NTT

トピックモデルの応用：相関・構造をもつトピックモデル

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”

- 機械学習の研究分野では、日々新しい、より柔軟で表現力の高い(≒複雑な☹)トピックモデルが提案されています

- このスライドでは、それらのうち、特に構造化に関する仕事を厳選してご紹介します

トピックモデルの大きな特長はモデルの単純さです

- 誤解を恐れずにいえば、単純な混合ガウシアンモデル(GMM)が理解できれば、LDAは理解できます

- GMMがその単純さゆえに非常に幅広いドメインの連続データで有効なように、LDAも幅広いドメインの離散データで有効です

トピックモデルの問題点も モデルの単純さです

- モデルが単純ということは、大胆な仮定を置いてデータを表現していることになります
- 実際のデータと明らかに合わない仮定の場合、これを正す必要があります

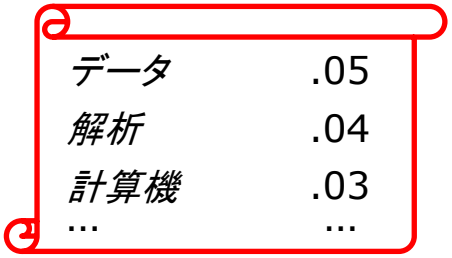
- 沢山の複雑化したトピックモデルが提案されています

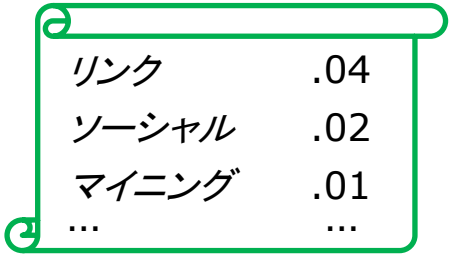
Correlated Topic Models
[Blei & Lafferty, 2007]

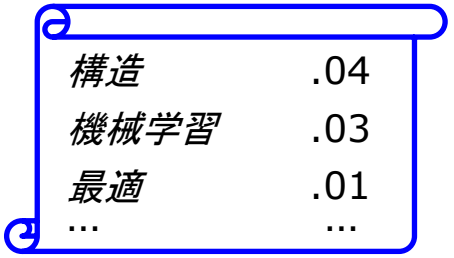
Blei and Lafferty,
"A Correlated Topic Model of Science",
The Annals of Applied Statistics,
Vol. 1(1), pp. 17-35, 2007.

トピックモデルの大前提の仮定: トピックは独立

- 簡単にいうと:「各トピック k の間には相関がない」
- 通常のGMMでも共有される考え方です
- これのおかげで各種モデル推論が簡単になっています







トピックは本当に独立なのか？

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

[Blei, 2003]

TREC APコーパス: AP通信ニュースの文書データセット

トピックは本当に独立なのか？

- 先の例だけで分かるように、これは成り立たないことが多々ありそうです

- すなわち、「本当は相関のあるトピック」を無理やり「相関のないトピック」に分割している可能性が高いです

提案法: Correlated Topic Models (CTM)

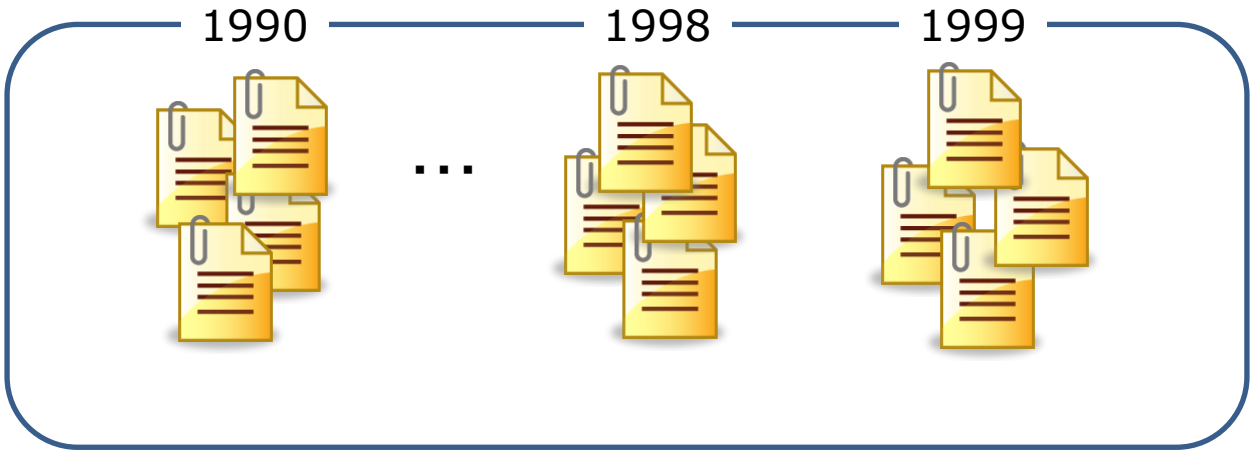
- ☺ 以後のトピックモデル研究に非常に大きな影響を与えたモデルです

- 科学誌ScienceのOCRデータを用いて、科学論文のトピック解析を行います
- トピック間の相関（正・負）をexplicitにモデル化します

- 推論は少々面倒になります

対象データ: Science誌

- 1880年にエジソンによって刊行された、非常に著名な科学論文誌
- OCRされた論文誌データ(JSTOR)を利用
- 実験では1990年代の論文を対象とします



