NTT

トピックモデルの応用：イントロダクション

NTT コミュニケーション科学基礎研究所
石黒　勝彦

2013/01/15-16 統計数理研究所　会議室1

このスライドの“トピック”

- まずは本日の講義のイントロダクションです

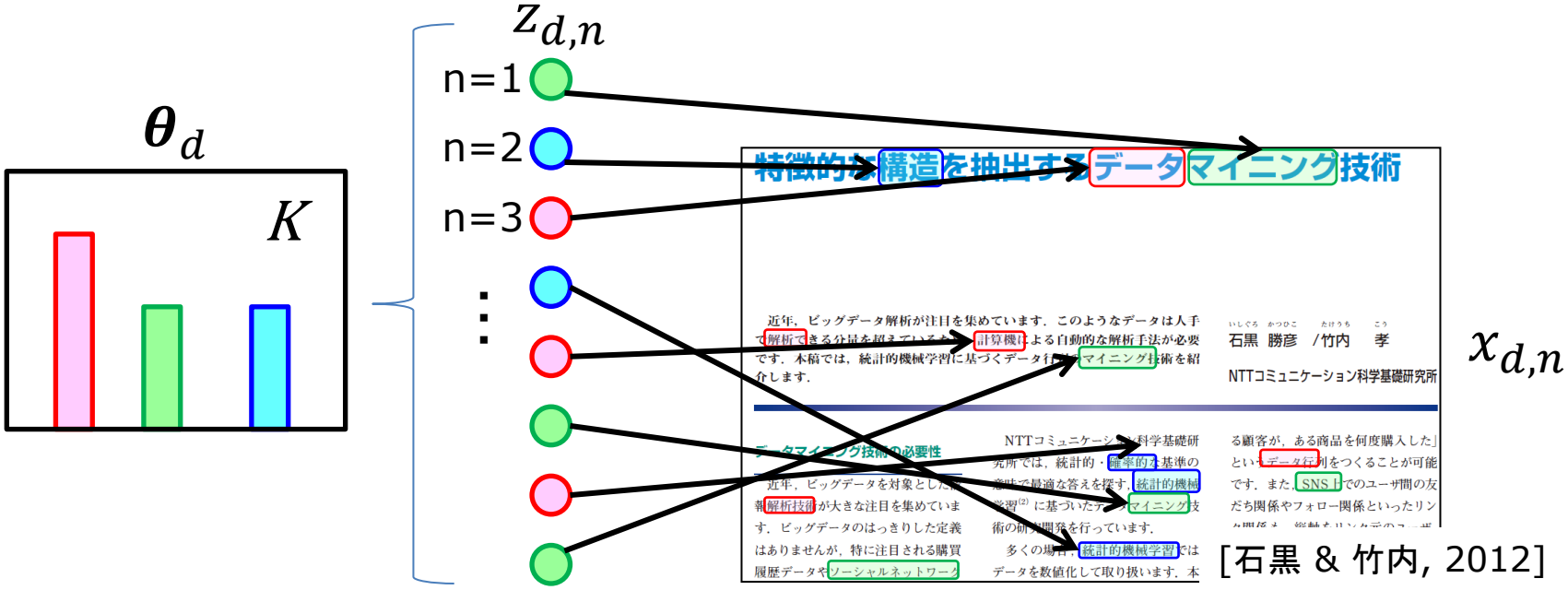
- LDAの復習
  - 使用する変数, モデルなど
  - 解法: Gibbs sampler, VB-EM
- 今日の講義全体に関する注意
  - notation policyなど
  - 参考文献

本日の予定

- LDAの拡張手法と応用について紹介します
  - 基本的には、様々な論文の説明になります
- 午前: トピックモデル(LDA)の拡張モデル
- 午後: トピックモデルの各種ドメインデータへの応用

Latent Dirichlet Allocaion [Blei, 2003]

- 様々な離散データに隠された潜在的なトピックを推定するベイジアンモデル



[石黒 & 竹内, 2012]

