第1章 統計的潜在意味解析とは

第1回「トピックモデルによる統計的潜在意味解析」読書会

@ksmzn

会場:株式会社 ALBERT 西新宿

June 4, 2015

自己紹介

Koshi @ksmzn

- 某大学 M2 → 社会人一年目
- リサンプリング法を研究してました
- SQL にまみれる日々

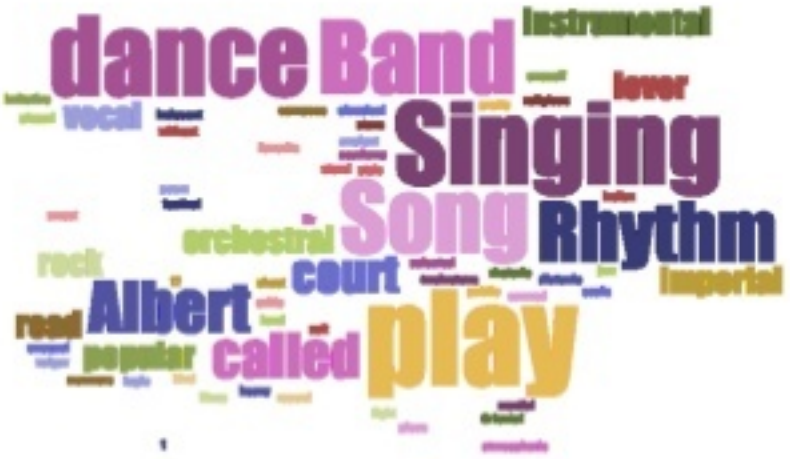
はじめに

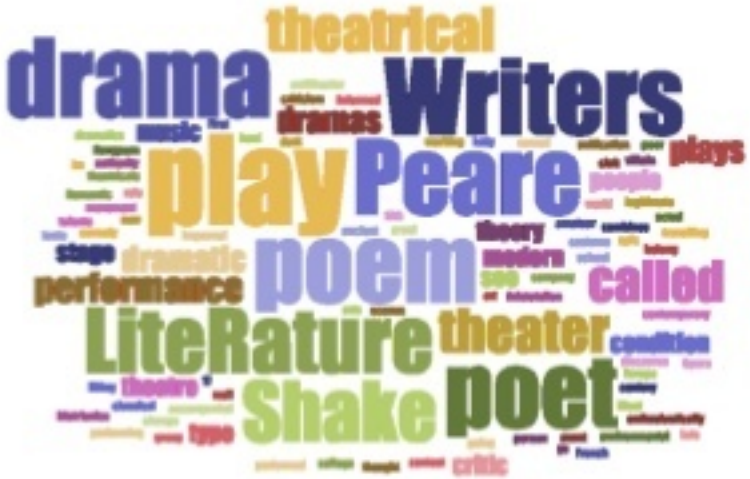
https://speakerdeck.com/yamano357/tokyowebmining46th

先日の TokyoWebmining での資料がとても参考になるので、見ましょう！！

目次

目次


dance
Band
Instrumental
lover
vocal
Singing
Song
Rhythm
orchestral
rock
court
imperial
Albert
play
read
popular
called


theatrical
drama
Writers
dramas
music
play
Peare
plays
people
theory
stage
dramatic
poem
called
performance
LiteRature
theater
poet
Shake
theatre
type
critic


field
throw
Play
equatic
SHOES
Baseball
Bat
Tennis
games
outdoor
balls
Catch
spring

潜在的意味・トピックと潜在的共起性

潜在的意味

- 「音楽」や「スポーツ」という単語が無かったとしても、単語群を見て想起できる
- 複数の単語の共起性によって創発される情報

トピック

- 潜在的意味のカテゴリをトピックと呼ぶ

→「単語の共起性をいかに数学的にモデル化するか？」

目次

潜在意味解析の歴史

- 行列分解 (1988)
  Latent Semantic Indexing/Analysis (LSI/LSA)
- 確率モデル (1998)
  Probabilistic LSI/LSA (PLSI/PLSA)
- 階層ベイズモデル (2003)
  Latent Dirichlet Allocation (LDA)
- 拡張モデル多数 (2004 ごろ)
- 大規模データのための高速化 (2007)

