NTT

トピックモデルの応用：教師情報・補助情報つきモデル

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”

- いわゆる文書データ以外の補助情報・クラス情報が得られる場合のトピックモデル活用法の例です

教師なし学習 (unsupervised learning)

- 「正解」信号となる情報がない設定でモデルを学習したりすることです
- LDA(トピックモデル)は一般に教師なし学習のフレームワークで使われます
  - 文書データだけが与えらえた状態で、まったく未知のトピックを学習しています
- 教師なし学習は基本的に難しいので、高い精度を出すLDAは重宝されます

しかし、教師情報・補助情報があるなら使えばいいのです

- 全てをLDAで、つまり教師なし学習でまとめる必要はありません

- 教師信号・補助情報があるならば、モデル全体の「部品」としてトピックモデルを利用すれば十分です

Supervised LDA
[Blei & McAuliffe, 2008]

Blei and McAuliffe,
"Supervised Topic Models",
in Advances in Neural Information Processing Systems 20
(Proc. NIPS), 2008.

補助情報が観測可能な文書データ

- 典型的には各種レビュー記事を想像すると分かり易いと思います
  - その他、文章のsentiment analysis, 学生のレポート採点データ、・・・
- 当然、文書をモデル化する「だけ」がお仕事のLDAでは表現できません

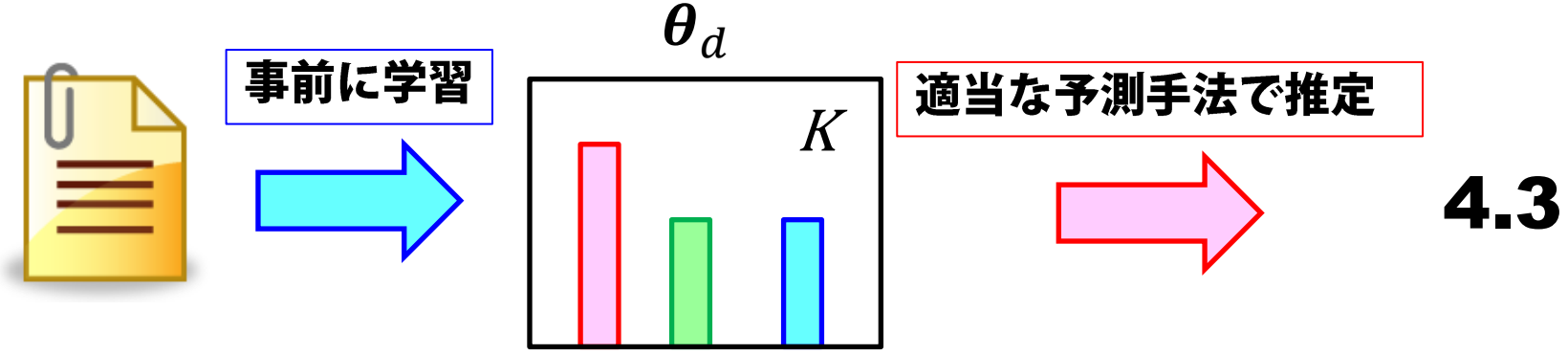
文書と単語の山

補助情報・評価情報・…

4.3

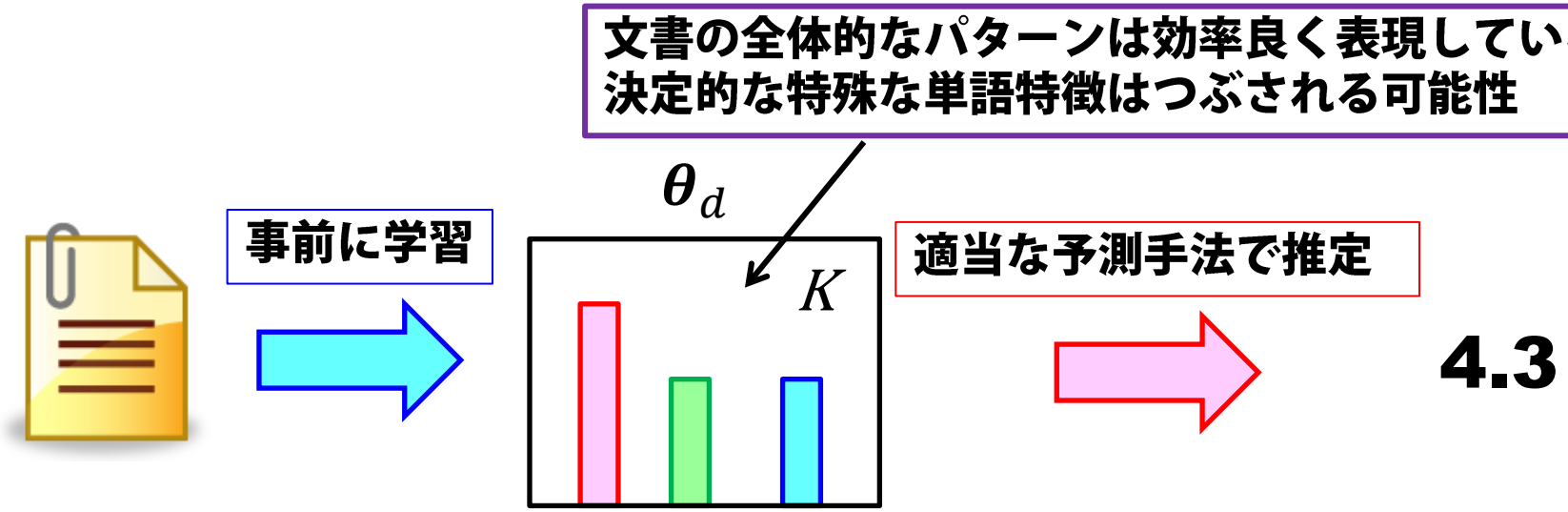
「２段構え」ではダメですか？

- LDAによるトピック学習ののち、各文書のトピックで補助情報を回帰・予測 [Blei, 2003]
- 期待: トピックは文書の中身を上手く圧縮表現しているから、上手く行くはず！ ☺
- 実際: 上手くいかない ☹



理由： 識別に必要な情報がつぶされる

- LDAを含む次元削減手法は、データの全体的な分布を効率よく表現するために、小さい情報量を削除する
- 評価値の予測に必要な特徴量がごく一部だと、「小さい方」に入って削除される可能性



