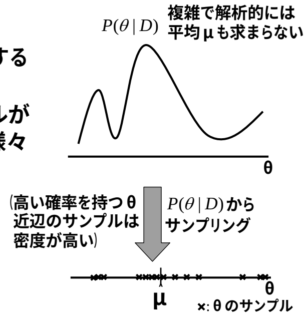
- LDAにはHDPがぴったり



推論：HDP ⇒ MCMC

- Monte Carlo法
  - 分布Fの性質を調べたいとする
    - 例えば、平均
  - 分布Fに従う大量のサンプルがあれば、サンプルを使って様々な性質が調べられる。
    - 例えば、サンプルの平均
- MCMC
  (Markov Chain Monte Carlo)
  - 事後分布のサンプルを生成する一般的な枠組み

[Garmerman 1997] [伊庭 2005]



サンプルをたくさん得ることができればサンプル平均は事後分布の平均！

推論：HDP-LDA

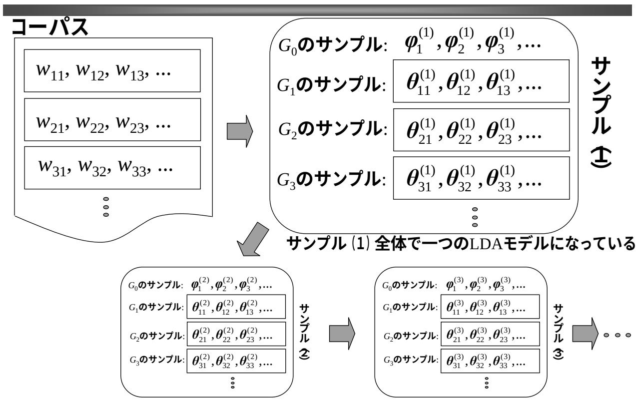
- DPMの事後分布サンプリング（＝CRP）

$$P(\theta_i \mid \theta_1,...,\theta_{i-1}, \boldsymbol{d}) \propto P(\theta_i, \boldsymbol{d} \mid \theta_1,...,\theta_{i-1})$$
$$\propto P(\boldsymbol{d} \mid \theta_i) \mathrm{P}(\theta_i \mid \theta_1,...,\theta_{i-1})$$

新しい料理：G_0と尤度（$P(\boldsymbol{d} \mid \theta_i)$）が考慮される

- HDPの事後分布サンプリング（＝CRF）
  - DPMの場合とほぼ同じ。
  - ただし、おまかせ料理（新しいトピックの導入）の場合には、親プロセスDP(γ ,H)から料理（トピック）を導入する。

HDP-LDAのサンプリング



実験

- [Teh et al. 2004]より
  - 学習・テストデータ（10-fold交差検定）
    - A corpus of nematode biology abstracts
      - 5,838 abstracts, 476,441単語（語彙サイズ=5,699）
  - テストセット・パープレキシティ
    - LDAの混合数を変化させて実験
    - HDPはMCMC

    } LDAの最高性能とHDPは同じ
  - HDPのT に関する事後分布
    - T =60〜70が高い確率
      （LDAは50〜80の混合数で最高性能）

    データの大きさ
    データの複雑さ } に応じて自動的に決まる

HDPはLDAの最適な混合数を正しく見積もっている！

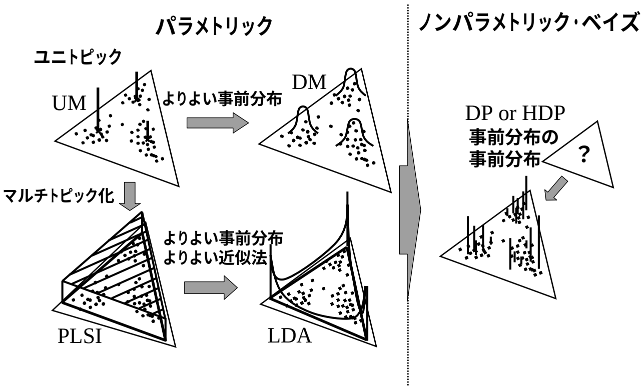
LDAのベストの混合数をうまく推定しているのは何故？

- 絶対値はたまたま？
  - 事前分布のハイパーパラメータにもちろん依存する
- しかし、
  - 十分に許容範囲の広い事前分布
    - データが自ら語る
  - DPのよい性質
    - データが多くなると新たなテーブル・料理が用意される確率が低くなる（パラメータはlog Nのオーダー）
    - 新しいデータが複雑だと（新しいテーブルを使った方が尤度が高くなる）、テーブルはどんどん増えていく

