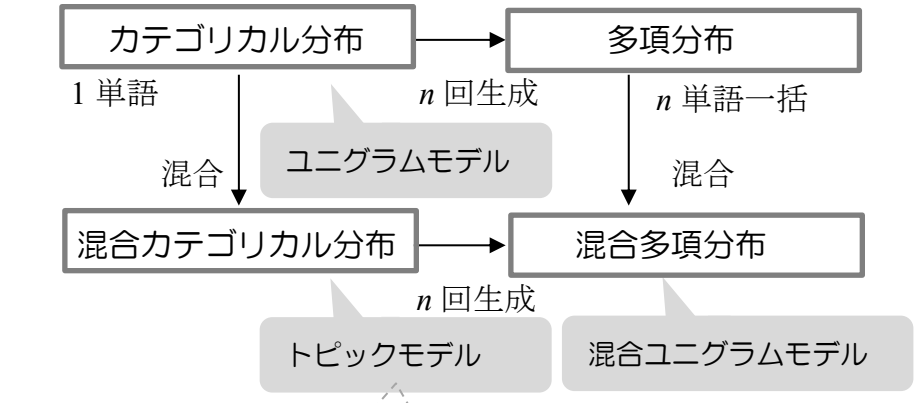
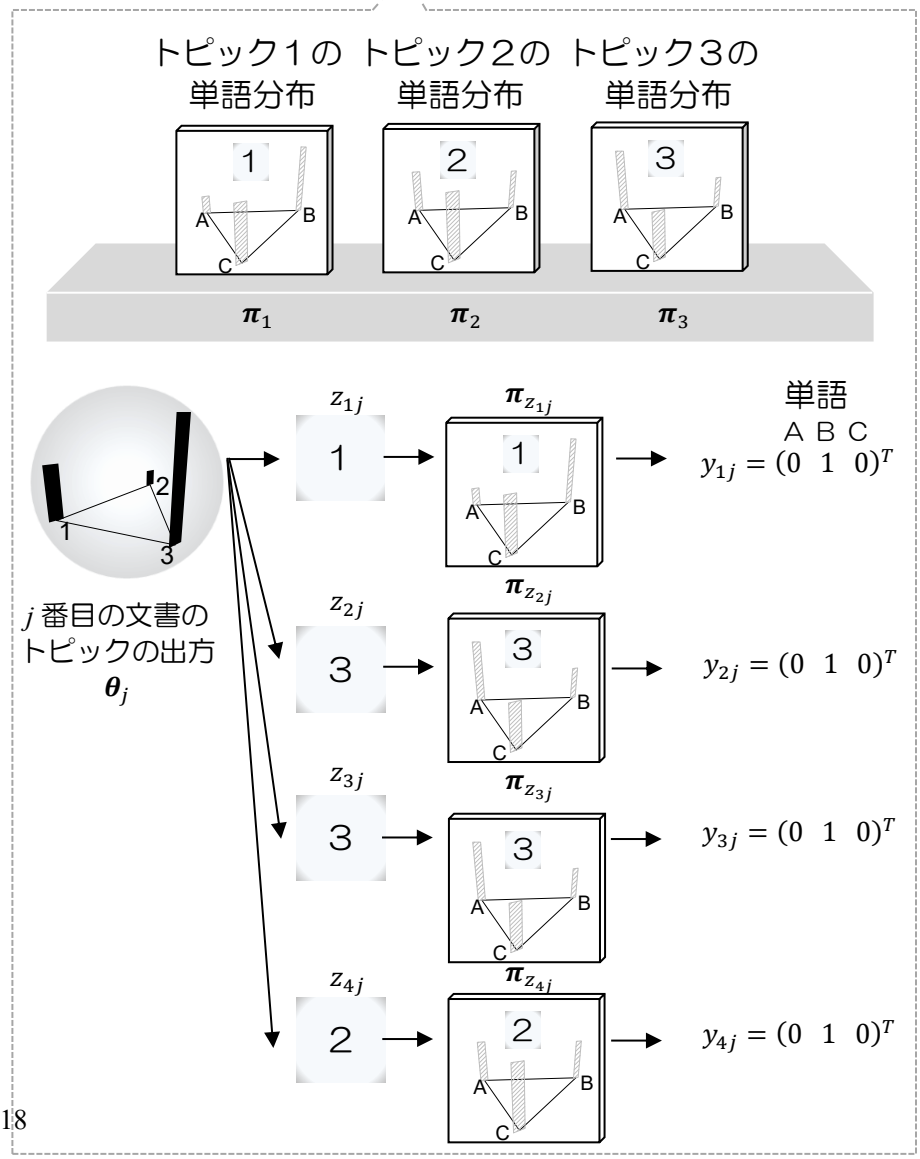
第4講　トピックモデル

学びのポイント

- 二つ以上の分布がある混合比に基づいて混ぜられた分布のことを**混合分布**と呼ぶことが分かり、基本的な分布について視覚的に理解をすることができる。
- 混合分布の密度式は、**「混ぜられる前の分布」×「その分布の混合比」**を各分布に対して計算し、それらを**足し合わせた**ものである、という点が理解できる。
- **混合正規分布、混合カテゴリカル分布、混合多項分布**において、その密度式を理解でき、混合比を変化させると分布の形状がどのように変化するのか、説明することができる。
- 自然言語処理をはじめとする、文書への統計的アプローチでは、文の中の階層構造（統語構造）を捨象し、Bag-of-Words と呼ばれる表現で議論が行われていることが分かる。
- 各**単語**を、ベクトルとして表記するやり方が分かる。
- 各**文書**を、単語の集合としてベクトル表記するやり方が分かる。
- **文書の集合**を、ベクトルとして表記するやり方が分かる。
- **ユニグラムモデル**が母集団に置いている仮定が理解でき、多項ロジスティック回帰において切片のみのモデルを仮定していることと同じであることを説明できる。
- **混合ユニグラムモデル**が母集団に置いている仮定が理解でき、各単語が従う単語分布が決定される際に、潜在的なトピックという概念を用いて分布を混ぜていることを説明できる。
- **トピックモデル**が母集団に置いている過程が理解でき、混合ユニグラムモデルとは異なり、各文書にトピックが複数混ざって存在できる仕組みになっていることを説明できる。
- トピックモデルの推定結果について、具体的な事例に即して推論を行うことができ、**潜在的なトピック（意味）の分析**の仕方を説明できる。

見取り図


カテゴリカル分布
多項分布
1 単語
n 回生成
n 単語一括
混合
ユニグラムモデル
混合
混合カテゴリカル分布
混合多項分布
n 回生成
トピックモデル
混合ユニグラムモデル


トピック1の 単語分布
トピック2の 単語分布
トピック3の 単語分布
1
2
3
A
B
C
$\boldsymbol{\pi}_1$
$\boldsymbol{\pi}_2$
$\boldsymbol{\pi}_3$
1
2
3
j 番目の文書の トピックの出方 $\boldsymbol{\theta}_j$
z_{1j}
z_{2j}
z_{3j}
z_{4j}
1
3
3
2
$\boldsymbol{\pi}_{z_{1j}}$
$\boldsymbol{\pi}_{z_{2j}}$
$\boldsymbol{\pi}_{z_{3j}}$
$\boldsymbol{\pi}_{z_{4j}}$
単語
A B C
$y_{1j} = (0 \ 1 \ 0)^T$
$y_{2j} = (0 \ 1 \ 0)^T$
$y_{3j} = (0 \ 1 \ 0)^T$
$y_{4j} = (0 \ 1 \ 0)^T$

■ 目標

ここでは、「言語統計学 B」で登場したカテゴリカル分布／多項分布を発展させたユニグラムモデル、混合ユニグラムモデル、トピックモデルという統計モデルを扱います。

これらのモデルは、文書（あるいは文書における単語）の生成過程に関するモデルです。「生成過程」というのは、単語がどのような確率で選ばれて登場するのかというプロセスのことです。単純に考えれば、単語というものは、「犬」「猫」「科学」「恋人」…と、数こそ多いものの離散的なカテゴリーを形成していますから単語の出やすさを表すカテゴリカル分布を想定し、そこからランダムに選ばれているという想定を置きたくなります。