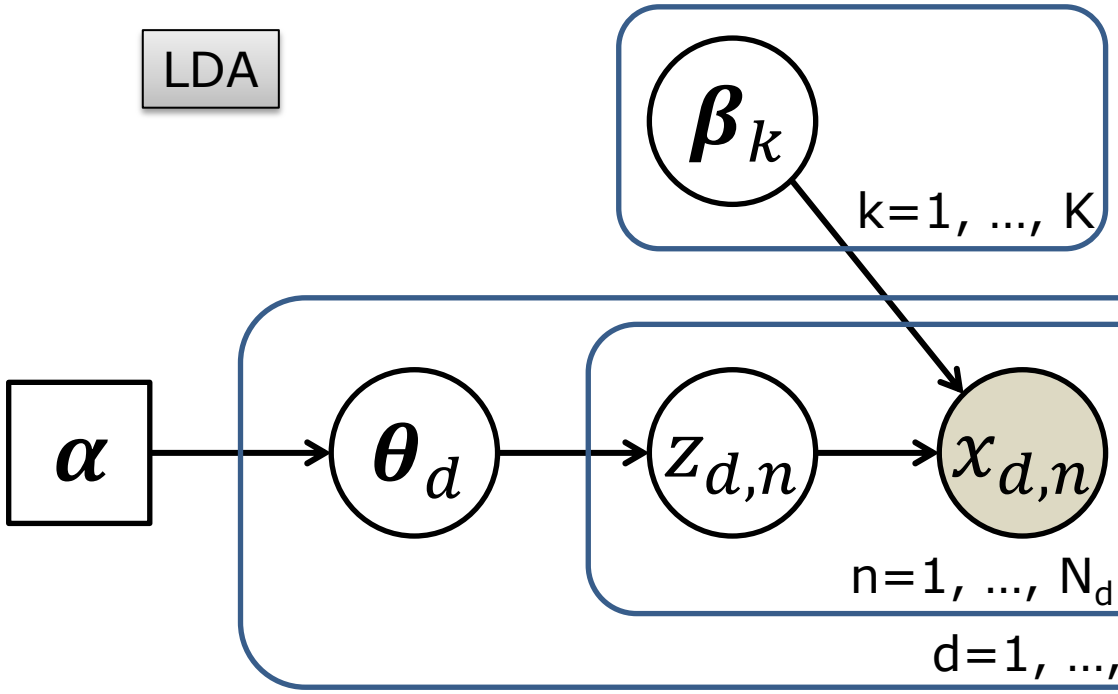
提案法: Supervised LDA

- 文書データそれぞれに対して、連続値の補助情報（教師情報）が付与されているモデル

- 補助情報の影響も考慮して、トピックを学習すると同時に補助情報の分布を学習

- ☺ 教師無し学習のモデルであるLDAを、教師情報有りデータへも応用する道を開いた重要な論文

LDA



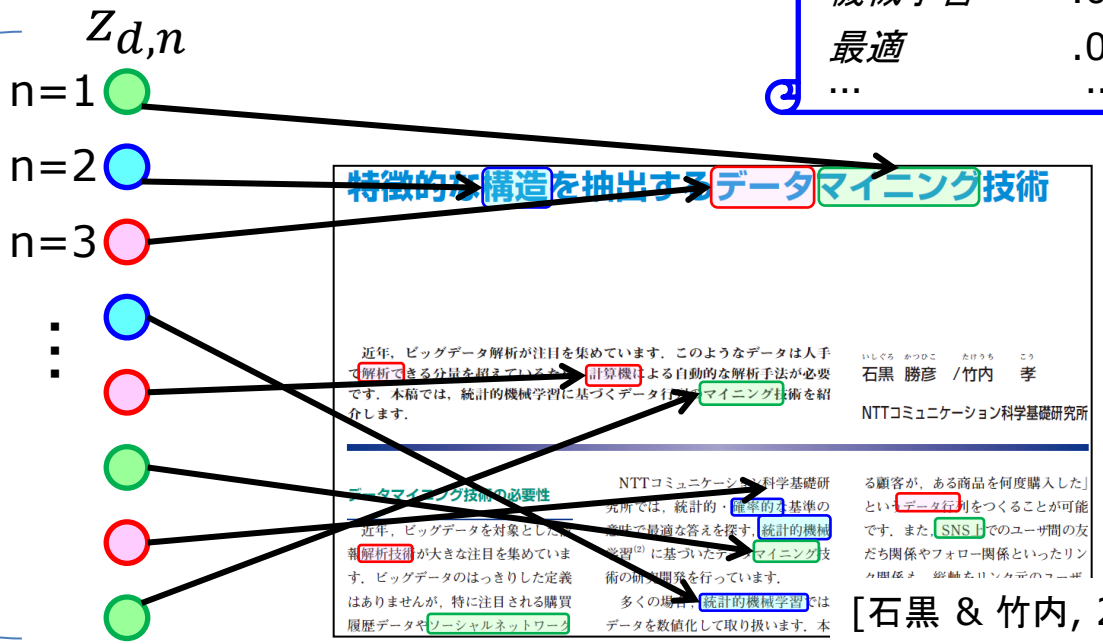
データ	.05
解析	.04
計算機	.03
…	…

$\boldsymbol{\beta}_k$

リンク	.04
ソーシャル	.02
マイニング	.01
…	…

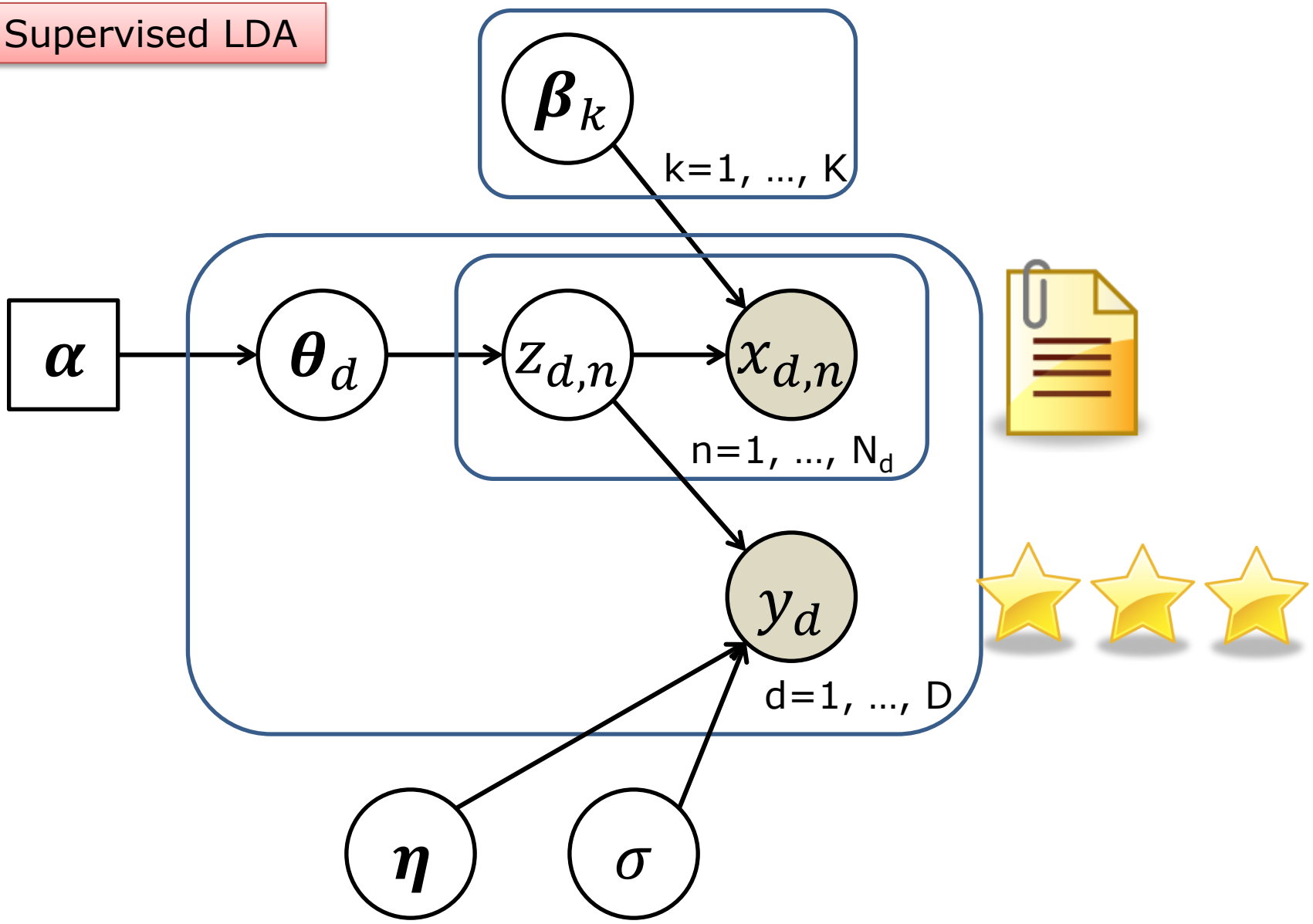
構造	.04
機械学習	.03
最適	.01
…	…

$\boldsymbol{\theta}_d$



[石黒 & 竹内, 2012]

Supervised LDA



生成モデル

for 法案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

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

response variable $y_d|\bar{\boldsymbol{z}}_d, \boldsymbol{\eta}, \sigma \sim \mathrm{N}(\boldsymbol{\eta}^T \bar{\boldsymbol{z}}_d, \sigma^2)$

for トピック k = 1, 2, …, K

topic-word proportion $\boldsymbol{\beta}_k$

提案法のポイント

- 観測量として、文書データ X と補助情報 Y があり、対等な関係になっています
- 推論時にはXとYの分布を同時に満足するようにZを学習するため、LDAからの２段構えよりも良いモデルが期待できます

LDA

$p(Z|X) \longrightarrow Y$

supervised LDA

$p(Z|X,Y)$

経験トピックに基づく補助情報の回帰

- 文書ごとの経験トピック分布（平均値）でパラメータを生成します
  - topic proportion θ_d を使わない理由は不明 ☹
- 正規分布を使っているので、補助情報は連続値と仮定しています

$$\bar{\mathbf{z}}_d = \frac{1}{N_d}\sum_{n=1}^{N_d} \mathbf{z}_{d,n}$$

$z_{d,n}$をK次元ベクトルとして見ています

$$y_d|\bar{\mathbf{z}}_d, \boldsymbol{\eta}, \sigma \sim \mathrm{N}\left(\boldsymbol{\eta}^T \bar{\mathbf{z}}_d, \sigma^2\right)$$