近代の科学は分野横断的です

- 先進のbioinformaticsは高度な統計学の知識とデータマイニングの手法が必要です
- 分子動力学法は物理学に則っていますが、化学・生物学の多様な系に応用されます

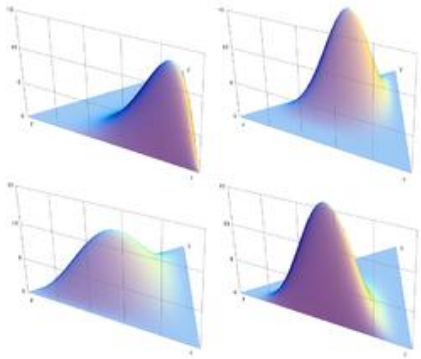
- Science誌は専門分野の論文誌ではないため、このような分野間の相関構造が強く表れるはずです

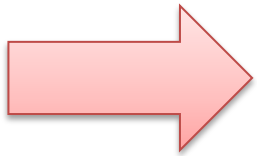
提案法のアイデア: とにかく簡単に相関を埋め込みます

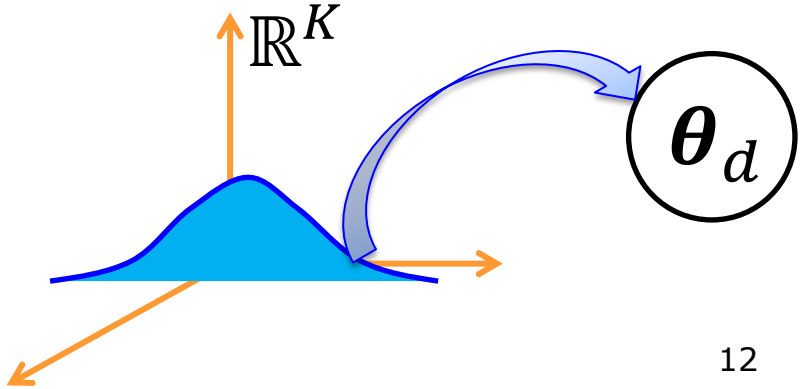
- 目的: 文書 d のtopic proportion θ_d を生成する際に、トピックの相関を埋め込む
- 解: 一番簡単な相関を持つ分布といえば多次元正規分布なので、素直にそれを使う

Dirichlet**分布: 足して1にするだけ**

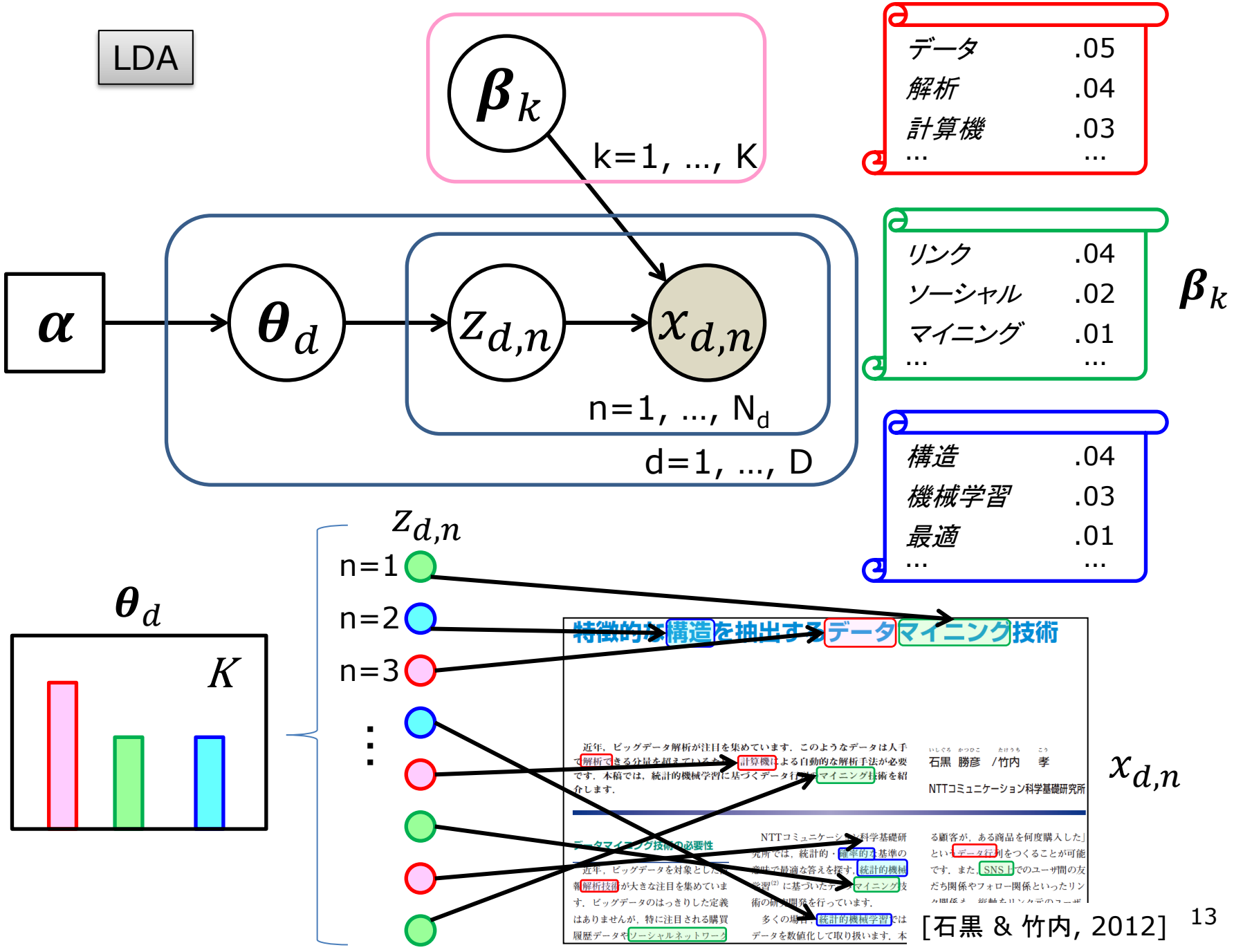
多次元正規分布: "一緒に値が動く" "片方が増えるともう片方が減る" などを表現できる





