## **PLSI**

# (probabilistic latent semantic indexing)

正田 備也

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→PLSI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。

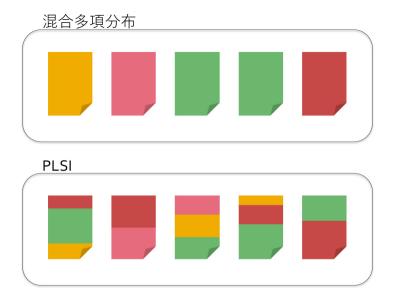


Figure: 混合多項分布と PLSI の違い



Figure: PLSI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

#### Contents

混合多項分布の問題点

**PLSI** 

## PLSI (probabilistic latent semantic indexing)

- ▶ LSI(latent semantic indexing) 確率モデル化したモデル
  - ▶ LSI については、説明を割愛(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
- ▶ PLSIにおける単語多項分布を、トピック (topic)と呼ぶ
  - ▶ PLSI は最もシンプルなトピックモデル