離散値への対応

- なお、正規分布を一般化線形モデルに変更することで離散値の補助情報にも対応できる･･･そうです

- 離散値への対応は別の手法(例えば[Lacoste-Julien, 2009] など)も参考になさった方が良いと思います

隠れ変数・パラメータの推定

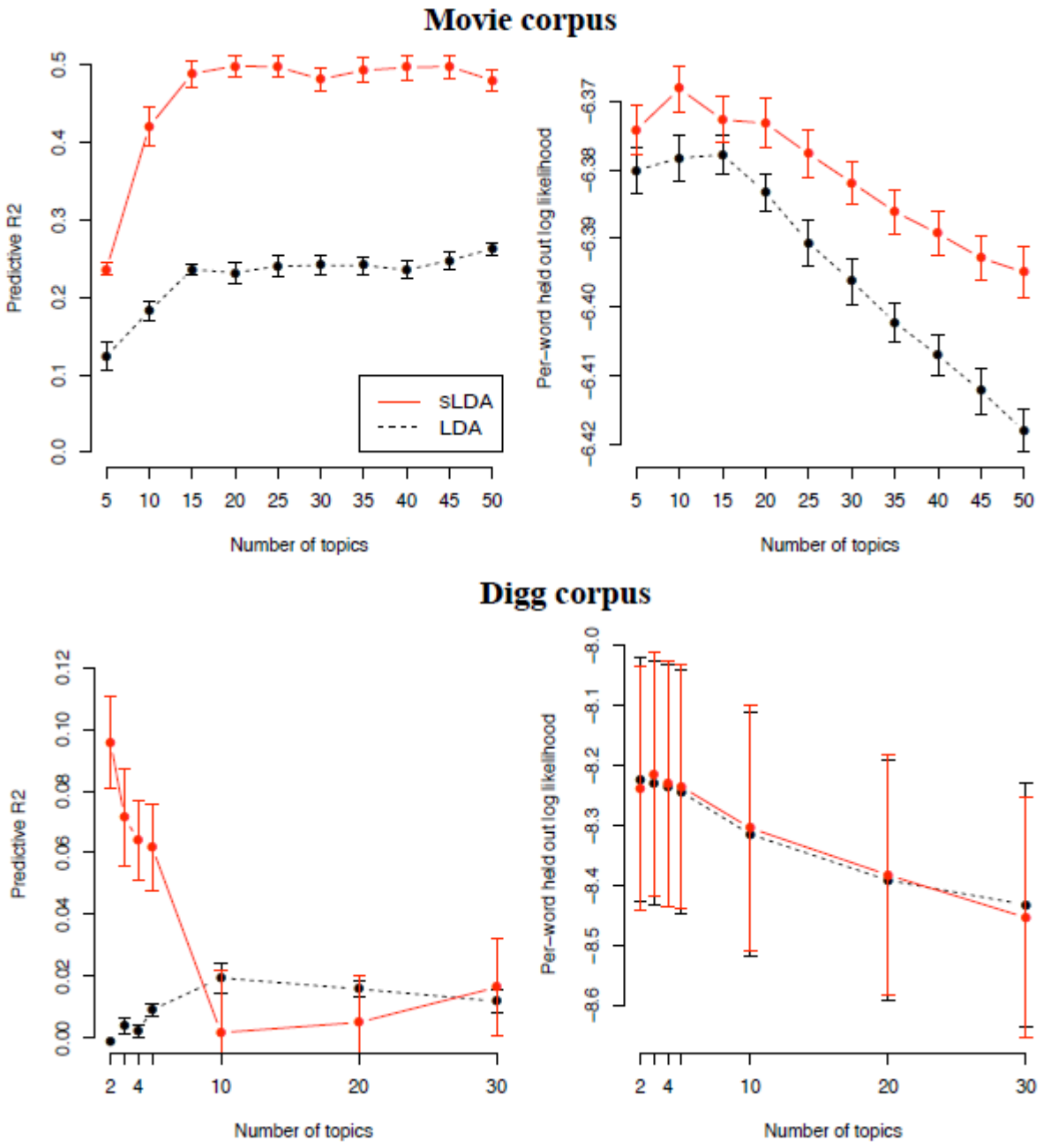
- 任意の手法で推定してかまいませんが、変分ベイズ法をお勧め

$$\log p(\boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2) \geq H(q)$$

$$+\sum_d E_q[\log p(\boldsymbol{\theta}_d|\boldsymbol{\alpha})] + \sum_{d,n} \mathrm{E}_q\left[\log p\left(z_{d,n}|\boldsymbol{\theta}_d\right)\right]$$

$$+\sum_d E_q[\log p(y_d|\bar{\boldsymbol{z}}_d, \boldsymbol{\eta}, \sigma^2)] + \sum_{d,n} \mathrm{E}_q\left[\log p\left(x_{d,n}|z_{d.n}, \boldsymbol{\beta}\right)\right]$$

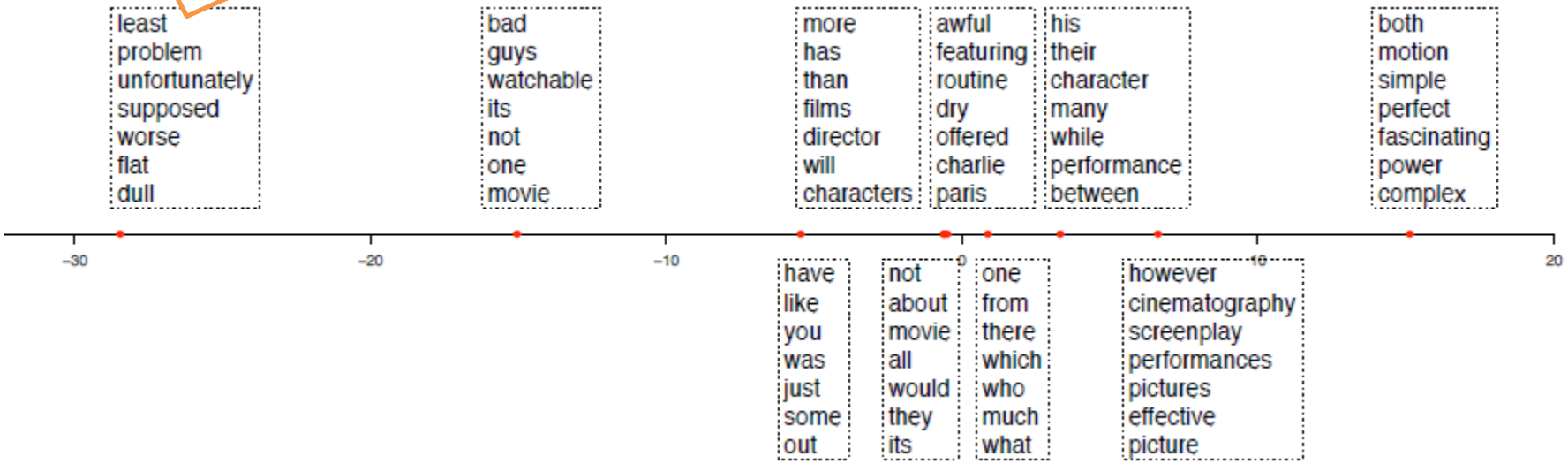
ポイントの部分で述べたように、XとYの両方が効きます



[Blei & McAuliffe, 2008]

η_kの期待値が負になるトピックの頻出単語

η_kの期待値が大きなトピックの頻出単語



$$y_d|\bar{\mathbf{z}}_d, \boldsymbol{\eta}, \sigma \sim \mathrm{N}\left(\boldsymbol{\eta}^T \bar{\mathbf{z}}_d, \sigma^2\right)$$

[Blei & McAuliffe, 2008]

まとめ: Supervised LDA

- 教師情報（観測可能な補助情報）有りのトピックモデルの先駆けです
- 非常に多くのモデルの基礎となっている、重要な拡張LDAモデルです

Ideal Point Topic Model
[Gerrish & Blei, 2011]

Gerrish and Blei,
"Predicting Legislative Roll Calls from Text",
in Proc. ICML, 2011.