まとめ 1/2

- 従来モデル：真の出現確率をただ一つ推定する
- トピックモデル：出現確率の変動を捕らえる
  - ユニトピックモデル
    - Unigram Mixtures: 基本的なトピックモデル
    - Dirichlet Mixtures: キャッシュモデル
  - マルチトピックモデル
    - Probabilistic LSI: 非生成モデル, 過適応
    - Latent Dirichlet Allocation: 生成モデル, ロバスト
  - ノンパラメトリックベイズ
    - Dirichlet Process Mixtures
    - Hierarchical Dirichlet Process（HDP-LDA）

まとめ 2/2



応用

- 統計的言語モデル
  - 「文」を出力する応用一般に使える可能性
- トピックに基づく統計的言語モデル
  - 「文書」を出力する応用一般に使える可能性
  - これまで文単位で処理したしていたシステムの拡張
    - 音声認識、統計的機械翻訳、形態素解析
  - もともと文書で行っていたシステムへの応用
    - スペルチェッカ、情報検索、WSD、文書クラスタリング

参考文献

[Bahl et al. 1983] L.R.Bahl, F.Jelinek and R.L.Mercer. 1983. A maximum likelihood approach to continuous speech recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 5(2):179-190.

[Bellegarda 1998] J.R.Bellegarda. 1998. A multispan language modeling framework for large vocabulary speech recognition. IEEE Trans. Speech Audio Process, 6(5):456-467.

[Blackwell&MacQueen 1973] D.Blackwell and J.MacQueen. 1973. Ferguson distributions via Polya urn schemes. Annals of Statistics, 1:353-355.

[Blei et al. 2003] D.M.Blei, A.Y.Ng and M.I.Jordan. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research. 3:993-1022.

[Carlin&Louis 2000] B.P.Carlin and T.A.Louis. 2000. Bayes and Empirical Bayes Methods for Data Analysis (2nd ed.). Chapman & Hall/CRC.

[Brown et al. 1993] P.Brown, J.Cock, S.Della Pietra, V.Della Peitra, F.Jelinek, J.Lafferty, R.Mercer and P.Roossin. 1990. A statistical approach to machine translation. Computational Linguistics, 16(2):79-85.

[Church&Gale 1991] K.W.Church and W.A.Gale. 1991. Probability scoring for spelling correction. Statistics and Computing, 1:93-103.

[Church&Gale 1995] K.W.Church and W.A.Gale. 1995. Poisson Mixtures. Natural Language Engineering, 1(2):163-190.

[DLR 1977] A.P.Dempster, N.M.Laird and D.B.Rubin. Maximum-likelihood from incomplete data via the EM algorithm. In Journal of the Royal Statistics Society B (methodological), 39:1-38.

[Efron&Morris 1975] B.Efron and C.Morris. 1975. Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association, 70:311-319.

[Escobar 1994] M.D.Escobar. 1994. Estimating normal means with a Dirichlet process prior. J. of the American Statistical Association, 89:268-277.

[Garmerman 1997] D.Gamerman. 1997. Markov Chain Monte Carlo. Chapman & Hall/CRC.

[Gelman et al. 2004] A.Gelman et al. 2004. Bayesian Data Analysis. 2nd ed. Pages 10-11. Chapman & Hall/CRC.

[Gildea&Hofmann 1999] Daniel Gildea and Thomas Hofmann. 1999. Topic-based language models using EM. EuroSpeech'99, pages 2167-2170.

[Girolami&Kaban 2003] M.Girolami and A.Kaban. 2003. On an equivalence between PLSI and LDA. In SIGIR'03, pages 433-434.

[Hofmann 1999] Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. SIGIR-99, pages 50-57.

[Iyer&Ostendorf 1999] R.M.Iyer and M.Ostendorf. 1999. Modeling long distance dependence in language: topic mixtures versus dynamic cache models. IEEE Tran. on Speech and Audio Processing, 7(1):30-39.

[Jaakkola 2000] T.S.Jaakkola. Tutorial on variational approximation methods. In Advanced Mean Field Methods: Theory and Practice. MIT Press.

[Jordan 2005] M.I.Jordan. 2005. Dirichlet processes, Chinese restaurant processes and all that. Tutorial presentation at the NIPS2005 Conference. http://www.cs.berkeley.edu/~jordan

[Katz 1987] S.M.Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recogniser. IEEE Trans. ASSP, 35(3):400-401.

[Kuhn&de Mori 1990] R.Kuhn and R. de Mori. 1990. A chche-based natural language model for speech recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 12(6):570-583.