特異値分解

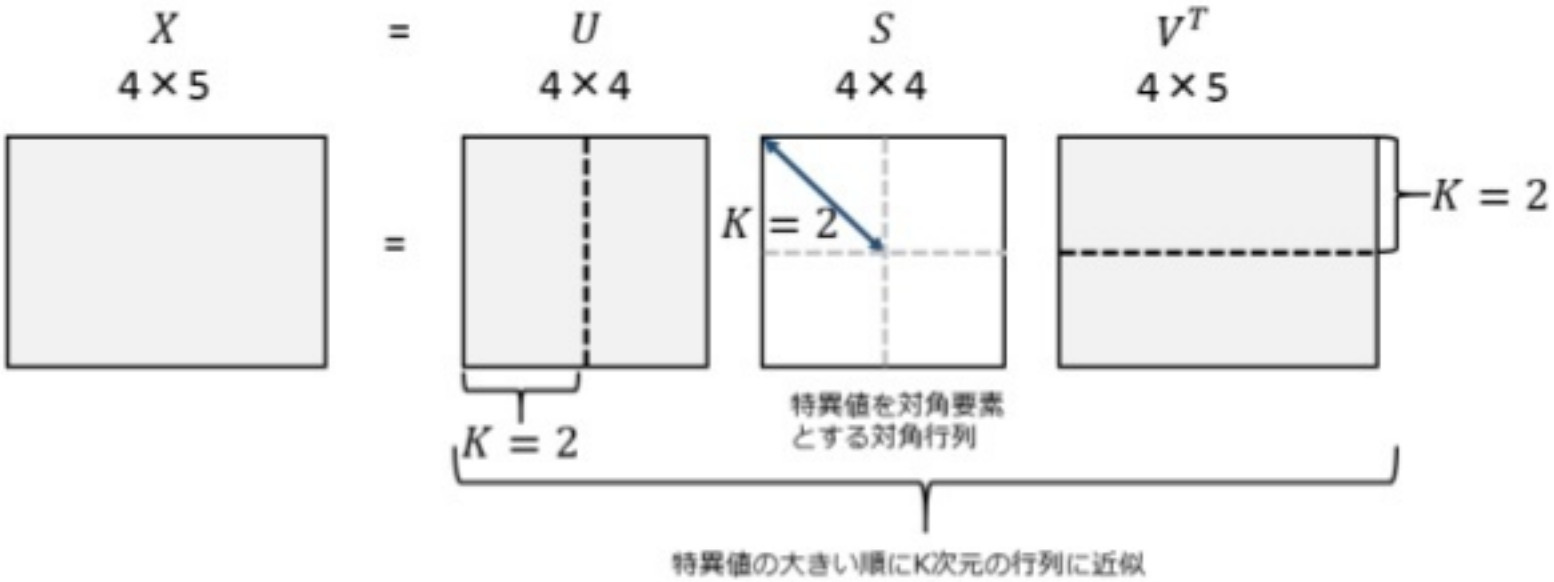
- 単語文書行列 X を 3 つの行列に分解

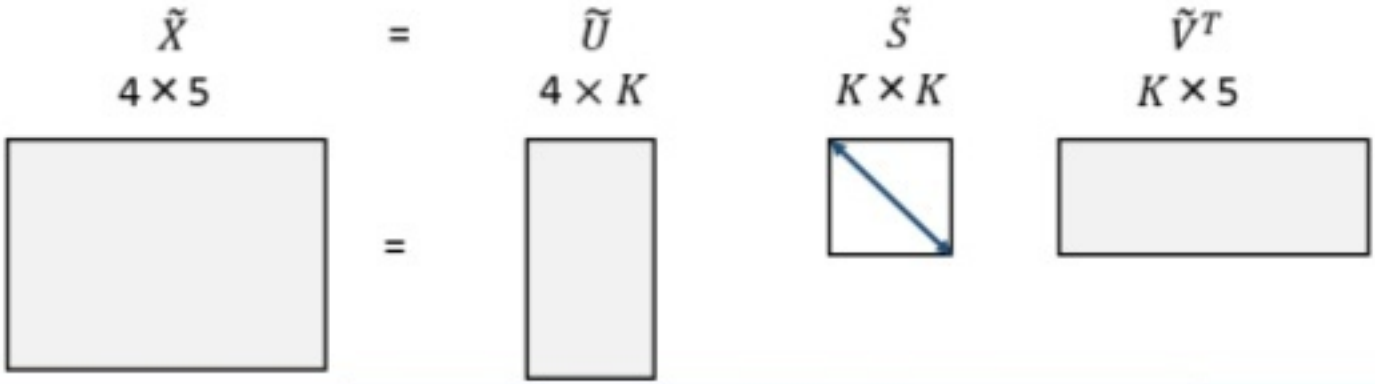
$$X = USV^{\mathrm{T}}$$

- U, S, V の各列ベクトルを特異値が大きい順に K 個用いて、$\tilde{U}, \tilde{S}, \tilde{V}$ を作り、ランク K の低ランク近似行列 $\tilde{X}$ を得る

$$\tilde{X} = \tilde{U}\tilde{S}\tilde{V}^{\mathrm{T}}$$

特異値分解


X
4×5
=
U
4×4
S
4×4
V^T
4×5
=
$K = 2$
$K = 2$
$K = 2$
特異値を対角要素
とする対角行列
特異値の大きい順にK次元の行列に近似


$\tilde{X}$
4×5
=
$\tilde{U}$
$4 \times K$
$\tilde{S}$
$K \times K$
$\tilde{V}^T$
$K \times 5$
=

特異値分解による潜在意味解析

文書に含まれている単語を抽出し、それらの頻度から単語文書行列 X を作成する

	drive	automobile	car	play	music
文書1	2	3	0	0	0
文書2	2	0	2	0	0
文書3	0	0	0	2	2
文書4	0	0	0	3	1

