

# PLSI

(probabilistic latent semantic  
analysis)

正田 備也

[masada@rikkyo.ac.jp](mailto:masada@rikkyo.ac.jp)

# Contents

## 混合多項分布の問題点

PLSI

# 混合多項分布

- ▶ 混合多項分布モデルでは、一つ一つの文書がそれ全体で、意味的なまとまりを持つ
  - ▶ ニュース記事であれば、記事まるごと、特定のカテゴリ（ex. 政治、経済、スポーツ、etc）に割り振られる。
- ▶ つまり、一つの文書内は意味的に均一だと、仮定している
- ▶ しかし、この仮定は現実の文書の実態に合わない
  - ▶ 文書は複数の話題を含みうるので。

# 混合多項分布の改良

- ▶ カテゴリの違いは、混合多項分布と同様、語彙集合上に定義された多項分布（単語多項分布）の違いとして表す
  - ▶ 政治について書かれたテキストと、スポーツについて書いたテキストとでは、どの単語がどのくらいの確率で出現するかが異なる、という考え方。
- ▶ そこで、一つの文書に含まれる単語トークン群が、唯一の単語多項分布からではなく、複数の単語多項分布から生成されると、仮定する→PLSA モデル
  - ▶ 同じ文書内に、異なる単語多項分布に由来する単語トークンが混ざっていてもよい、という考え方。

## 混合多項分布



## PLSI



Figure: 混合多項分布と PLSA の違い

Shanghai is the largest city in China, located on its eastern coast at the outlet of the Yangtze River. Originally a fishing and textiles town, Shanghai grew in importance in the 19th century. In 2005 Shanghai became the world's busiest cargo port. The city is an emerging tourist destination renowned for its historical landmarks such as the Bund and Xintiandi, its modern and

Figure: PLSA では同じ文書の単語トークンが複数の単語多項分布に由来しうる

# Contents

混合多項分布の問題点

PLSI

# PLSA (probabilistic latent semantic analysis)

- ▶ LSA(latent semantic analysis) を probabilistic にしたモデル
  - ▶ LSA については次スライドの図を参照（実態は単なる SVD）
- ▶ 同じ文書内でも、単語トークンが異なる単語多項分布から生成される
- ▶ どの単語多項分布がどのくらいの確率で使われるかが、文書によって異なる
- ▶ PLSA における単語多項分布を、トピック (topic) と呼ぶ
  - ▶ PLSA は最もシンプルなトピックモデル



# LSA の概念図

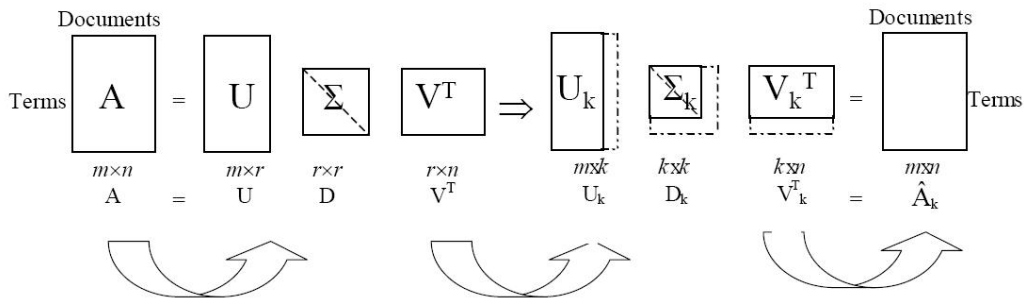


Figure: LSA の概念図

- ▶ 左から順に、データ行列の特異値分解、低ランク近似、元のデータ行列の再現
- ▶  $m$  が語彙サイズ、 $n$  が文書数、 $k$  がトピック数 ( $r$  は元のデータ行列のランク)

# 確率モデルとしてのPLSA

- ▶ PLSA は、行列分解ではなく、観測データの生成モデル
- ▶ 文書  $d$  の  $i$  番目の単語として  $w$  が現れる確率  $p_d(x_i = w)$  を、PLSA では以下のようにモデリングする

$$p_d(x_i = w) = \sum_{z_i=1}^K p(x_i = w | z_i) p_d(z_i) \quad (1)$$

- ▶  $p_d(z = k)$  は、文書  $d$  内の単語がトピック  $k$  を扱っている確率
- ▶  $p(x = w | z = k)$  は、トピック  $k$  を扱うときに単語  $w$  が使われる確率