このシンプルな仮定のもと単語の生成過程をモデル化したものが**ユニグラムモデル**です。これは、多項ロジスティック回帰の切片モデルであり、その構造の単純さからこの講義でもまずこのモデルに注目をしながら話を展開していきたいと考えています。

しかし、実際の研究の場面では、ユニグラムモデルよりもう少し複雑な仮定を考えた方が切れの良い分析ができる場合もあります。例えば、単語の出やすさは、その文章のトピックに応じて変わるという仮定を置くこともできます。つまり、文学系の文書では「愛」とか「夫婦」とか「友情」といった単語が生成されやすいことでしょう。しかし、一方で生物の教科書では、「求愛」とか「つがい」とか「群れ」といった言葉が出る確率が高まっていると考えられます。このようにその文書がどういうトピックに属するかで、カテゴリカル分布が変わるという想定を置いたものを**混合ユニグラムモデル**と呼びます。

しかし、この混合ユニグラムモデルよりも近年盛んに用いられるモデルとして**トピックモデル**と呼ばれるモデルがあります。これは、各文書にトピックが一つと考える混合ユニグラムモデルとは対照的に、一つの文書の中に複数のトピックが混在していると考えるモデルです。この講義の最終目的地は、みなさんにこのトピックモデルの考え方の基礎を、その応用事例とともに学んでもらうことにあります。

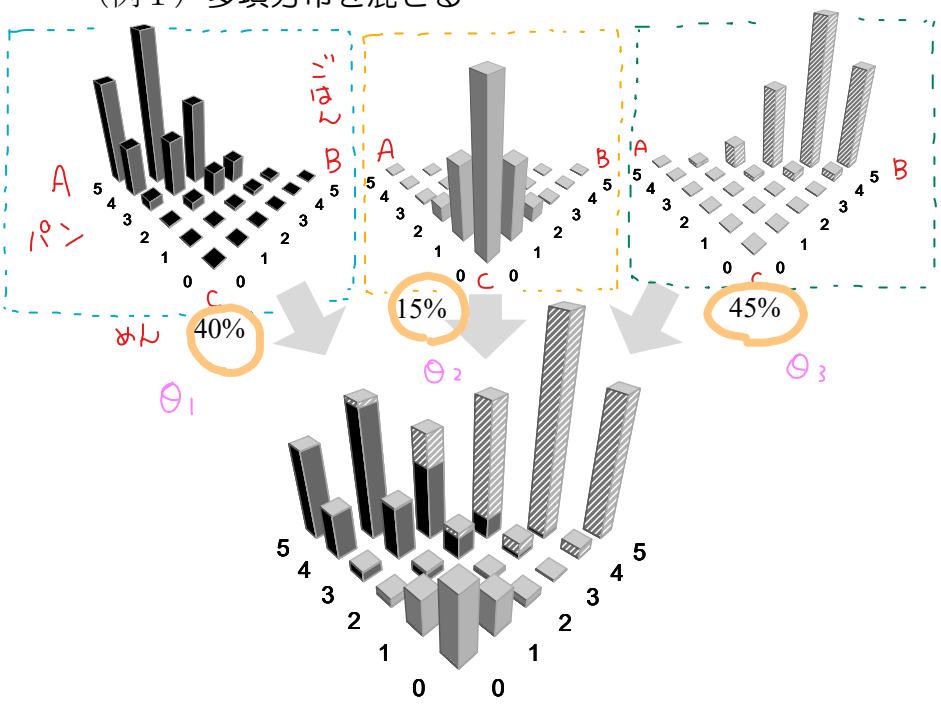
また、この混合ユニグラムモデルとトピックモデルのベースには二つ以上の分布を混ぜるという考え方が存在します。このような二つ以上の分布が混ざったものを**混合分布**と言い、話の前座として、この講義の前半では、混合分布と呼ばれるものがどのような性質を持つのか、視覚的な説明も交えながら、ゆっくりと解説をしていきます。