- 「car」で検索しても、文書 1 は発見できない
- 「automobile」でも、文書 2 は発見できない

→単語の持つ潜在的な意味を考える
→特異値分解

特異値分解の結果

	drive	automobile	car	play	music
文書1	2.38	2.29	0.85	0	0
文書2	1.32	1.27	0.47	0	0
文書3	0	0	0	2.36	1.37
文書4	0	0	0	2.67	1.55

文書 1・2 ともに、「car」「automobile」の頻度が 0 でない！
→「drive」との共起性から、潜在的な意味が抽出されている

$\tilde{V}$の情報

	drive	automobile	car	play	music
Topic 1	0	0	0	0.86	0.5
Topic 2	0.7	0.67	0.25	0	0

各列ベクトルは、複数の単語の共起性を表している。
→潜在トピック

$\tilde{U}$ の情報

	Topic1	Topic2
文書1	0	0.87
文書2	0	0.48
文書3	0.66	0
文書4	0.75	0

各列ベクトルは、文書とトピックの共起性を表している。
→間接的に、文書と単語の共起性を抽出できる

LSIの問題点

- $\tilde{U}, \tilde{S}$ の解釈が難しい
- 特異値分解の性質により、トピックの軸が互いに直交するため、トピックに対し非常に強い制約となる

→PLSI, 階層ベイズモデル, etc...