数理モデルによる政治解析

- Quantitative political science: 政治系の話題においても数理手法を用いた解析・研究がおこなわれています

先日のアメリカ大統領選挙

"New York Timesの選挙予測専門家、ネイト・シルバーは昨夜、大統領選の勝敗を全50州で的中させた。その一方で、いわゆる政治専門家たちの予想はほとんどが外れた。…(中略)…シルバーは今回も彼の作った数理的予測モデルが古臭い専門家の勘や生半可な統計に基づく推測より圧倒的に優れていたことを証明した。"

G. Ferenstein, TechCrunch Japan, Nov 8 2012.
(http://jp.techcrunch.com/archives/20121107pundit-forecasts-all-wrong-silver-perfectly-right-is-punditry-dead/)

議員の投票行動モデル: ideal point model [Clinton, 2004]

- 議員の政治的信条と法案をそれぞれ潜在空間に射影、お互いの位置関係で賛成確率をモデル化

- ☹ 法案の中身（文章データ）を利用してない
- ☹ 新規法案に対する投票予測が不可能

提案法: Ideal point topic model

- 投票行動のみを扱っていたIdeal point modelを拡張

- 各法案の内容をトピックモデルでモデル化

- ☺ 文書のトピック・言葉づかいと投票行動の関係を調査できる
- ☺ 新規法案でもトピック(文書内容)から各議員の投票行動を予測できる

対象データ

- 1997-2011年のU.S. Congress Roll call data
  - 各法案(bill, resolution)はn-gramを”単語”とするBow表現
  - 各議員(legislator)は0/1のvotesを持つ

法案 dのBoW表現 $x_{d,n}$　**法案 dに対する議員 uの投票結果 $v_{d,u}$**

法案　d

u = 1: yes(1)　u = 2: no(0)　u = 3: no(0)

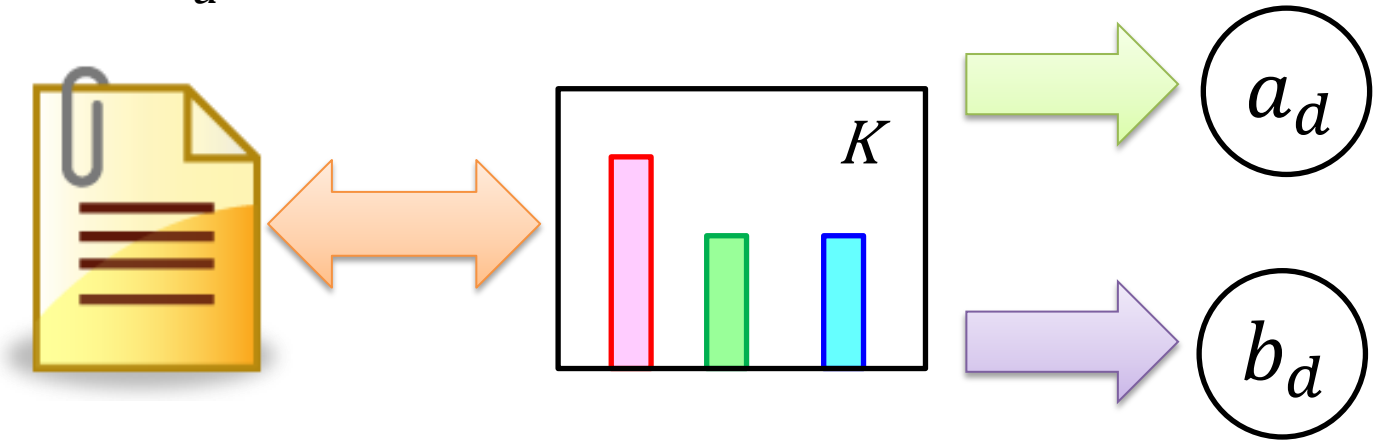
Ideal topic model: 法案の投票傾向を２つのパラメータで表現

- a_d bill difficulty: 誰でも賛成(反対)するような法案だと正の(負の)大きい値をとる
- b_d bill discriminaty: 法案の"立ち位置"
- y_u: 議員の理想とする"立ち位置"

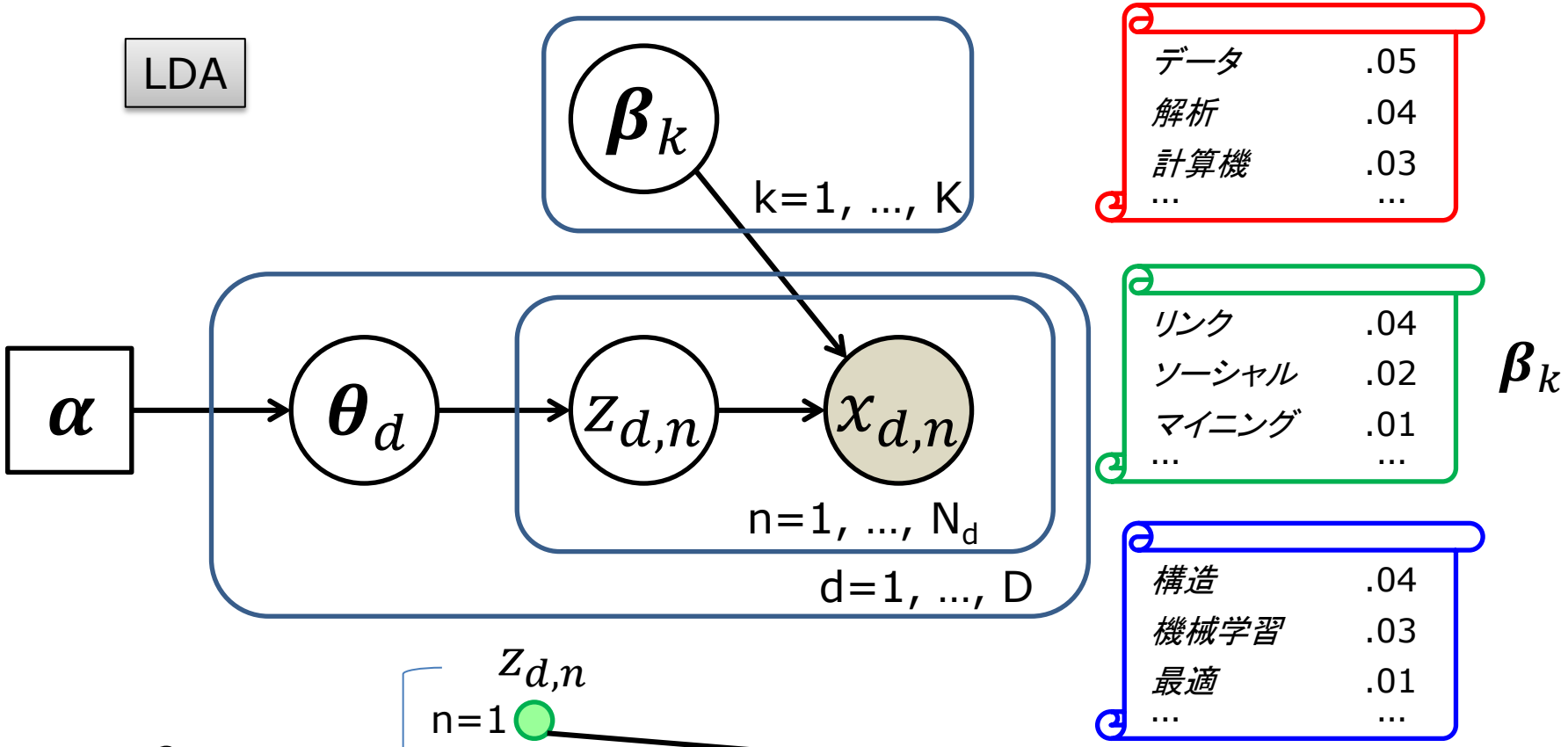
$$p(v_{du} = 1) = \sigma(a_d + b_d y_u) \qquad \sigma(x) = \frac{\exp(x)}{1 + \exp(x)}$$