🕮 ノート０　あらすじ：トピックモデルのねらい

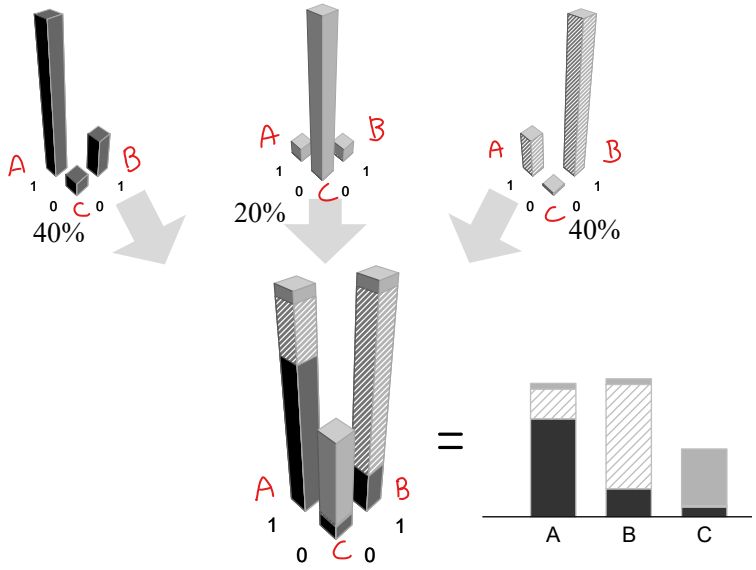
（１）モデル拡張の動機：分布の混合

複数の単純な事象が混ざったため複雑に見える事象がある。データから混ざり方と混ざる前の構造を推測したい。

（例１）多項分布を混ぜる



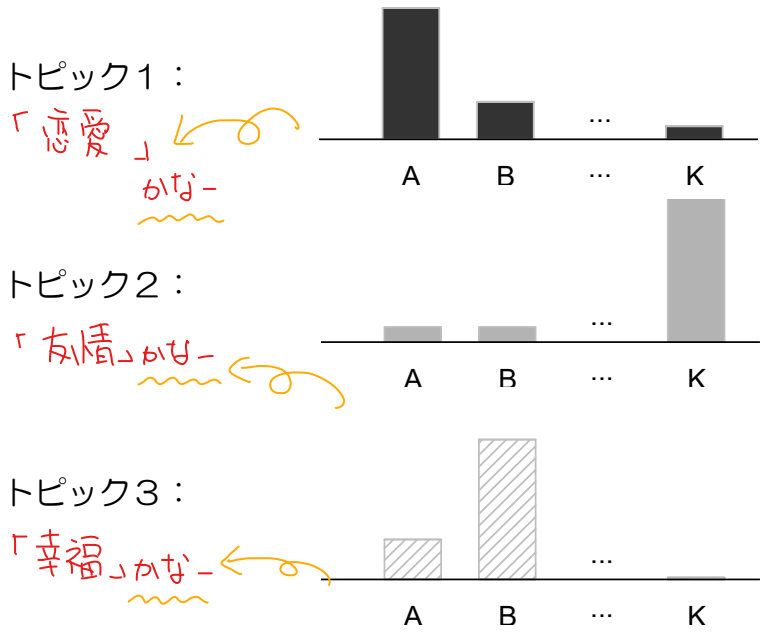
（例２）カテゴリカル分布を混ぜる



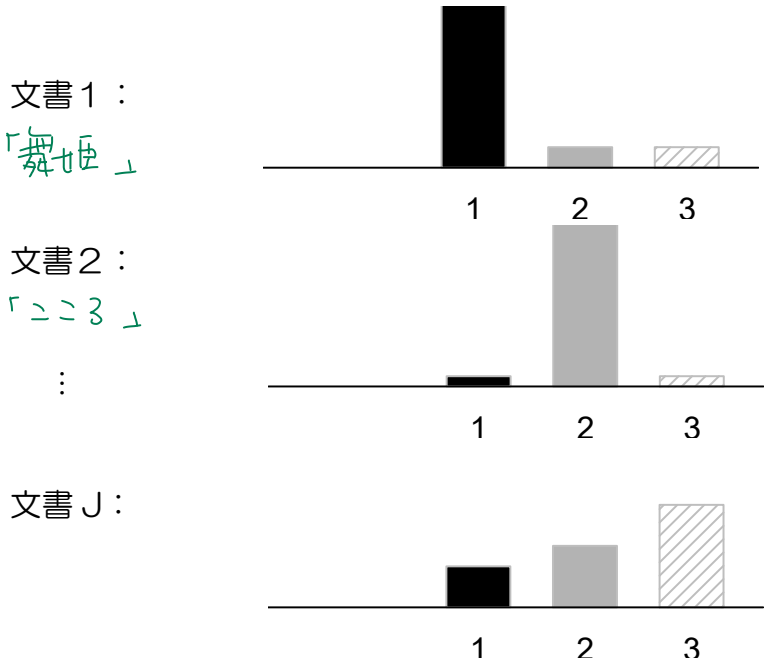
（2）トピックモデル

各「文書j」ごとに、混ざり方が異なるという仮定のもと、混ざり方（トピック分布）、混ざる前の構造（単語分布）を想定し、推定する統計モデル。

① 混ざる前の構造の分析：単語分布



② 混ざり方の分析：トピック分布



📖 ノート１　混合分布

正規分布に従うと仮定してきたけど、観測できないｘごとに正規分布になってるんじゃない？

（１）（有限）混合モデル (Finite) Mixture Model

これは、（有限の）複数のモデルをある割合（混合比）に基づいて混ぜた確率モデルのこと。

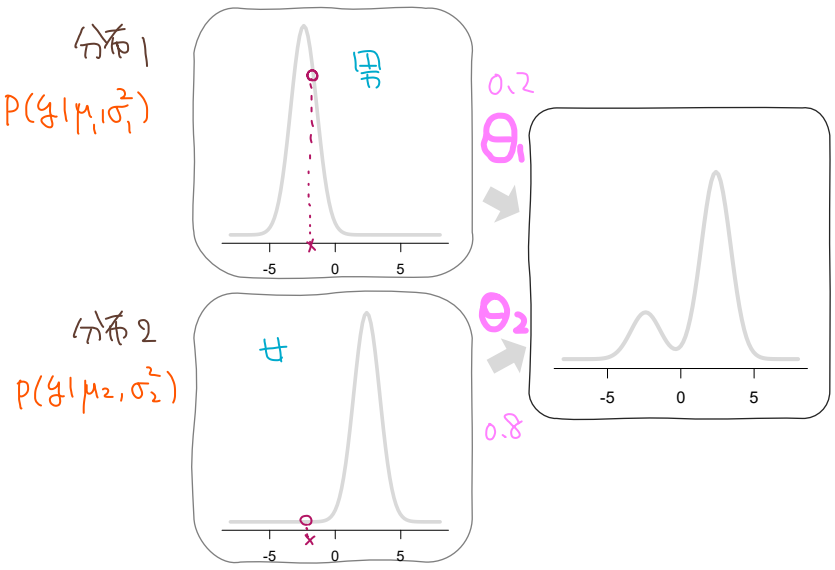
この１ートでは
$\theta_1, \theta_2, \cdots$ という記号を使う

（２）連続型：混合正規分布

これは、（有限の）複数の正規分布をある割合（混合比）に基づいて混ぜた確率分布のこと。

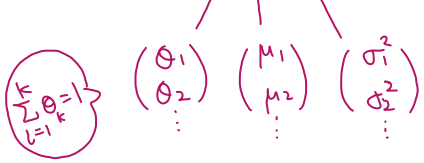
（ケース１）二個混ぜた場合

$$p(y|\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \theta_1 \times p(y|\mu_1, \sigma_1^2) + \theta_2 \times p(y|\mu_2, \sigma_2^2)$$



（ケース２）K 個混ぜた場合

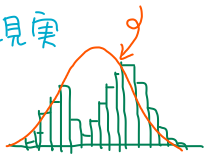
$$p(y|\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_{k=1}^{K} \theta_k \times p(y|\mu_k, \sigma_k^2)$$



Ⓟ 懸念事項

① 想定：正規分布

② 現実



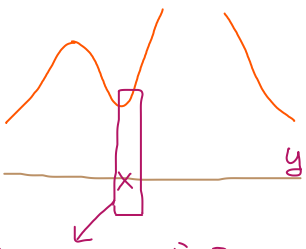
③ よりよい想定

２つの正規分布が混ざっているのでは？

⇒ 男女別にデータを取ったら二つに分けられた

⇒ 何らかの現実的な制約から「分ける」ための情報が得られないとき，混合分布が登場

Ⓟ 混合分布の密度



yが−1を取る密度は？

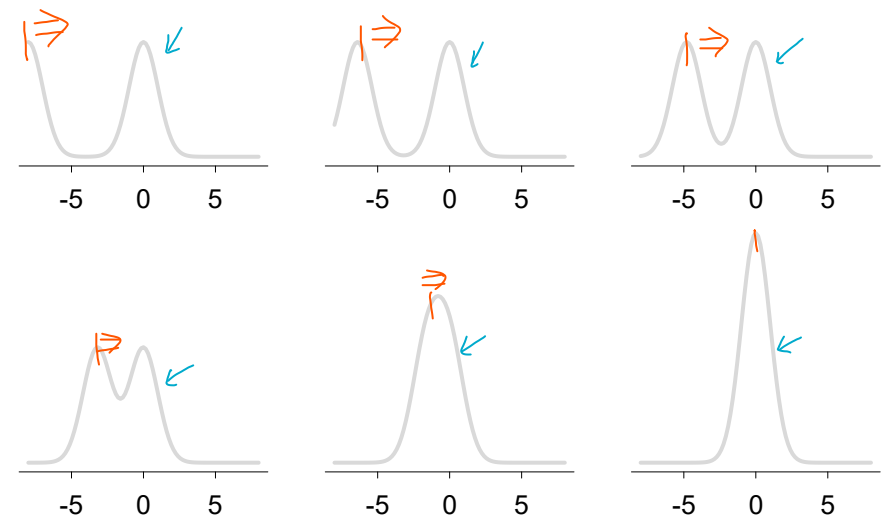
(ケース1)「男性」に所属
θ_1（全体の20%）
$p(y|\mu_1,\sigma_1^2)$

＋

(ケース2)「女性」に所属
θ_2（全体の80%）
$p(y|\mu_2,\sigma_2^2)$

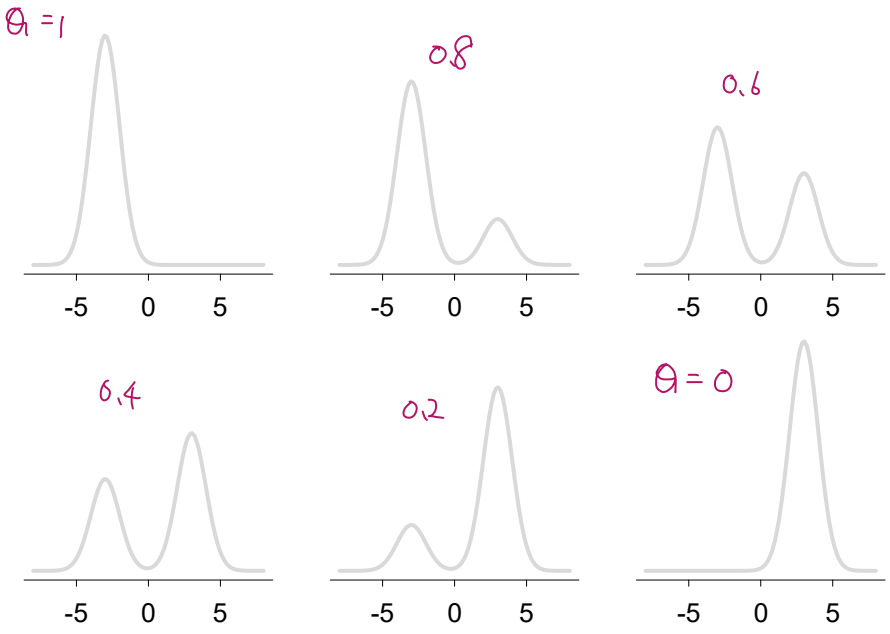
混合正規分布の例１：中心の位置をずらした例

$$0.5 \times p(y|\mu = k, \sigma^2 = 1) \quad + \quad 0.5 \times p(y|\mu = 0, \sigma^2 = 1)$$