目次

確率的潜在変数モデル

確率的潜在変数モデル

1. 観測できない潜在変数を仮定する数理モデル
2. 潜在変数をデータから推定することで、データ間の類似性とその意味を解析する

例：データ間の類似性

- $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$：観測変数
- $\boldsymbol{z} = \{z_1, z_2, \ldots, z_n\}$：潜在変数
- $\boldsymbol{\phi} = \{\phi_1, \phi_2, \ldots, \phi_K\}$：
  どのように類似しているのかを表す確率変数

$$z_1 = z_2 = k \quad \Rightarrow$$

x_1 と x_2 は ϕ_k の意味で類似している

目次

確率的生成モデルとグラフィカルモデル

確率的生成モデル

- データの生成過程を確率モデルで表現した数理モデル

グラフィカルモデル

- 確率的生成モデルを視覚的に表現するもの

生成過程

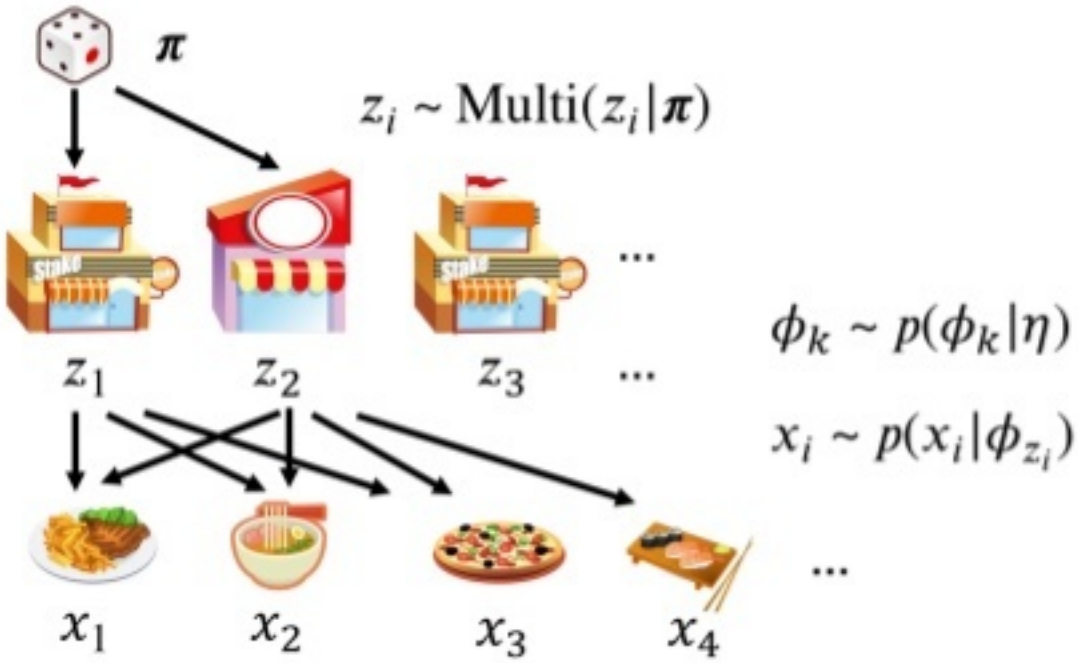
ある確率変数 $x_i (x = 1, \ldots, n)$ が確率分布 $p(x_i|\phi)$ に従うとき,