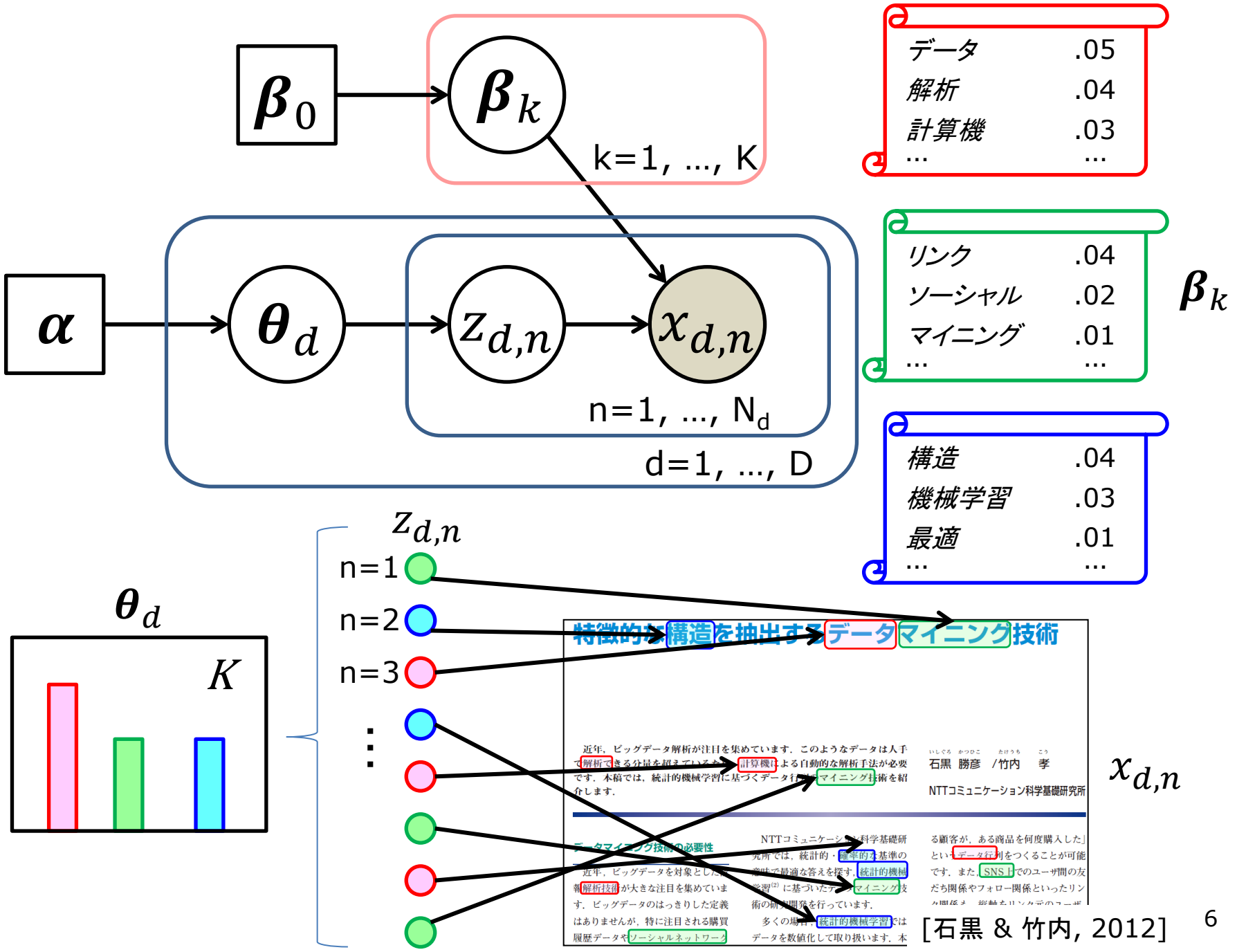
インデックス、定数、変数

- 文書インデックス d
- トピックインデックス k
- 単語インデックス n

- 文章数 D
- トピック数 K
- 文書 d 中の単語数 N_d
- 単語の種類 V

- 観測された単語 x
- 単語のtopic assignment z
- 文書のtopic proportion θ
- トピックのword proportion β