混合正規分布の例２：混合比を変えた例

$$\theta_1 \times p(y|\mu = -3, \sigma^2 = 1) \quad + \quad \theta_2 \times p(y|\mu = 0, \sigma^2 = 1)$$



（3）離散型 1：混合カテゴリカル分布

これは、（有限の）複数のカテゴリカル分布をある割合（混合比）に基づいて混ぜた確率分布のこと。

Ⓟ ベクトルの記法

$\boldsymbol{y} \rightarrow \mathbf{y}$

$\boldsymbol{\pi} \rightarrow \boldsymbol{\pi}$

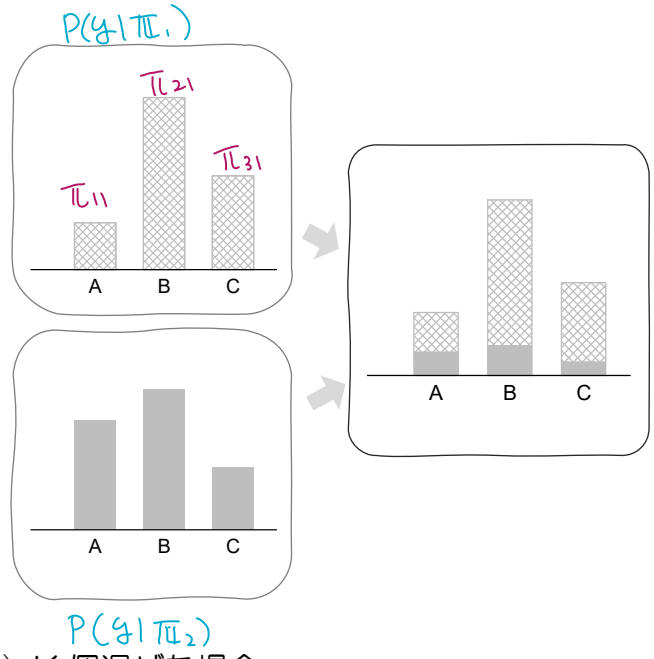
$\boldsymbol{\phi} \rightarrow \boldsymbol{\phi}$

白抜き文字

（ケース１）二個混ぜた場合

$\boldsymbol{\pi}_1 = \begin{pmatrix} \pi_{11} \\ \pi_{21} \\ \pi_{31} \end{pmatrix}$

$$p(\boldsymbol{y}|\theta_1, \theta_2, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = \theta_1 \times p(\boldsymbol{y}|\boldsymbol{\pi}_1) + \theta_2 \times p(\boldsymbol{y}|\boldsymbol{\pi}_2)$$



（ケース２）K 個混ぜた場合

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \theta_k \times p(\boldsymbol{y}|\boldsymbol{\pi}_k)$$

「単語データ」を従属変数にする

i 番目の単語を次のようにベクトルで表記する。

$$\boldsymbol{y_i} = \overset{\begin{matrix}\text{A} & \text{B} & \text{C}\end{matrix}}{(0 \quad 1 \quad 0)^T} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \end{pmatrix} \begin{matrix} \text{A} \\ \text{B} \\ \text{C} \end{matrix}$$

Ⓟ 復習：カテゴリカル分布の密度

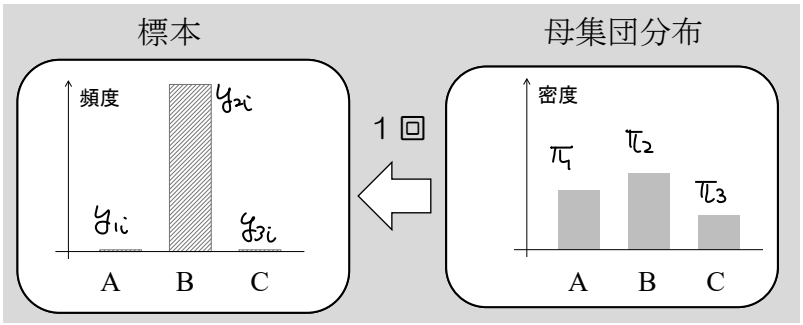
y_{1i} → π_1
y_{2i} → π_2
y_{3i} → π_3

⇓ 場合わけせずに表現したい！

$\pi_1^{y_{1i}}$ $\pi_2^{y_{2i}}$ $\pi_3^{y_{3i}}$

(例) π_1^0 π_2^1 π_3^0
$= 1 \times \pi_2 \times 1$
$= \pi_2$

密度関数１：カテゴリカル分布



① データ

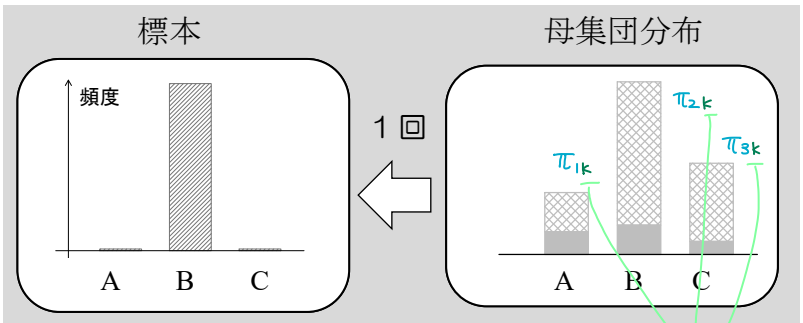
$$\boldsymbol{y}_i = \begin{pmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{Vi} \end{pmatrix}$$

② 確率密度関数

$$p(\boldsymbol{y}_i|\boldsymbol{\pi}) = \pi_1^{y_{1i}}\pi_2^{y_{2i}}\cdots\pi_V^{y_{Vi}}$$

$$= \prod_{v=1}^{V} \pi_v^{y_{vi}} \qquad \cdots（式１）$$

密度関数２：混合カテゴリカル分布



$$p(\boldsymbol{y}_i|\boldsymbol{\theta},\boldsymbol{\pi}) = \theta_1 \times \left(\pi_{11}^{y_{1i}}\pi_{21}^{y_{2i}}\cdots\pi_{V1}^{y_{Vi}}\right)$$
$$+ \theta_2 \times \left(\pi_{12}^{y_{1i}}\pi_{22}^{y_{2i}}\cdots\pi_{V2}^{y_{Vi}}\right)$$
$$+ \cdots$$
$$+ \theta_K \times \left(\pi_{1K}^{y_{1i}}\pi_{2K}^{y_{2i}}\cdots\pi_{VK}^{y_{Vi}}\right)$$

どのグループ（トピック）に属するのかを表わす添え字

$$= \sum_{k=1}^{K} \theta_k \left(\prod_{v=1}^{V} \pi_{vk}^{y_{vi}}\right) \qquad \cdots（式２）$$

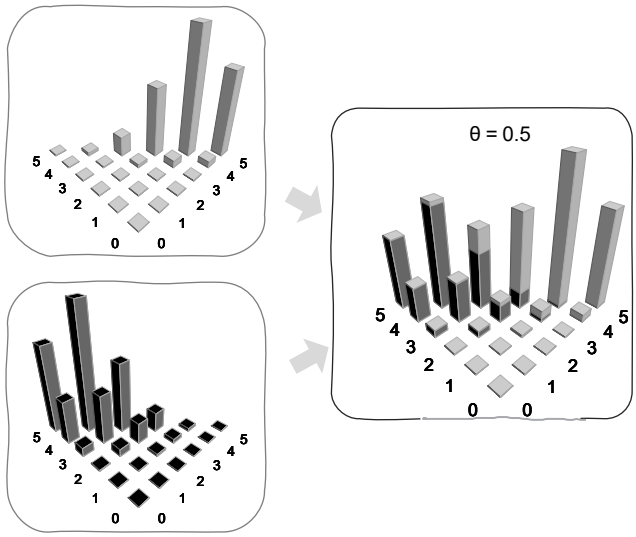
（4）離散型２：混合多項分布

これは、（有限の）複数の多項分布をある割合（混合比）に基づいて混ぜた確率分布のこと。

（ケース１）二個混ぜた場合

混合比 × 混合する前の密度

$$p(\boldsymbol{y}|\theta_1,\theta_2,\boldsymbol{\pi}_1,\boldsymbol{\pi}_2,n) = \theta_1 \times p(\boldsymbol{y}|\boldsymbol{\pi}_1,n) + \theta_2 \times p(\boldsymbol{y}|\boldsymbol{\pi}_2,n)$$