$$x_i \sim p(x_i|\phi) \qquad (i = 1, \ldots, n)$$

と記述する。
↓
確率変数 x_i の値が, 確率分布 $p(x_i|\phi)$ から生成されたことを示す。

サイコロで考える

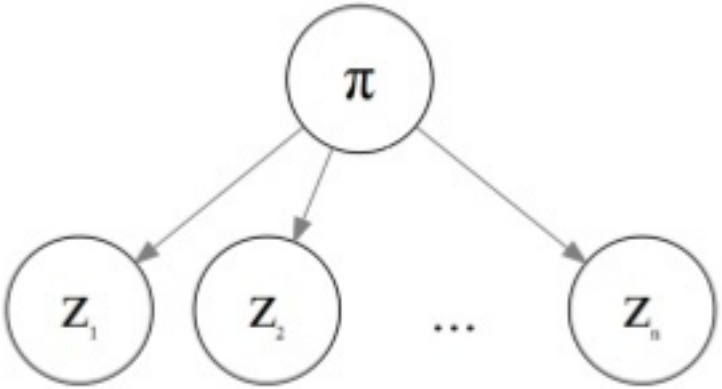
K 個の目が出るサイコロを n 回振ったときに出る目を生成モデルとして考える

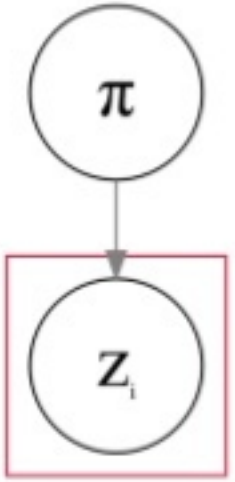


グラフィカルモデル

グラフィカルモデル

- 確率変数間の条件付き依存構造のグラフ表現
- サイコロ生成モデルのπとz_iの関係をグラフィカルモデルで以下のように表す





ベイズの定理と条件付き独立性

グラフィカルモデルは、ベイズの定理や条件付き独立性によって同時確率を展開するのに役立つ。

ベイズの定理

$$p(x|y) = \frac{p(y|x)\,p(x)}{p(y)}$$

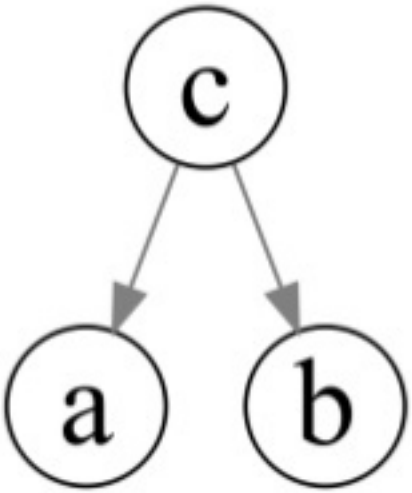
条件付き独立性

z が与えられた下での x と y の条件付き確率分布を $p(x|z)$, $p(y|z)$ とし, (x, y) の条件付き同時分布を $p(x, y|z)$ とする。

このとき、すべての x, y に対し $p(x, y|z) = p(x|z)p(y|z)$ が成り立つとき、「z が与えられた下で x と y は条件付き独立である」といい、$x \perp\!\!\!\perp y|z$ と表す