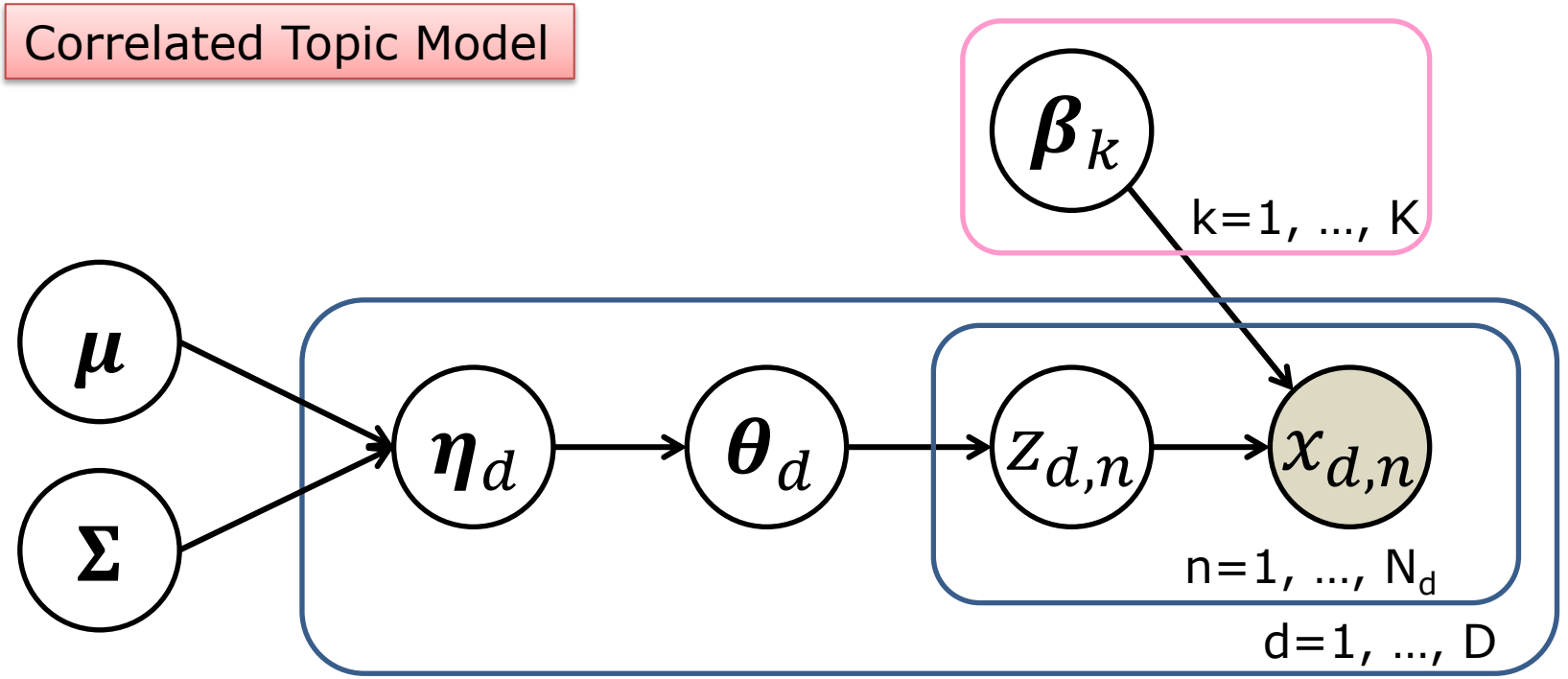


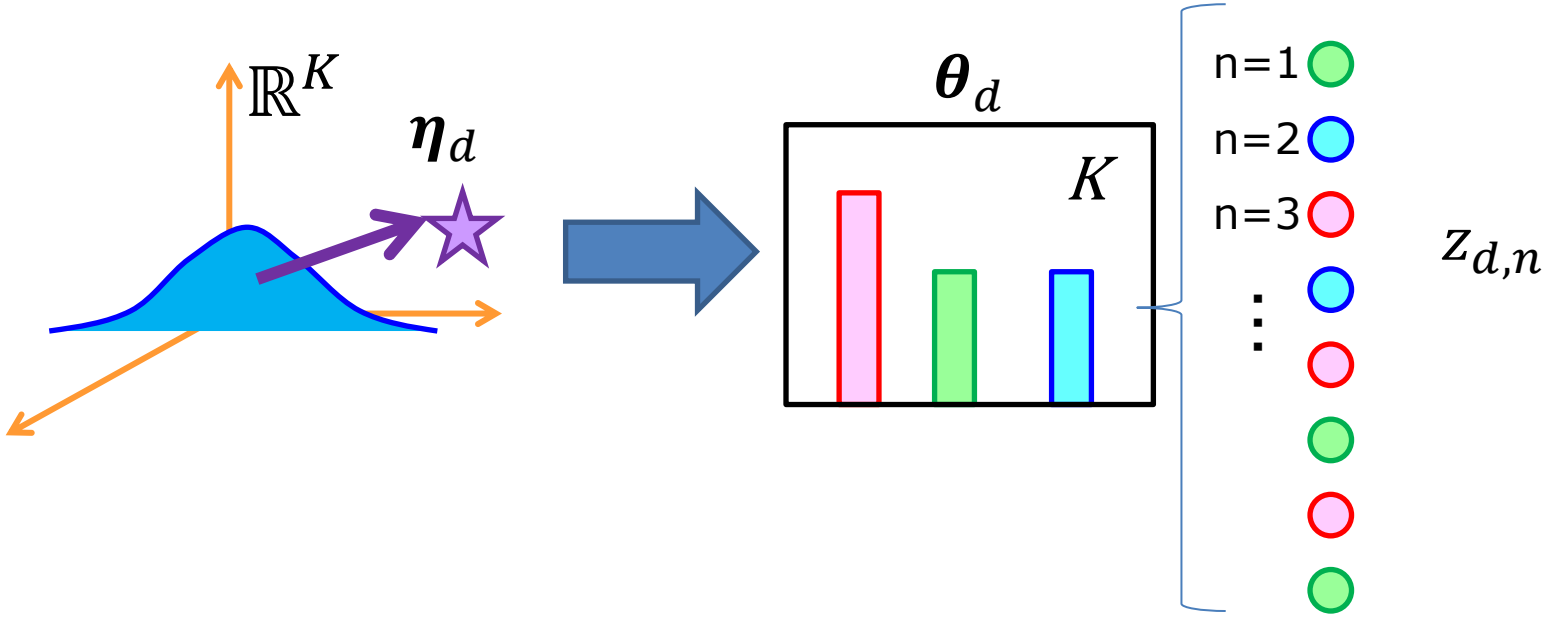
LDA



[石黒 & 竹内, 2012]

Correlated Topic Model





生成モデル

for 文書d = 1, 2, …, D_t

topic proportion

$$\boldsymbol{\eta}_d|\boldsymbol{\mu},\boldsymbol{\Sigma}\sim\mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$$

$$\boldsymbol{\theta}_d|\boldsymbol{\eta}_d=\pi(\boldsymbol{\eta}_d)$$

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

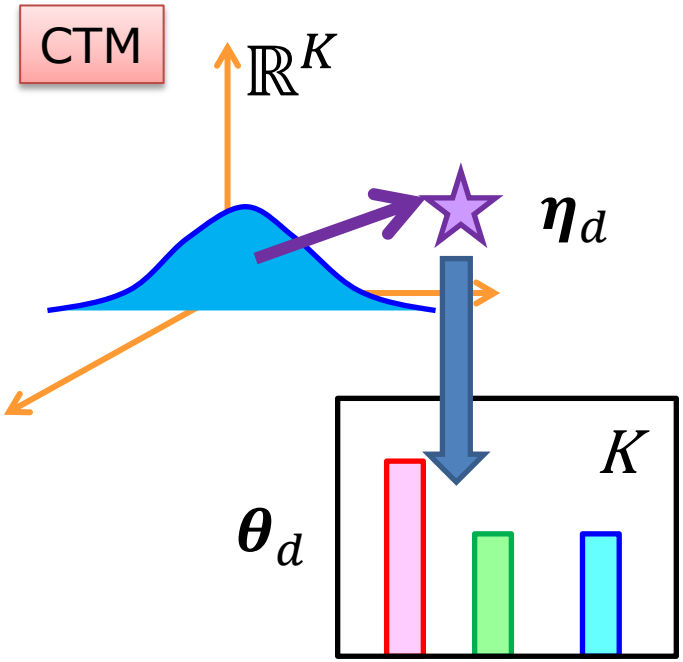
topic-word proportion $\boldsymbol{\beta}_k$