提案法のアイデア: 文書トピックによるideal point modelの制御

- 法案のパラメータは、内容＝文書のトピックによって制御されると考える
  - 例1) WBC優勝を褒め称える議案は皆賛成するのでa_dが大きな正の値
  - 例2) 予算関係は起案者の意見が色濃くでるのでb_dが特徴的な値をとる



LDA

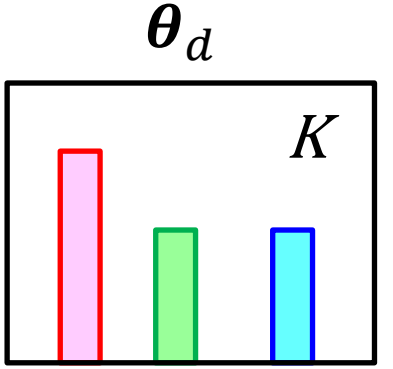


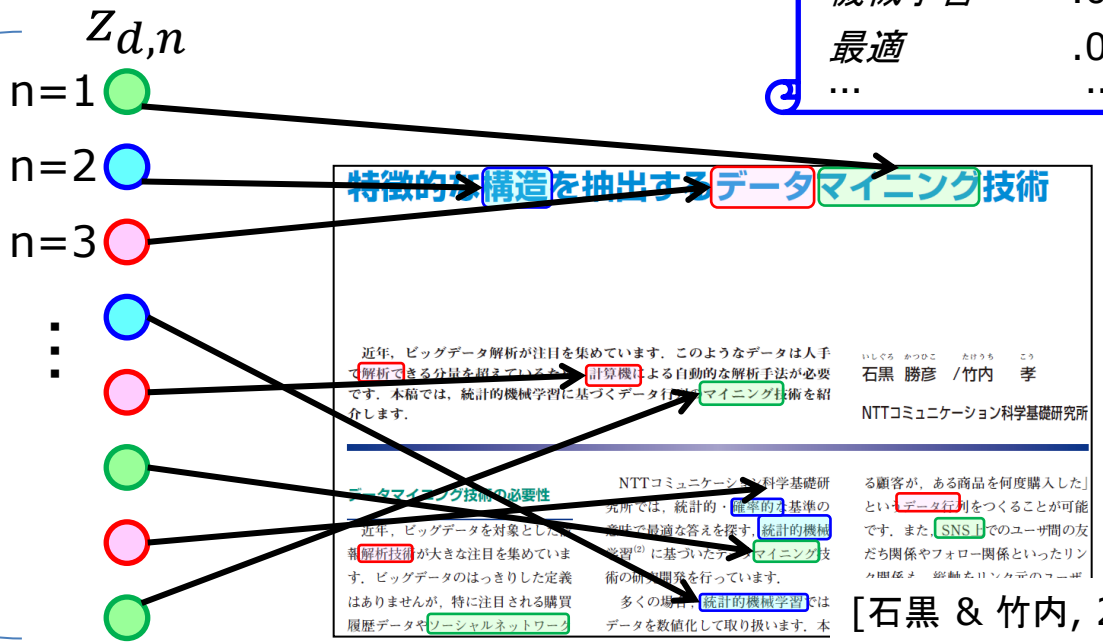
データ	.05
解析	.04
計算機	.03
…	…

リンク	.04
ソーシャル	.02
マイニング	.01
…	…

$\boldsymbol{\beta}_k$

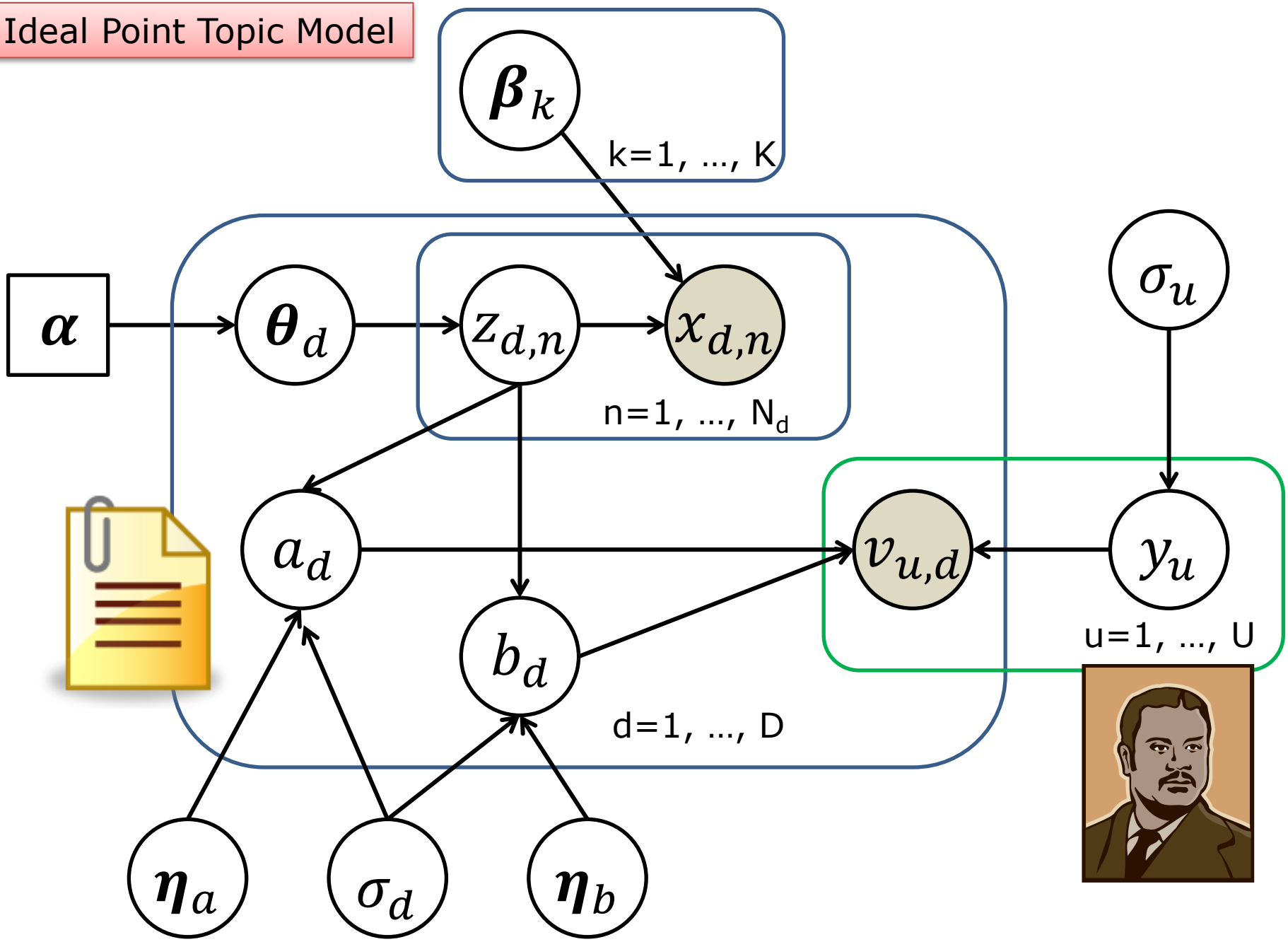
構造	.04
機械学習	.03
最適	.01
…	…





[石黒 & 竹内, 2012]

Ideal Point Topic Model



生成モデル

for 法案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 $\boldsymbol{\beta}_k$

生成モデル

for 議員 u = 1, 2, …, U

legislator ideal point $y_u|\sigma_u \sim \boldsymbol{N}(0, \sigma_u)$

for 法案 d = 1, 2, …, D_t

bill difficulty param. $a_d|\bar{\boldsymbol{z}}_d, \boldsymbol{\eta}_a, \sigma_d \sim \mathrm{N}\left(\boldsymbol{\eta}_a{}^T \bar{\boldsymbol{z}}_d, \sigma_d{}^{\boldsymbol{2}}\right)$

bill discrimination param. $b_d|\bar{\boldsymbol{z}}_d, \boldsymbol{\eta}_b, \sigma_d \sim \mathrm{N}\left(\boldsymbol{\eta}_b{}^T \bar{\boldsymbol{z}}_d, \sigma_d{}^{\boldsymbol{2}}\right)$