（ケース２）K 個混ぜた場合

$$p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\pi},n) = \sum_{k=1}^{K} \theta_k \times p(\boldsymbol{y}|\boldsymbol{\pi}_k,n)$$

「文書データ」を従属変数にする

j 番目の文書（N_j個の単語の集合）をベクトルで表記する。

j番目の文書の1番目の単語

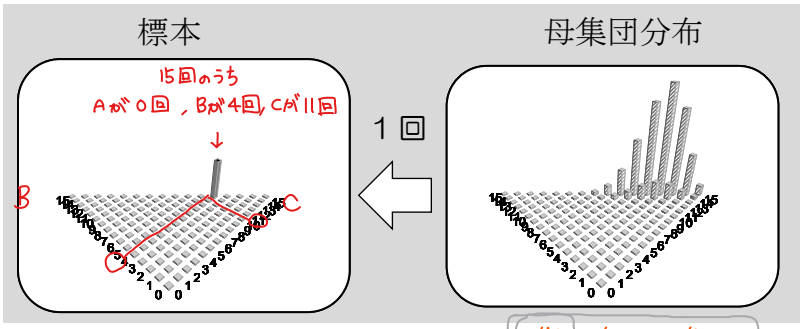
$$\boldsymbol{y_j} = (\boldsymbol{y}_{1j} \quad \boldsymbol{y}_{2j} \quad \cdots \quad \boldsymbol{y}_{Nj})^T = \begin{pmatrix} \boldsymbol{y}_{1j} \\ \boldsymbol{y}_{2j} \\ \vdots \\ \boldsymbol{y}_{ij} \\ \vdots \\ \boldsymbol{y}_{Nj} \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{y}_{1j} \\ \boldsymbol{y}_{2j} \\ \vdots \\ (0 \quad 1 \quad \cdots \quad 0)^T \\ \vdots \\ \boldsymbol{y}_{Nj} \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}_{1j} \\ \boldsymbol{y}_{2j} \\ \vdots \\ (y_{1ij} \quad y_{2ij} \quad \cdots \quad y_{vij} \quad \cdots \quad y_{Vij})^T \\ \vdots \\ \boldsymbol{y}_{Nj} \end{pmatrix}$$

A B C … V

0か1どれか1つが"1"で、残りは"0".

密度関数１：多項分布



① データ

$$\boldsymbol{y}_j = \begin{pmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{Nj} \end{pmatrix}$$

y_{1j}：1番目の単語
y_{2j}：2番目の単語
y_{ij}：i番目の単語
y_{Nj}：N_j番目の単語

j番目の諸

$$\pi_1^{y_{1j}} \pi_2^{y_{2j}} \cdots \pi_V^{y_{Vj}}$$

$$\pi_1^{y_{1j}+y_{2j}+\cdots+y_{Vj}} = \pi_1^{N_j}$$

② 確率密度関数

組み合わせを表わす部分

$$p(\boldsymbol{y}_j|\boldsymbol{\pi}) = \frac{N_j!}{N_{1j}!\,N_{2j}!\cdots N_{Vj}!} \times \prod_{i=1}^{N_j} \pi_1^{y_{1ij}} \pi_2^{y_{2ij}} \cdots \pi_V^{y_{Vij}}$$

行ごとの計算

総積記号で書き直した

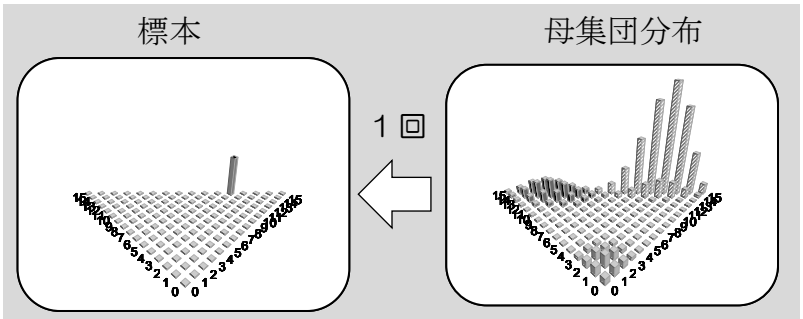
$$= (\text{定数}) \times \prod_{i=1}^{N_j} \prod_{v=1}^{V} \pi_v^{y_{vij}}$$

$$= (\text{定数}) \times \prod_{v=1}^{V} \pi_v^{N_{vj}} \qquad \cdots (式3)$$

$N_{vj} = y_{vij} + y_{vaj} + \cdots + y_{vNj}$

列ごとの計算

密度関数２：混合多項分布



$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\pi}, n) = \sum_{k=1}^{K} \theta_k \times \left((\text{定数}) \times \prod_{v=1}^{V} \pi_v^{N_{vj}} \right) \qquad \cdots (式4)$$

(式3)

（式３）の計算の詳細

$$\prod_{i=1}^{N_j} \pi_1^{y_{1ij}} \pi_2^{y_{2ij}} \cdots \pi_V^{y_{Vij}}$$

$$= \begin{array}{l} \quad (\quad \pi_1^{y_{11j}} \quad \pi_2^{y_{21j}} \quad \cdots \quad \pi_V^{y_{V1j}} \quad) \\ \times (\quad \pi_1^{y_{12j}} \quad \pi_2^{y_{22j}} \quad \cdots \quad \pi_V^{y_{V2j}} \quad) \\ \times \cdots \\ \times (\quad \pi_1^{y_{1ij}} \quad \pi_1^{y_{2ij}} \quad \cdots \quad \pi_1^{y_{Vij}} \quad) \\ \times \cdots \\ \times (\quad \pi_1^{y_{1N_jj}} \quad \pi_2^{y_{2N_jj}} \quad \cdots \quad \pi_V^{y_{VN_jj}} \quad) \end{array}$$

$$= \pi_1^{y_{11j}+y_{12j}+\cdots+y_{1N_jj}} \pi_2^{y_{21j}+y_{22j}+\cdots+y_{2N_jj}} \cdots \pi_V^{y_{V1j}+y_{V2j}+\cdots+y_{VN_jj}}$$