πはsoft-max関数

$$\pi(\boldsymbol{v})=\frac{\exp(v_k)}{\sum_l\exp(v_l)}$$

トピック間の相関: 多次元正規分布

- 共分散行列 Σ の効果でトピック分布に相関が生まれます
- 正規分布から生成される量はそのままは使えないので、Soft-maxで足して1にします



$$\eta_d|\mu,\Sigma \sim \mathrm{N}(\mu,\Sigma)$$

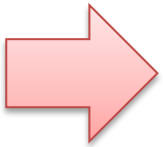
$$\theta_d|\eta_d = \pi(\eta_d)$$

πはsoft-max関数

$$\pi(v) = \frac{\exp(v_k)}{\sum_l \exp(v_l)}$$

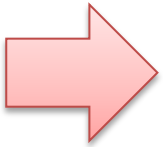
隠れ変数・パラメータの推定：難しくなります

- 原因1: Soft-max関数のため、共役性(conjugate)を利用できません ☹



(collapsed) Gibbs sampling**が非効率になるため、変分ベイズ法が候補になります**

- 原因2: 変分下限の評価も難しくなります



２段階での近似が必要になります

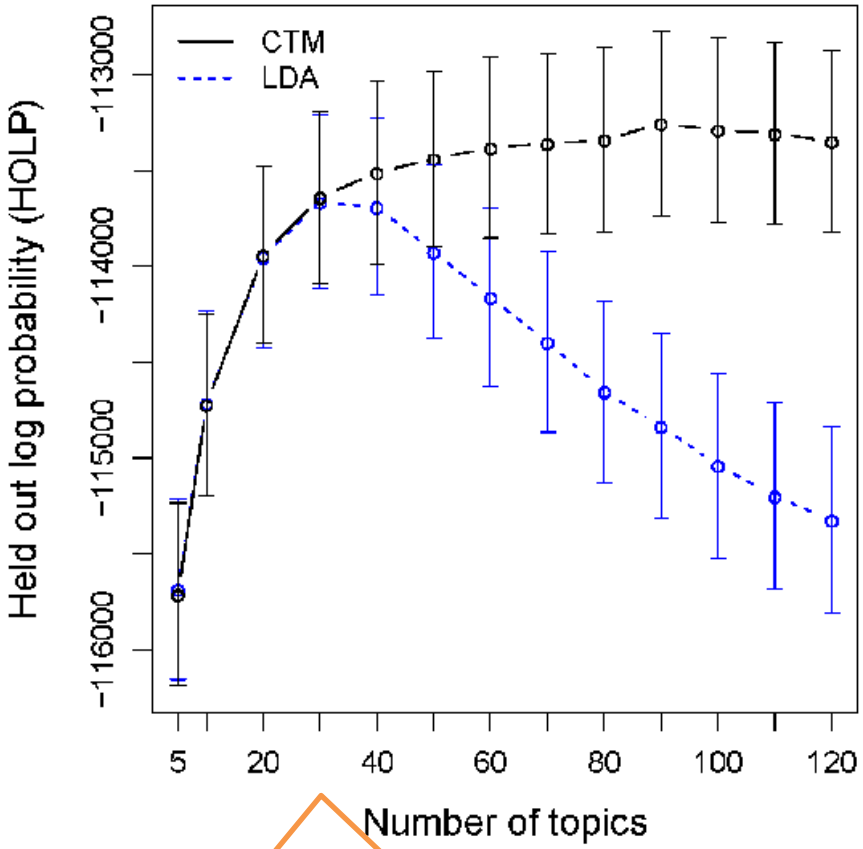
変分下限の評価

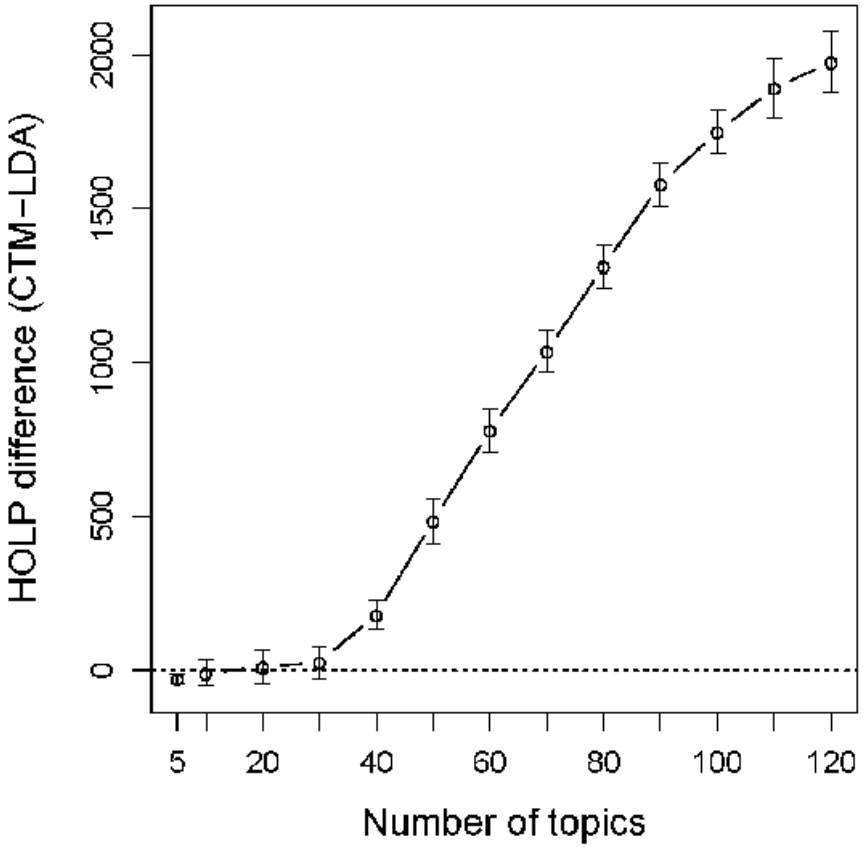
- 通常通り、Jensenによる下限を与えたあとに、さらに近似が必要です
- 詳しくは[Blei & Lafferty, 07]のAppendixをご覧ください

$$\log p(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta}) \geq \mathrm{E}_q[\log p(\boldsymbol{\eta}|\boldsymbol{\mu},\boldsymbol{\Sigma})] + \sum_{d,n} \mathrm{E}_q\left[\log p\left(z_{d,n}|\boldsymbol{\eta}\right)\right]$$

$$+ \sum_{d,n} \mathrm{E}_q\left[\log p\left(x_{d,n}|z_{d,n},\boldsymbol{\beta}\right)\right] + H(q)$$