for 議員 u = 1, 2, …, U

vote observation

$$p\left(v_{d,u} = 1|a_d, b_d, y_u\right) = \sigma(a_d + b_d y_u)$$

σはlogistic function $\sigma(x) = \dfrac{\exp(x)}{1 + \exp(x)}$

経験トピックに基づく Ideal pointパラメータモデリング

- 実際には θ_d ではなく、経験トピック分布でパラメータを生成します

$$\bar{\mathbf{z}}_d = \frac{1}{N_d}\sum_{n=1}^{N_d} \mathbf{z}_{d,n}$$

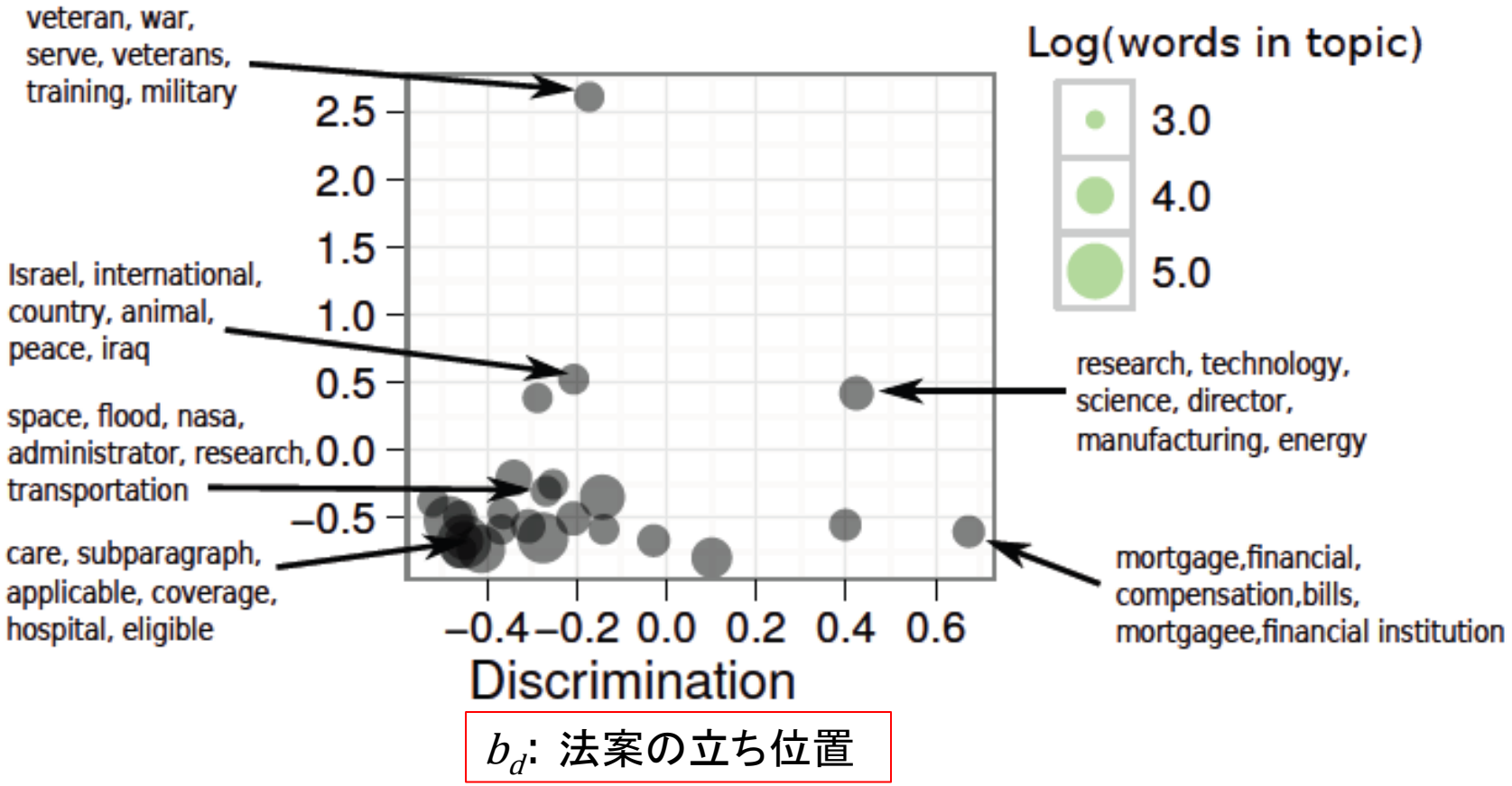
$z_{d,n}$をK次元ベクトルとして見ています

bill difficulty $a_d|\bar{\mathbf{z}}_d, \boldsymbol{\eta}_a, \sigma_d \sim \mathrm{N}\left(\boldsymbol{\eta}_a{}^T \bar{\mathbf{z}}_d, \sigma_d{}^2\right)$

bill discriminarity $b_d|\bar{\mathbf{z}}_d, \boldsymbol{\eta}_b, \sigma_d \sim \mathrm{N}\left(\boldsymbol{\eta}_b{}^T \bar{\mathbf{z}}_d, \sigma_d{}^2\right)$

隠れ変数・パラメータの推定

- 任意の手法で推定してかまいませんが、変分ベイズ法をお勧め
- 解の不定性
  - b ＊ yのせいで、b, yの値の符号は反転可能
  - そこで、共和党・民主党の重鎮だけは極端な正負の値に固定します

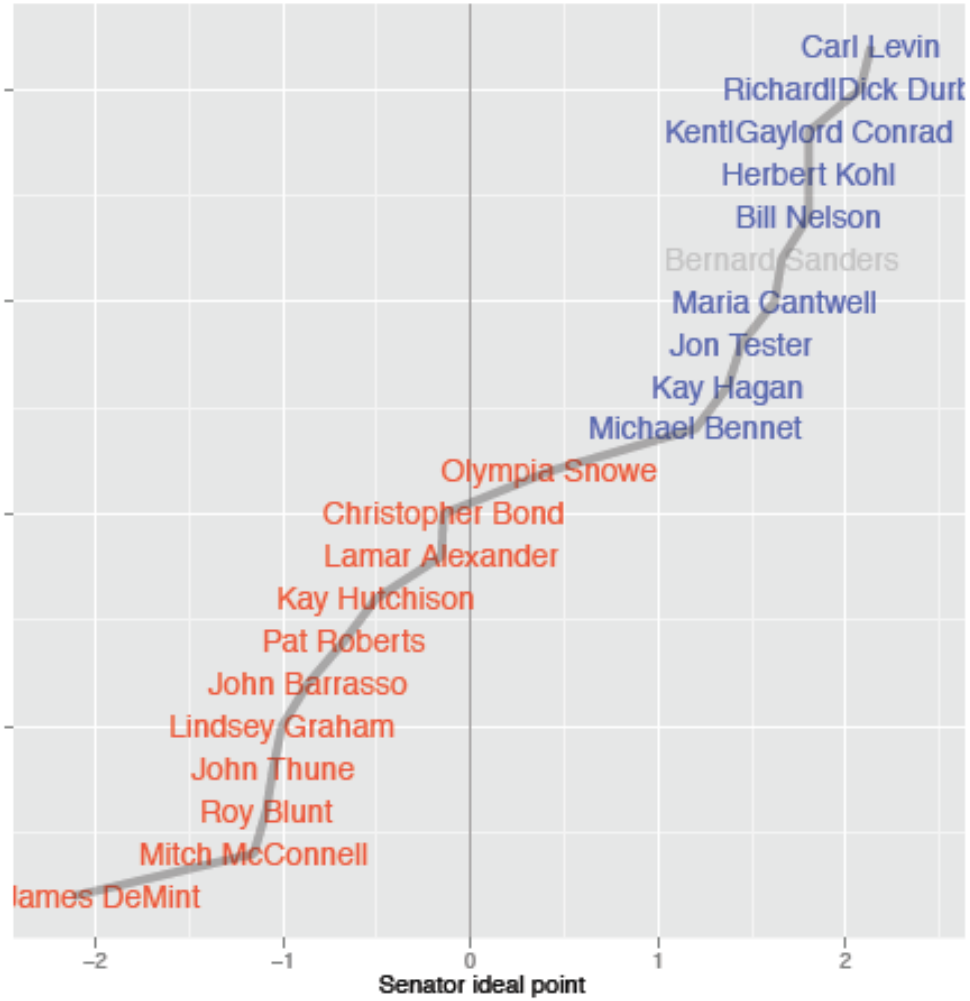


a_d: 法案のバイアス項

b_d: 法案の立ち位置

[Gerrish & Blei, 2011]

$$p(v_{du} = 1) = \sigma(a_d + b_d y_u)$$



$$p(v_{du} = 1) = \sigma(a_d + b_d y_u)$$

[Gerrish & Blei, 2011]

まとめ: Ideal Point Topic model

- 法案に対する投票行動のモデル化です
- 法案の文書内容だけでなく、各法案に対する議員の投票結果が観測できています
- トピックモデルを法案の内容を表す“部品”として、既存のideal point modelと組み合わせています

その他の教師あり・補助情報ありトピックモデル

- Flaherty et al., “A Latent Variable Model for Chemogenomic Profiling”, Bioinformatics, Vol. 21(15), pp.3286-3293, 2005.
- Lacoste-Julien et al., “DiscLDA: Discriminative Learning for Dimensionality Reductioin and Classification”, Advances in Neural Information Processing Systems 21 (Proc. NIPS), 2009.
- Ramage et al., “Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora”, in Proc. EMNLP, 2009.
- Gerrish and Blei, “How They Vote: Issue-Adjusted Models of Legislative Behavior”, in Proc. NIPS, 2012.

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.
- [Blei & McAuliffe, 2008] Blei and McAuliffe, “Supervised Topic Models”, Advances in Neural Information Processing Systems 20 (Proc. NIPS), 2008.
- [Lacoste-Julien, 2009] Lacoste-Julien et al., “DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification”, Advances in Neural Information Processing Systems 21 (Proc. NIPS), 2009.
- [Gerrish & Blei, 2011] Gerrish and Blei, “Predicting Legislative Roll Calls from Text”, in Proc. ICML, 2011.
- [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.