$$= \pi_1^{N_{1j}} \pi_2^{N_{2j}} \cdots \pi_V^{N_{Vj}}$$

$$= \prod_{v=1}^{V} \pi_v^{N_{vj}}$$

最 初 の等号：総積記号を分解して書き下した。

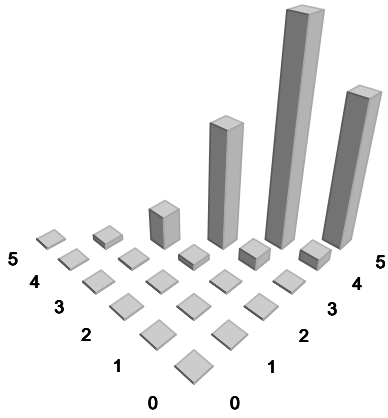
二番目の等号：縦方向に見方を変えてπごとに見直した。

三番目の等号：j 番目の文書に出てくる単語vの数をN_{vj}とおいた。ここで、y_{vij}は、0 か 1 の二通りしかないことに注意！

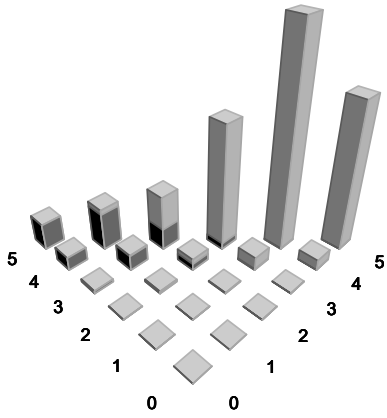
四番目の等号：総積記号を使って、コンパクトに表現した。

混合多項分布の例（二つの多項分布を混ぜた例）

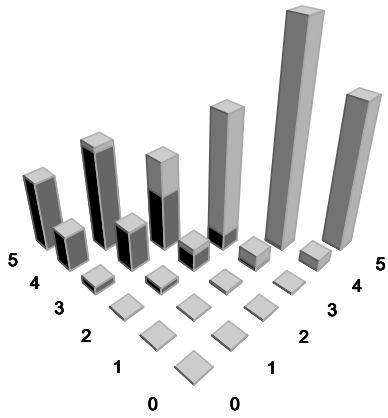
θ = 0



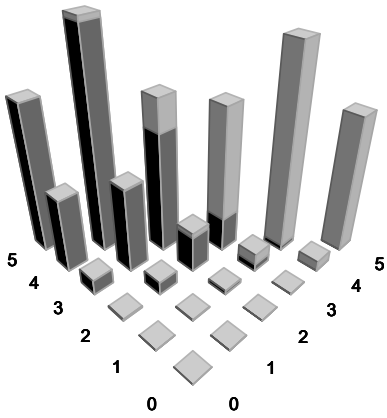
θ = 0.2



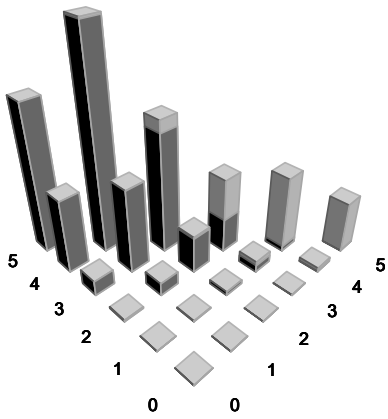
θ = 0.4



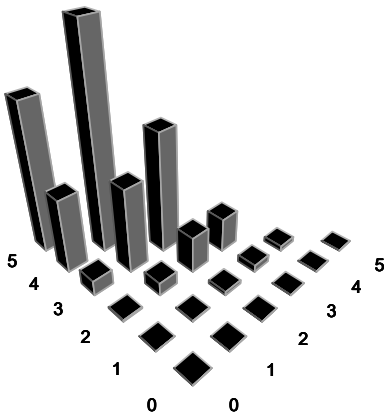
θ = 0.6



θ = 0.8



θ = 1



📖 ノート２　母集団：統計モデル

得られた文書（データ）を生成する確率モデルを作ろう！

↳ カテゴリカル分布に従うOOの集合

（１）ユニグラムモデル ＝ 多項ロジスティック切片モデル

文書 j の i 番目のデータ（単語）が独立に同一のカテゴリカル分布から生まれると仮定。

$$\boldsymbol{y}_{ij} \sim Cat(\boldsymbol{\phi})$$

Ⓟ 記法：πとφ

π　pi

φ　phi

ここからは，総積記号とまぎらわしくなるので πの代わりに φを用いる

Bag-of-Words

これは、「文」から「構造」を取り除いて作られた、「語」の集合。重複が許される多重集合である。

(1) [What I found] was [a huge dog [with a long tail]] .

(2) { a, a, dog, found, huge, I, long, tail, was, what, with }

「文書データ」を従属変数にする

Bag-of-words 表現によって、「文書データ」をベクトルと見なす。

(3) $\boldsymbol{y}_j$ ＝ （j番目の文書）

$$= \begin{pmatrix} \boldsymbol{y}_{1j} \\ \boldsymbol{y}_{2j} \\ \boldsymbol{y}_{3j} \\ \boldsymbol{y}_{4j} \\ \boldsymbol{y}_{5j} \\ \boldsymbol{y}_{6j} \\ \boldsymbol{y}_{7j} \\ \boldsymbol{y}_{8j} \\ \boldsymbol{y}_{9j} \\ \boldsymbol{y}_{10j} \\ \boldsymbol{y}_{11j} \end{pmatrix} = \begin{pmatrix} & \text{a} & \text{dog} & \text{find} & \text{huge} & \text{I} & \text{long} & \text{tail} & \text{be} & \text{what} & \text{with} \\ (0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0)^T \\ (0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0)^T \\ (0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0)^T \\ (0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0)^T \\ (1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0)^T \\ (0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0)^T \\ (0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0)^T \\ (0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1)^T \\ (1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0)^T \\ (0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0)^T \\ (0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0)^T \end{pmatrix}$$

（i番目の文書のj番目の単語）

（データが一単語だけのとき）

i 番目の単語が産出される確率：

$$p(\boldsymbol{y}_{ij}|\boldsymbol{\phi}) = \phi_1^{y_{1ij}}\phi_2^{y_{2ij}}\cdots\phi_V^{y_{Vij}}$$

$$= \prod_{v=1}^{V}\phi_v^{y_{vij}}$$

（データがn単語あるとき）

1 番目から n 番目の単語が同時に産出される確率：

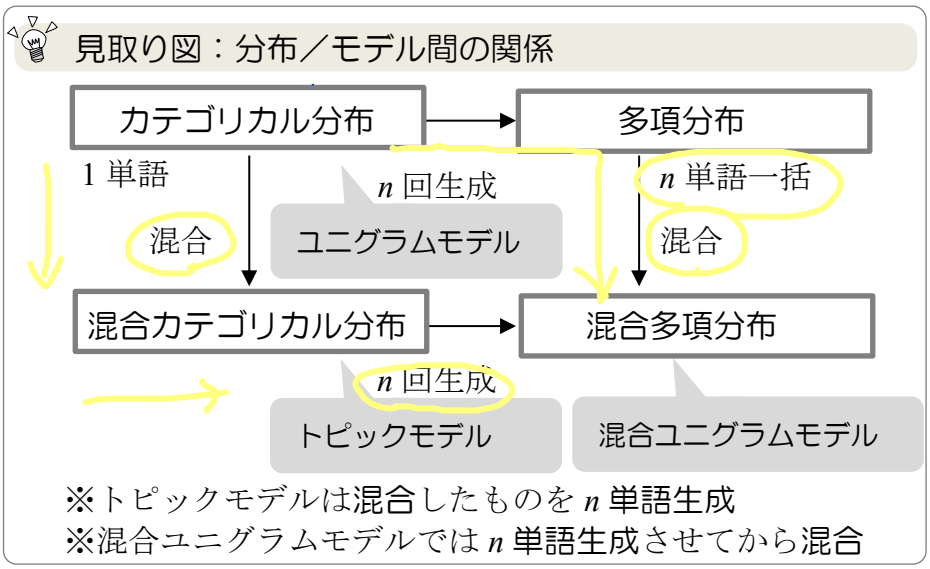
$$p(\boldsymbol{y}_j|\boldsymbol{\phi}) = p(\boldsymbol{y}_{1j}|\boldsymbol{\theta},\boldsymbol{\phi})p(\boldsymbol{y}_{2j}|\boldsymbol{\theta},\boldsymbol{\phi})\cdots p(\boldsymbol{y}_{nj}|\boldsymbol{\theta},\boldsymbol{\phi})$$

$$= \left(\prod_{v=1}^{V}\phi_v^{y_{v1j}}\right)\left(\prod_{v=1}^{V}\phi_v^{y_{v2j}}\right)\cdots\left(\prod_{v=1}^{V}\phi_v^{y_{vnj}}\right)$$

$$= \phi_1^{y_{11j}}\phi_2^{y_{21j}}\cdots\phi_V^{y_{V1j}} \times \phi_1^{y_{12j}}\phi_2^{y_{22j}}\cdots\phi_V^{y_{V2j}} \times \phi_1^{y_{1nj}}\phi_2^{y_{2nj}}\cdots\phi_V^{y_{Vnj}}$$

$$= \phi_1^{y_{11j}+y_{12j}+\cdots+y_{1nj}}\phi_2^{\sum_{i=1}^{n}y_{2ij}}\cdots\phi_V^{\sum_{i=1}^{n}y_{Vij}}$$

$$= \left(\prod_{v=1}^{V}\phi_v^{\sum_{i=1}^{n}y_{vij}}\right)$$ N_v



※トピックモデルは混合したものを n 単語生成

※混合ユニグラムモデルでは n 単語生成させてから混合

（2）混合ユニグラムモデル

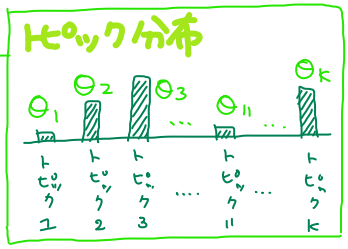
これは、単語分布が文書のトピックに応じ変化するという想定した混合カテゴリカル分布を基にした統計モデル。