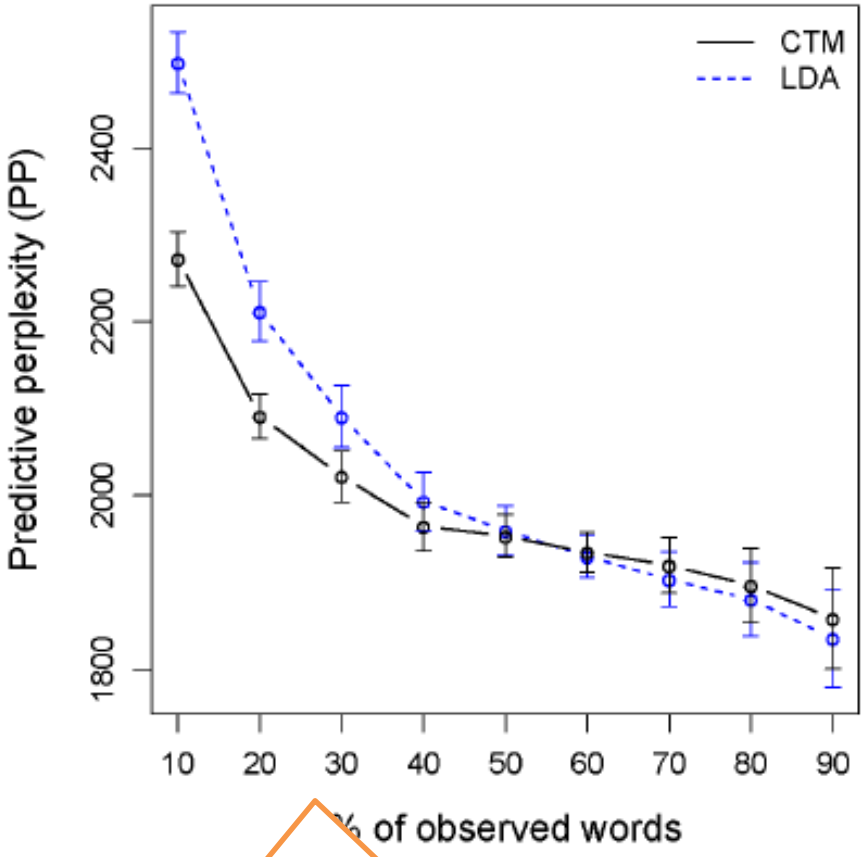
この評価でさらに近似が必要

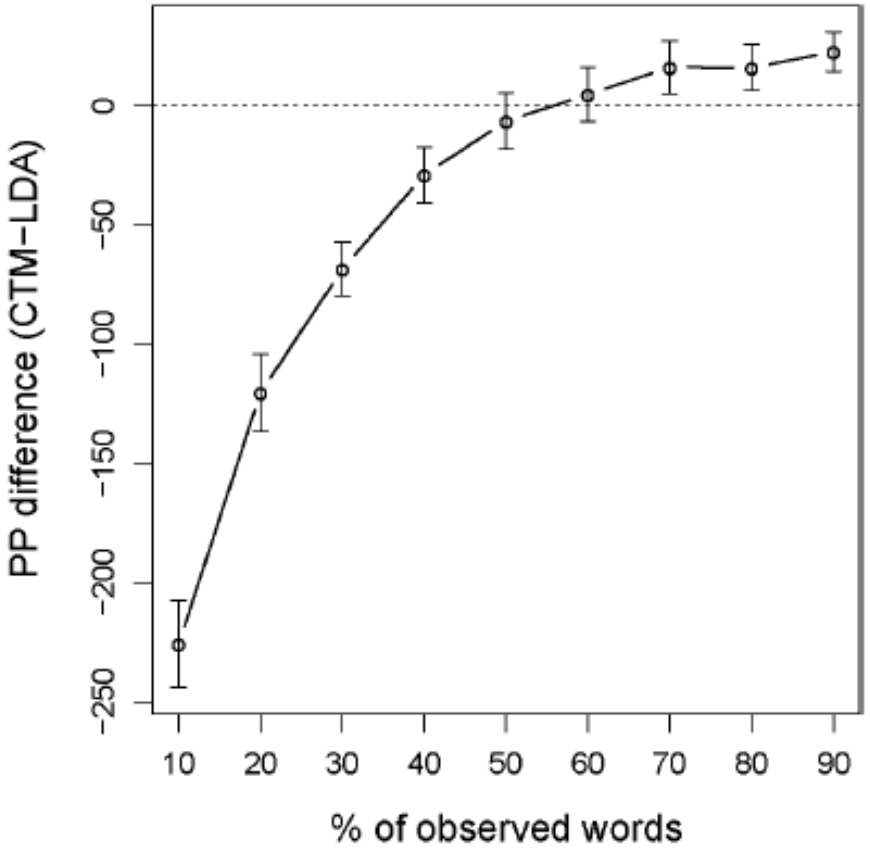