[Minka 2001] Thomas P. Minka. 2001. Using lower bounds to approximate integrals. http://www.stat.cmu.edu/minka/papers/rem.html.

[Minka 2003] Thomas P. Minka. 2003. Estimating a Dirichlet distribution.

http://www.stat.cmu.edu/minka/papaers/dirichlet.html.
[Ney et al. 1997] H.Ney, S.Martin, and F.Wessel. 1997. Statistical language modeling using leaving-one-out. Chapter 6 in Corpus-Based Methods In Language and Speech Processing, edited by S.Young and G.Bloothooft, pages 174-207. Kluwer Academic.
[Nigam et al. 2000] K.Nigam, A.Mccallum, S.Thrun, and T.Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. Machine Learning, 39(2/3):103-134.
[Ponte&Croft 1998] J.M.Ponte and W.B.Croft. 1998. A language modeling approach to information retrieval. In SIGIR'98, pages 275-281.
[Rasmussen 2000] C.E.Rasmussen. 2000. The infinite Gaussian mixture model. in Advances in Neural Information Processing System 12, pages 554-560.
[Rosenfeld 1996] R. Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modeling. In Computer, Speech and Language, 10:187-228.
[Sethuraman 1994] J. Sethuraman. 1994. A constructive definition of Dirichlet priors. Statistica Sinica, 4:639-650.
[Sjolander et al. 1996] K Sjolander, K.Karplus, M.Brown, R.Hunghey, A.Krogh, I.Mian and D.Haussler. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. In Computer Applications in Biosciences, Vol.12:327-345.
[Teh et al. 2004] Y.W.Teh, M.I.Jordan, M.J.Beal and D.M.Blei. 2004. Hierarchical Dirichlet processes. Technical Report 653, Dept. of Statistics, UC at Berkeley. http://www.stat.berkeley.edu/~jordan/653.pdf
[伊庭 2005] 伊庭,種村,大森,和合,佐藤,高橋. 2005. 計算統計 II(統計科学のフロンティア 12). 岩波書店.
[上田 2002] 上田修功. 2002. ベイズ学習 I, II, III, IV. 電子情報通信学会誌, 85(4:265-271), 85(6:421-426), 85(7:504-509), 85(8:633-638).
[上田 2003] 上田修功. 2003. テキストモデリングの新展開. 言語処理学会第 9 回年次大会チュートリアル資料, pages 1-17.
[北 1999] 北研二. 1999. 確率的言語モデル. 東京大学出版会.
[久保川 2004] 久保川達也. 2004. スタインのパラドクスと縮小推定の世界. モデル選択(甘利他編,統計科学のフロンティア 3)の第３章. 岩波書店.
[貞光他 2005] 貞光九月, 三品拓也, 山本幹雄. 2005. 混合ディリクレ分布を用いたトピックに基づく言語モデル. 電子情報通信学会論文誌 D-II, J88(9):1771-1779.
[貞光 2006] 貞光九月. 2006. 階層ベイズモデルを用いた混合ディリクレモデルのスムージング法. 筑波大学システム情報工学研究科コンピュータサイエンス専攻修士論文.
[三品&山本 2004] 三品拓也, 山本幹雄. 2004. 確率的 LSA に基づく ngram モデルの変分ベイズ学習を利用した文脈適応化. 電子情報通信学会論文誌 D-II, J87(7):1409-1417.
[三品他 2004] 三品拓也, 貞光九月, 山本幹雄. 2004. 確率的 LSA を用いた日本語同音異義語誤りの検出・訂正. 情報処理学会論文誌, 45(9):2168-2176.
[中里他 2005] 中里, 貞光, 冨山, 山本, 板橋. 2005. 生成文書モデルを用いた文書読み上げ音声認識. 情報処理学会研究報告, 2005-SLP-57-4:19-24.
[毎日新聞社 1991-2004] 毎日新聞社. 1991-2004. CD-毎日新聞'91-'99, 2000, 2004 データ集. 日外アソシエーツ(株).
[古澄 2005] 古澄英男. 2005. ディリクレ過程事前分布を用いた構造変化のベイズ分析. ベイズ計量経済分析(和合肇編著,2005,東洋経済新報社)の第８章, pages 235-258.
[持橋&菊井 2006] 持橋大地, 菊井玄一郎. 無限混合ディリクレ文書モデル. 情報処理学会 自然言語処理研究会 NL-172, to appear.
[山本 1999] 山本幹雄. 1999. 統計的言語モデル-理論と実験-. 言語処理学会第 9 回年次大会チュートリアル資料.