【ステップ１】トピックの生成　　←　混合比

j 番目の文書のトピックz_jは、1 から K の中から一つランダムに定まり、この分布をトピック分布と呼ぶ。

$$z_j \sim Cat(\boldsymbol{\theta})$$

$$p(z_{ij} = k|\boldsymbol{\theta}_j) = \theta_k$$



【ステップ２】単語の生成　　←　混合される前の分布

j 番目の文書の i 番目の単語が得られる確率はそのトピックz_jに応じたカテゴリカル分布（単語分布）に沿う。

$$\boldsymbol{y}_{ij} \sim Cat\left(\boldsymbol{\phi}_{z_j}\right)$$

$$p(\boldsymbol{y}_{ij}|\boldsymbol{\theta}_j, \boldsymbol{\phi}, \boldsymbol{z}_{ij} = k) = \phi_{1k}^{y_{1ij}} \phi_{2k}^{y_{2ij}} \cdots \phi_{Vk}^{y_{Vij}}$$

同一文書（トピックz_j）下で n 回繰り返されると考える。

$$\boldsymbol{y}_j \sim Multi\left(\boldsymbol{\phi}_{z_j}, N_j\right)$$

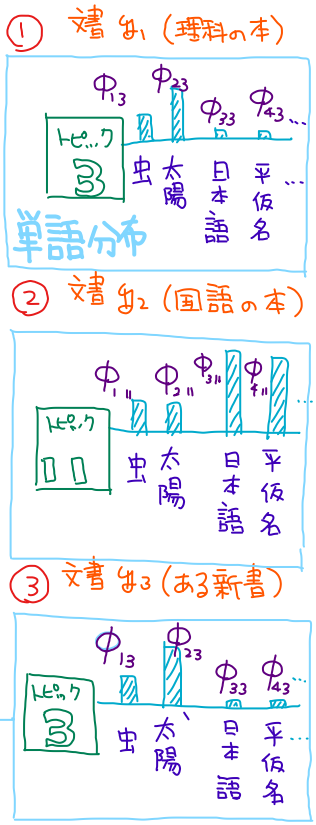
$$p(\boldsymbol{y}_j|\boldsymbol{\theta}_j, \boldsymbol{\phi}, N_j, z_{ij} = k) = \phi_{1k}^{N_{1j}} \phi_{2k}^{N_{2j}} \cdots \phi_{Vk}^{N_{Vj}}$$

※ N_{vj}：j番目の文書に含まれた単語vの生起回数

【ステップ３】確率密度関数

j 番目の文書が得られる確率密度式を、ステップ 1 とステップ 2 の積の和として求める。

$$p(\boldsymbol{y}_j|\boldsymbol{\theta}, \boldsymbol{\phi}, N_j) = \sum_{k=1}^{K} \theta_k \times p(\boldsymbol{y}_j|\boldsymbol{\phi}_k, N_j)$$



データの表記１：「文書（作品）」と「文書集合（作品集）」

文書が集まった「文書集合（作品集）」が我々が手にするデータ。これを$\boldsymbol{y}$とし、その中の各「文書（作品）」を$\boldsymbol{y_1}, \boldsymbol{y_2}, \cdots$のように表す。

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_J} \end{pmatrix} = \begin{pmatrix} \text{羅生門} \\ \text{トロッコ} \\ \vdots \\ \text{蜘蛛の糸} \end{pmatrix}$$

データの表記２：「単語」と「文書（作品）」

文書$\boldsymbol{y_j}$は、j 番目の文書の i 番目の単語を表すベクトルを考え、これを$\boldsymbol{y}_{ij}$とおく。

$$\boldsymbol{y_1} = \text{羅生門} = \begin{pmatrix} \boldsymbol{y}_{11} \\ \boldsymbol{y}_{21} \\ \boldsymbol{y}_{31} \\ \vdots \\ \boldsymbol{y}_{n_1 1} \end{pmatrix} = \begin{pmatrix} & \text{老婆} & \text{男} & \text{門} & & \text{線路} \\ (0 & 0 & 1 & \cdots & 0)^T \\ (0 & 1 & 0 & \cdots & 0)^T \\ (0 & 1 & 0 & \cdots & 0)^T \\ & & \vdots & & \\ (1 & 0 & 0 & \cdots & 0)^T \end{pmatrix}$$

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \boldsymbol{y_3} \\ \vdots \\ \boldsymbol{y_J} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} \boldsymbol{y}_{11} \\ \boldsymbol{y}_{21} \\ \boldsymbol{y}_{31} \\ \vdots \\ \boldsymbol{y}_{n_1 1} \end{pmatrix} \\ \boldsymbol{y_2} \\ \boldsymbol{y_3} \\ \vdots \\ \boldsymbol{y_J} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} & 1 & 2 & 3 & & \text{V} \\ & \text{老婆} & \text{男} & \text{門} & & \text{線路} \\ (0 & 0 & 1 & \cdots & 0)^T \\ (0 & 1 & 0 & \cdots & 0)^T \\ (0 & 1 & 0 & \cdots & 0)^T \\ & & \vdots & & \\ (1 & 0 & 0 & \cdots & 0)^T \end{pmatrix} \\ \boldsymbol{y_2} \\ \boldsymbol{y_3} \\ \vdots \\ \boldsymbol{y_J} \end{pmatrix}$$

（3）トピックモデル

これは、単語の生成過程のモデルとして、「文書j」ごとに異なる混ざり方をした単語分布を想定する統計モデル。

【ステップ1】トピックの生成　　← 混合比

j番目の文書の単語iのトピックz_{ij}は、文書ごとに定められるトピック分布からランダムに決まる。

$$z_{ij} \sim Cat(\boldsymbol{\theta}_j)$$

$$p(z_{ij} = k|\boldsymbol{\theta}_j) = \theta_{jk}$$

添え字がついているので文書ごとに変化するトピック分布を仮定

【ステップ2】単語の生成　　← 混合された後の分布

j番目の文書に含まれるi番目の単語が$\boldsymbol{y}_{ij} = (y_{1ij}\ y_{2ij}\ \cdots\ y_{Vij})^T$になる確率はトピック$z_{ij}$の単語分布が規定。

$$\boldsymbol{y}_{ij} \sim Cat\left(\boldsymbol{\phi}_{z_{ij}}\right)$$

$$p(\boldsymbol{y}_{ij}|\boldsymbol{\theta}_j, \boldsymbol{\phi}) = \sum_{k=1}^{K} p(z_{ij} = k|\boldsymbol{\theta}_j) p(y_{ij}|\boldsymbol{\phi}_k)$$

$$= \sum_{k=1}^{K} \theta_k (\phi_{1k}^{y_{1ij}} \phi_{2k}^{y_{2ij}} \cdots \phi_{Vk}^{y_{Vij}})$$