LDAは K＝30でピーク:
無理に独立なトピックを
仮定することの弊害が出てくる

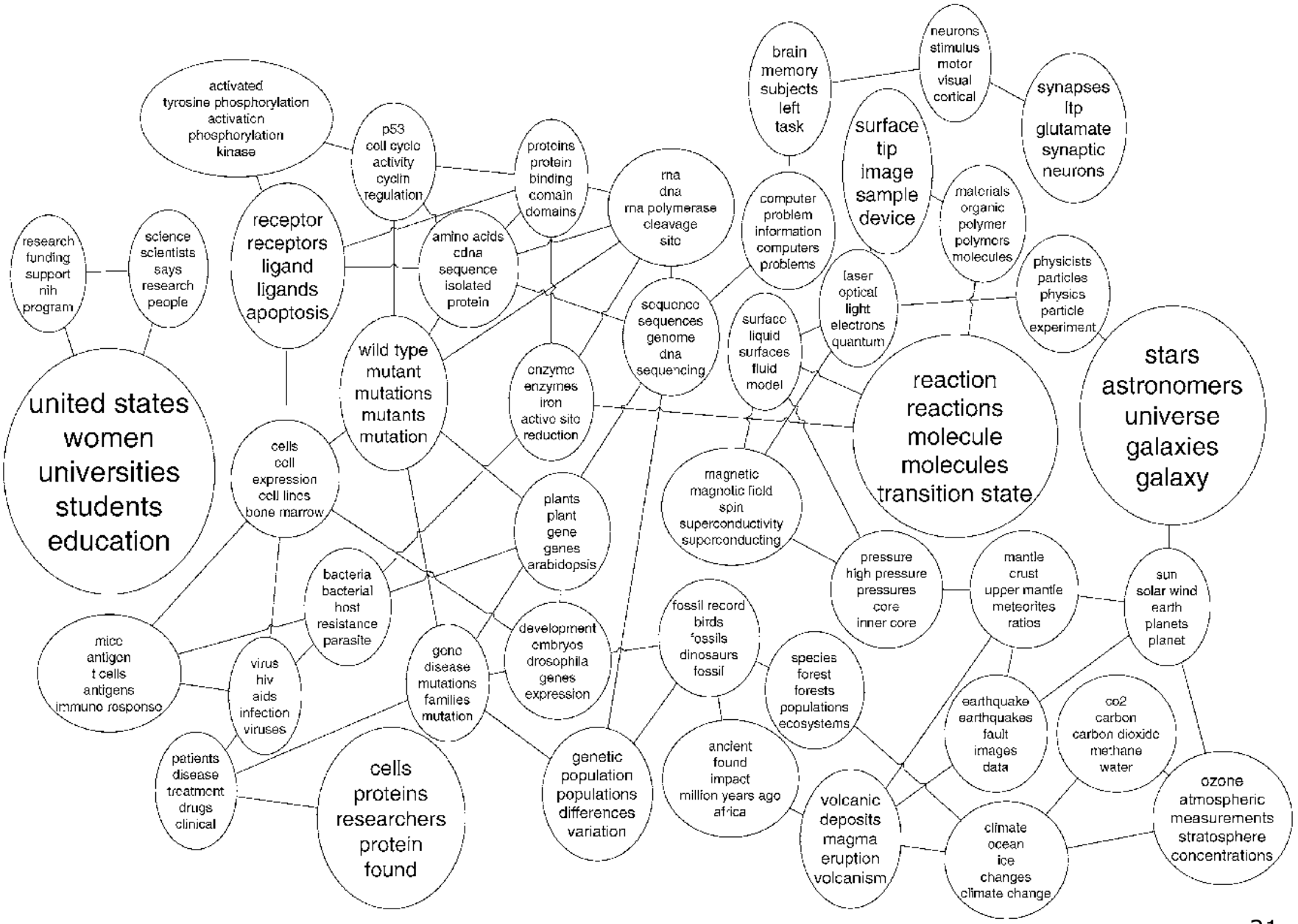
[Blei & Lafferty, 2007]