tail-to-tail 型

tail-to-tail 型



条件付き独立性:$a \perp\!\!\!\perp b|c$

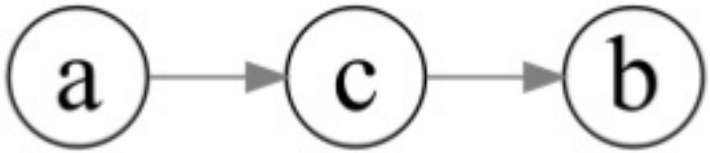
$$\Rightarrow p(a,b|c) = p(a|c)p(b|c)$$

グラフに対応する同時分布

$$p(a,b,c) = p(a|c)p(b|c)p(c)$$

head-to-tail 型

head-to-tail 型



条件付き独立性:$a \perp\!\!\!\perp b|c$

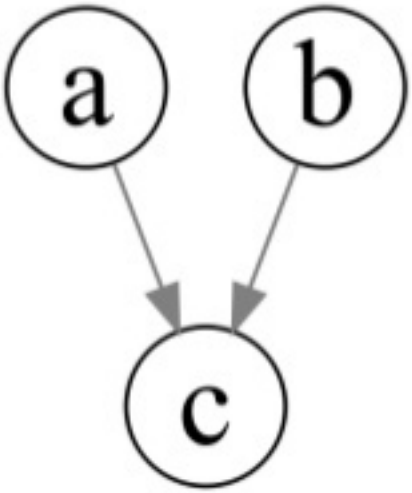
$$\Rightarrow p(a,b|c) = p(a|c)p(b|c)$$

グラフに対応する同時分布

$$p(a,b,c) = p(b|c)p(c|a)p(a)$$

head-to-head型

head-to-head型



条件付き独立性:$a \not\perp b|c$

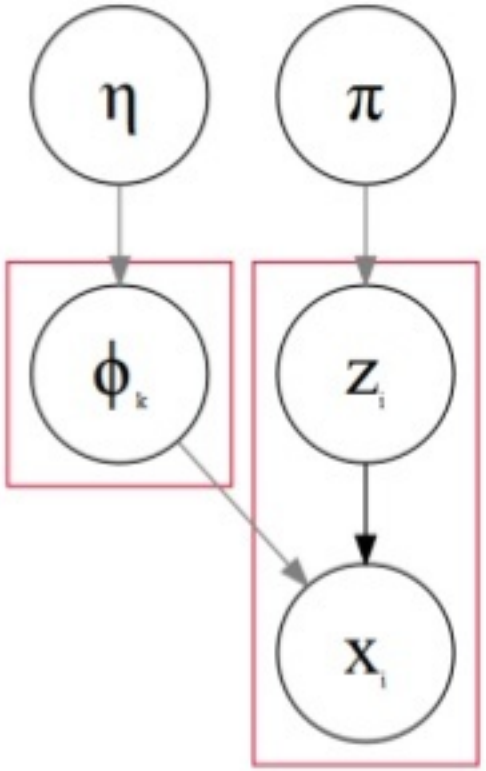
$$\Rightarrow p(a,b|c) \neq p(a|c)p(b|c)$$

グラフに対応する同時分布

$$p(a,b,c) = p(c|a,b)p(a)p(b)$$

サイコロ生成モデルの同時分布

同時分布 $p(\boldsymbol{x}, z, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$ を展開する



- $\boldsymbol{\pi}$ の生成確率は $p(\boldsymbol{\pi})$
- η の生成確率は $p(\eta)$
- $\boldsymbol{\pi}$ が与えられた下で z は tail-to-tail 型なので、

$$p(\boldsymbol{z}|\boldsymbol{\pi}) = \prod_{i=1}^{n} p(z_i|\boldsymbol{\pi})$$

- η が与えられた下で $\boldsymbol{\phi}$ は tail-to-tail 型なので、

$$p(\boldsymbol{\phi}|\eta) = \prod_{k=1}^{K} p(\phi_k|\eta)$$

サイコロ生成モデルの同時分布

- z と $\boldsymbol{\phi}$ が与えられた下で $\boldsymbol{x}$ は tail-to-tail 型なので、

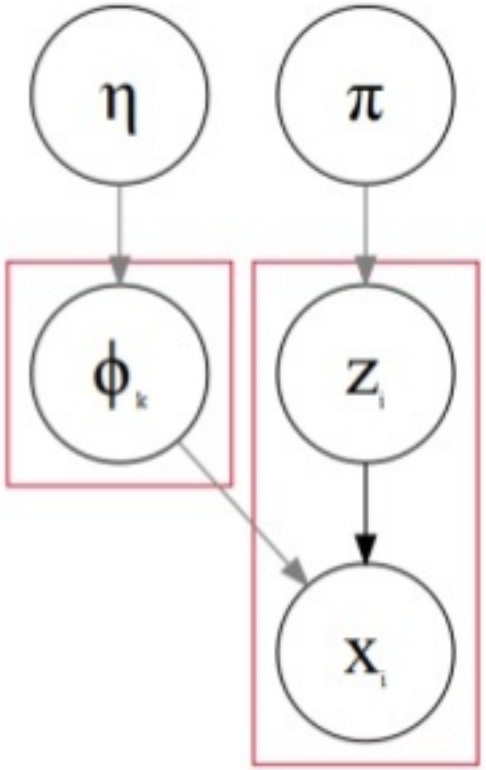
$$p(\boldsymbol{x}|z, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = p(\boldsymbol{x}|z, \boldsymbol{\phi}) = \prod_{i=1}^{n} p(x_i|z_i, \boldsymbol{\phi})$$