- θの事前分布パラメータ α
- βの事前分布パラメータ β_0


β_0
β_k
k=1, …, K
α
θ_d
$z_{d,n}$
$x_{d,n}$
n=1, …, N_d
d=1, …, D
データ .05
解析 .04
計算機 .03
… …
リンク .04
ソーシャル .02
マイニング .01
… …
β_k
構造 .04
機械学習 .03
最適 .01
… …
$z_{d,n}$
n=1
n=2
n=3
θ_d
K
特徴的な構造を抽出するデータマイニング技術
$x_{d,n}$

[石黒 & 竹内, 2012]

生成モデル

for 文書d = 1, 2, …, D_t

topic proportion $\boldsymbol{\theta}_d|\boldsymbol{\alpha}\sim\mathrm{Dir}(\boldsymbol{\alpha})$

for 単語n = 1, 2, …, N_d

topic-word assignment

$$z_{d,n}|\boldsymbol{\theta}_d\sim\mathrm{Mult}(\boldsymbol{\theta}_d)$$

word observation

$$x_{d,n}|z_{d,n},\{\boldsymbol{\beta}_k\}\sim\mathrm{Mult}(\boldsymbol{\beta}_{z_{d,n}})$$

for トピック k = 1, 2, …, K

topic-word proportion $\boldsymbol{\beta}_k|\boldsymbol{\beta}_0\sim\mathrm{Dir}(\boldsymbol{\beta}_0)$

topic-word proportionは事前分布を仮定しない場合も多々あります

Gibbs sampler

- 最も正確な解を得ることができる解法です
- 各文書dの単語nを一つずつトピックkに割り当てていきます

$$p\left(z_{d,n}=k|x_{d,n}=w,\boldsymbol{X}_{\neg(d,n)},\boldsymbol{Z}_{\neg(d,n)},\boldsymbol{\alpha},\boldsymbol{\beta_0}\right)\propto \frac{m_{dk}+\alpha_k}{\sum_{k'}(m_{dk'}+\alpha_{k'})}\frac{m_{kw}+\beta_{0,w}}{\sum_{w'}\left(m_{kw'}+\beta_{0,w'}\right)}$$

文書 dから トピック kが 生成される確率

トピック kから 単語 wが 生成される確率

m_{dk}: 文書 d 内でトピック k にアサインされた観測量のカウント($x_{d,n}$を除く)

m_{kw}: 文書全体でトピック k にアサインされた観測量のうち単語 w だった観測量のカウント($x_{d,n}$を除く)

変分ベイズ法(VB-EM)

- 文書データの周辺化尤度をJensen不等式で下から押さえて、それを最大化する変分事後分布$q()$を求める手法です

$$\log p(\boldsymbol{X}) = \iint \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta})\, d\boldsymbol{Z} d\boldsymbol{\theta}$$
$$\geq \iint q(\boldsymbol{Z}, \boldsymbol{\theta}|\boldsymbol{X}) \log \frac{p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta})}{q(\boldsymbol{Z}, \boldsymbol{\theta}|\boldsymbol{X})}\, d\boldsymbol{Z} d\boldsymbol{\theta}$$

$$q\left(z_{d,n} = k\right) \propto \beta_{k,x_{d,n}} \exp\left(\Psi\left(\hat{\theta}_{d,k}\right)\right)$$

$$\hat{\theta}_{d,k} = \alpha_k + \sum_n q\left(z_{d,n} = k\right)$$

Notationについて

- 直観的な理解のために、できる限り同じ意味を持つ量は同じ名前の変数・インデックスで表します
- また、一部のモデルについてはよりわかりやすく等価なモデルで説明します

- したがって、原論文とは変数の名前やグラフィカルモデルの形が違う可能性があります

全体を通じての重要な参考文献

- Blei et al, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
  - LDAの原論文です

全体を通じての重要な参考文献

- ビショップ、”パターン認識と機械学習”, 丸善出版, 2012
  - “PRML本”: 和書の中では、現状、機械学習に関するもっとも良い本の一つです

引用及び参考文献

- [Blei, 2003] Blei et al, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
- [PRML] Blei and Lafferty, “A Correlated Topic Model of Science”, The Annals of Applied Statistics, Vol. 1(1), pp. 17-35, 2007.
- [石黒 & 竹内, 2012] 石黒, 竹内, “特徴的な構造を抽出するデータマイニング技術”, NTT技術ジャーナル, Vol. 24, No. 9, 2012.