混合比

トピックz_{ij}の単語分布

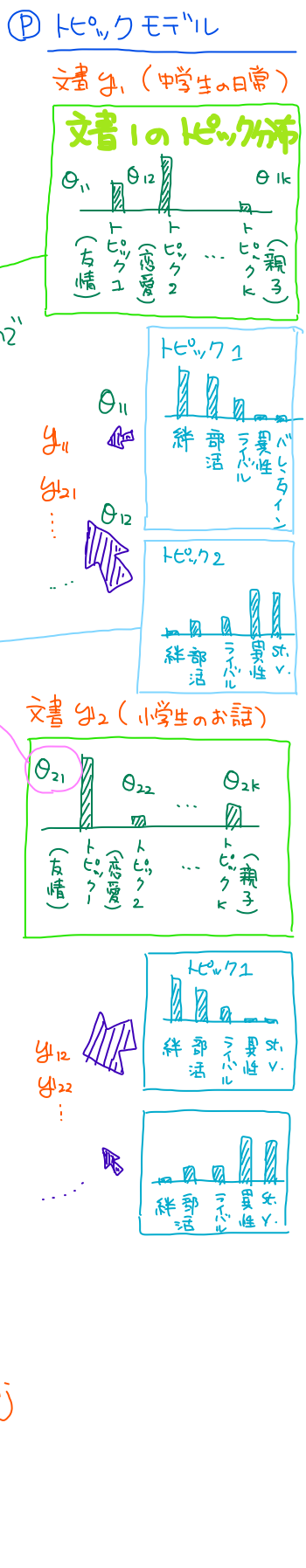
【ステップ3】確率密度関数

j番目の文書が得られる確率は、同一文書（トピックz_j）下でステップ2がn回繰り返されたものと考える。

$$p(\boldsymbol{y}_j|\boldsymbol{\theta}_j, \boldsymbol{\phi}, N_j) = \left[\sum_{k=1}^{K} \theta_k (\phi_{1k}^{y_{11j}} \phi_{2k}^{y_{21j}} \cdots \phi_{Vk}^{y_{V1j}})\right] \times \left[\sum_{k=1}^{K} \theta_k (\phi_{1k}^{y_{12j}} \phi_{2k}^{y_{22j}} \cdots \phi_{Vk}^{y_{V2j}})\right] \times \cdots \times \left[\sum_{k=1}^{K} \theta_k (\phi_{1k}^{y_{1Nj}} \phi_{2k}^{y_{2Nj}} \cdots \phi_{Vk}^{y_{VNj}})\right]$$

$$= \prod_{i=1}^{N_j} \left[\sum_{k=1}^{K} \theta_k (\phi_{1k}^{y_{1ij}} \phi_{2k}^{y_{2ij}} \cdots \phi_{Vk}^{y_{Vij}})\right]$$

←単語1　←単語2　↑単語 Nj



混合ユニグラムモデルとトピックモデルの比較

混合ユニグラムモデル

①トピックz_j　　文書jごとに決定

②トピックが生じる比率$\boldsymbol{\theta}$　　文書jには無関係

トピックモデル

①トピックz_{ij}　　文書jの各単語iごとに決定

②トピックが生じる比率$\boldsymbol{\theta_j}$　　文書jごとに決定

混合ユニグラムモデルの基本的な考え方

（例）世の作品に「恋愛」と「友情」という 2 つのトピックがあると想定する（3 つ目、例えば「生と死」を加えても話は同じ）。

「j 番目の文書の i 番目の単語が『男』である確率」

＝「① j 番目の文書のトピックが『恋愛』で、② この『恋愛』の単語分布から『男』が出る確率

＋

① j 番目の文書のトピックが『友情』で、② この『友情』の単語分布から『男』が出る確率」

$$=\theta_{\text{恋愛}} \times p(y_{ij}|z_j = \text{恋愛}) + \theta_{\text{友情}} \times p(y_{ij}|z_j = \text{友情})$$

↑混合比　↑確率密度　＋　↑混合比　↑確率密度

トピックモデルの基本的な考え方

「j 番目の文書の i 番目の単語が『男』である確率」

＝「① j 番目の文書の i 番目の単語のトピックが『恋愛』で、② この『恋愛』の単語分布から『男』が出る確率

＋

① j 番目の文書の i 番目の単語のトピックが『友情』で、② この『友情』の単語分布から『男』が出る確率」

$$=\theta_{\text{恋愛}} \times p(y_{ij}|z_{ij} = \text{恋愛}) + \theta_{\text{友情}} \times p(y_{ij}|z_{ij} = \text{友情})$$

↑混合比　↑確率密度　＋　↑混合比　↑確率密度

例題 **TASA コーパスの分析**

Steyvers and Griffiths (2007)で紹介されている TASA コーパス（Touchstone Applied Science Associates によって作成された教育的内容の 37,000 個の文書コーパス） のトピックモデルによる解析結果（※300 個のトピックを想定している）を見て、77, 82, 166 のトピックがどのようなカテゴリーを形成しているか、その潜在的意味の候補を推測してみよう。なお、問題の作成上、修正を施した箇所がある。

単語の通し番号ごとにトピックが決まる

Document #29795

Bix Beiderbecke, at age[060] fifteen[207], sat[174] on the slope[071] of a bluff[055] overlooking[027] the Mississippi[137] river[137]. He was listening[077] to music[077] coming[009] from a passing[043] riverboat. The music[077] had already captured[006] his heart[157] as well as his ear[119]. It was jazz[077]. Bix Beiderbecke had already had music[077] lessons[077]. He showed[002] promise[134] on the piano[077], and his parents[035] hoped[268] he might consider[118] becoming a concert[077] pianist[077]. But Bix was interested[268] in another kind[050] of music[077]. He wanted[268] to play[077] the cornet. And he wanted[268] to play[077] jazz[077].

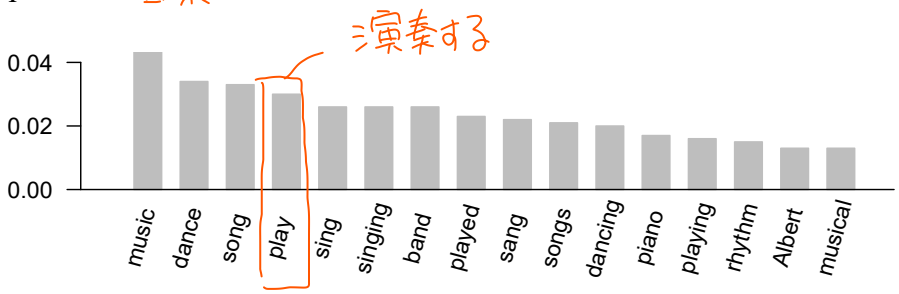
Document #1883

There is a simple[050] reason[106] why there are so few periods[078] of really great theater[082] in our whole western[046] world. Too many things[300] have to come right at the very same time. The dramatists must have the right actors[082], the actors[082] must have the right playhouses, the playhouses must have the right audiences[082]. We must remember[288] that plays[082] exist[143] to be performed[077], not merely[050] to be read[254]. (even when you read[254] a play[082] to yourself, try[288] to perform[062] it, to put[174] it on a stage[078], as you go along.)

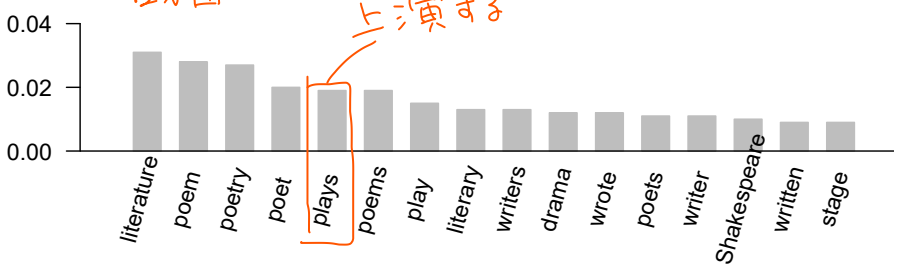
Document #21359

Jim has a game[166] book[254]. Jim reads[254] the book[254]. Jim sees[081] a game[166] for one. Jim plays[166] the game[166]. Jim likes[081] the game[166] for one. The game[166] book[254] helps[081] Jim. Don comes[040] into the house[038]. Don and Jim read[254] the game[166] book[254]. The boys[020] see a game[166] for two. The two boys[020] play[166] the game[166]. The boys[020] play[166] the game[166] for two. The boys[020] like the game[166].

Topic 77 音楽



Topic 82 戯曲



Topic 166 球技