- 同時分布は以下のように展開できる

$$\begin{aligned} p(\boldsymbol{x}, z, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) &= p(\boldsymbol{x}|z, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) p(z, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) \\ &= p(\boldsymbol{x}|z, \boldsymbol{\phi}) p(z|\boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\phi}|\eta) p(\eta) \\ &= \prod_{i=1}^{n} p(x_i|z_i, \boldsymbol{\phi}) \prod_{i=1}^{n} p(z_i|\boldsymbol{\pi}) p(\boldsymbol{\pi}) \prod_{k=1}^{K} p(\phi_k|\eta) p(\eta) \end{aligned}$$

サイコロ生成モデルの条件付き分布1

条件付き分布 $p(z|\boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta)$ を計算する

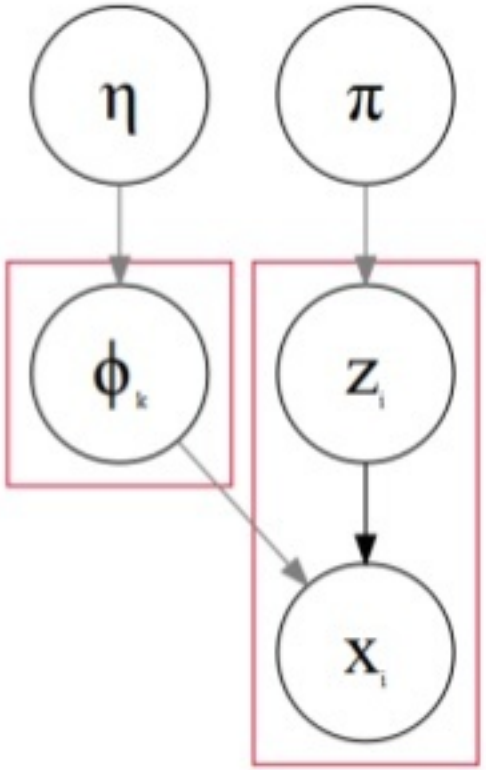


- $\boldsymbol{\pi}$ および $\boldsymbol{x}$ は z と繋がっているので依存関係がある
- $\boldsymbol{x}$ が与えられているので、$\boldsymbol{\phi}$ は z に対し独立にならない (head-to-head 型)
- $\boldsymbol{\phi}$ が与えられているので、η と $\boldsymbol{x}$ は条件付き独立 (head-to-tail 型)。従って、η と z も条件付き独立
- よって、

$$p(z|\boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\phi}, \eta) = p(z|\boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\phi})$$

サイコロ生成モデルの条件付き分布2

条件付き分布 $p(\boldsymbol{\phi}|\boldsymbol{x}, \boldsymbol{\pi}, z, \eta)$ を計算する



- η および $\boldsymbol{x}$ は $\boldsymbol{\phi}$ と繋がっているので依存関係がある
- $\boldsymbol{x}$ が与えられているので、$\boldsymbol{\phi}$ は z に対し独立にならない (head-to-head 型)
- z が与えられているので、$\boldsymbol{\pi}$ と $\boldsymbol{x}$ は条件付き独立 (head-to-tail 型)。従って、$\boldsymbol{\pi}$ と $\boldsymbol{\phi}$ も条件付き独立
- よって、

$$p(\boldsymbol{\phi}|\boldsymbol{x}, z, \boldsymbol{\pi}, \eta) = p(\boldsymbol{\phi}|\boldsymbol{x}, z, \eta)$$

まとめ

1. 潜在的意味のカテゴリをトピックと呼ぶ
2. 特異値分解を行い、文書の潜在的な意味を解析した
3. グラフィカルモデルを書くことで、同時分布の展開が容易になった

ご清聴ありがとうございました.