[Blei & Lafferty, 2007]

単語の観測量が少ないときにLDA よりも良い予測精度を記録



[Blei & Lafferty, 2007]

まとめ: Correlated Topic Models

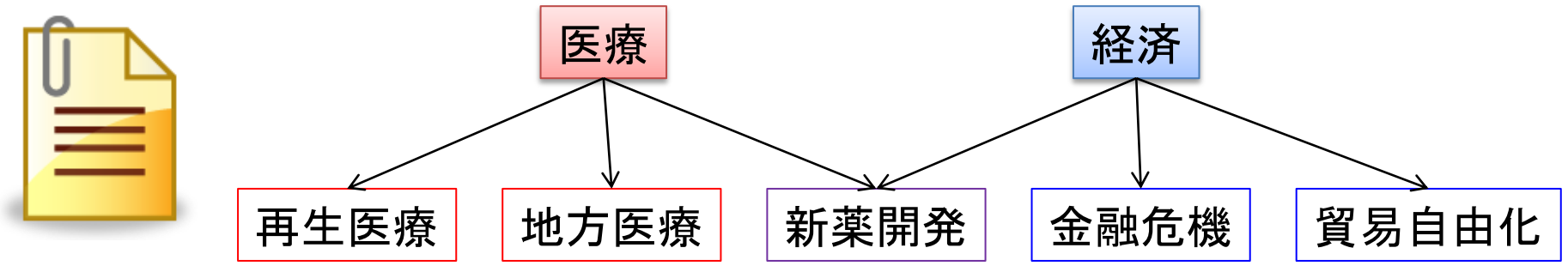
- トピック分布に、トピック間の相関を導入したモデルです
- 多次元正規分布でトピック間の関係を表現します
- 非常に有名で、後の各種トピックモデルに大きな影響を与えた仕事です。必須です。

PAM:
Pachinko Allocation Model
[Li & McCallum, 2006]

Li and McCallum,
"Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations",
in Proc. ICML, 2006.

CTMはトピックの間の階層構造が表せない

- 各トピックが同じレベルにあるからです
- トピックの階層構造（上下関係・包含関係）が適したデータの存在は容易に想像されます

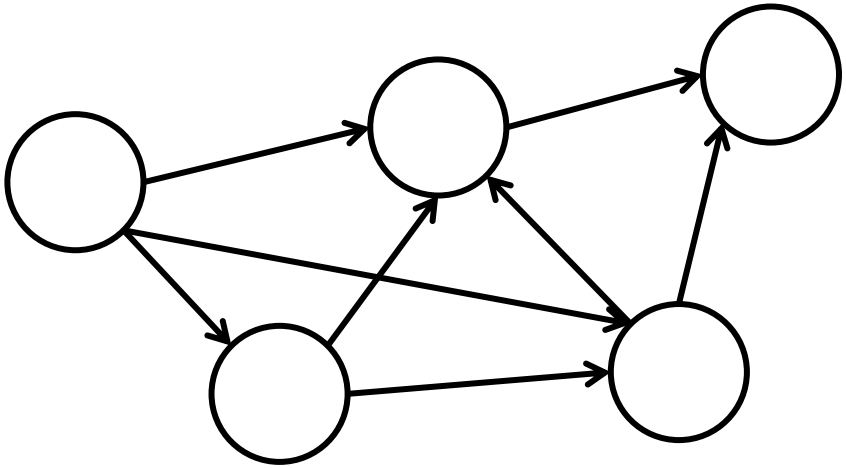


提案法: Pachinko Allocation Model (PAM)

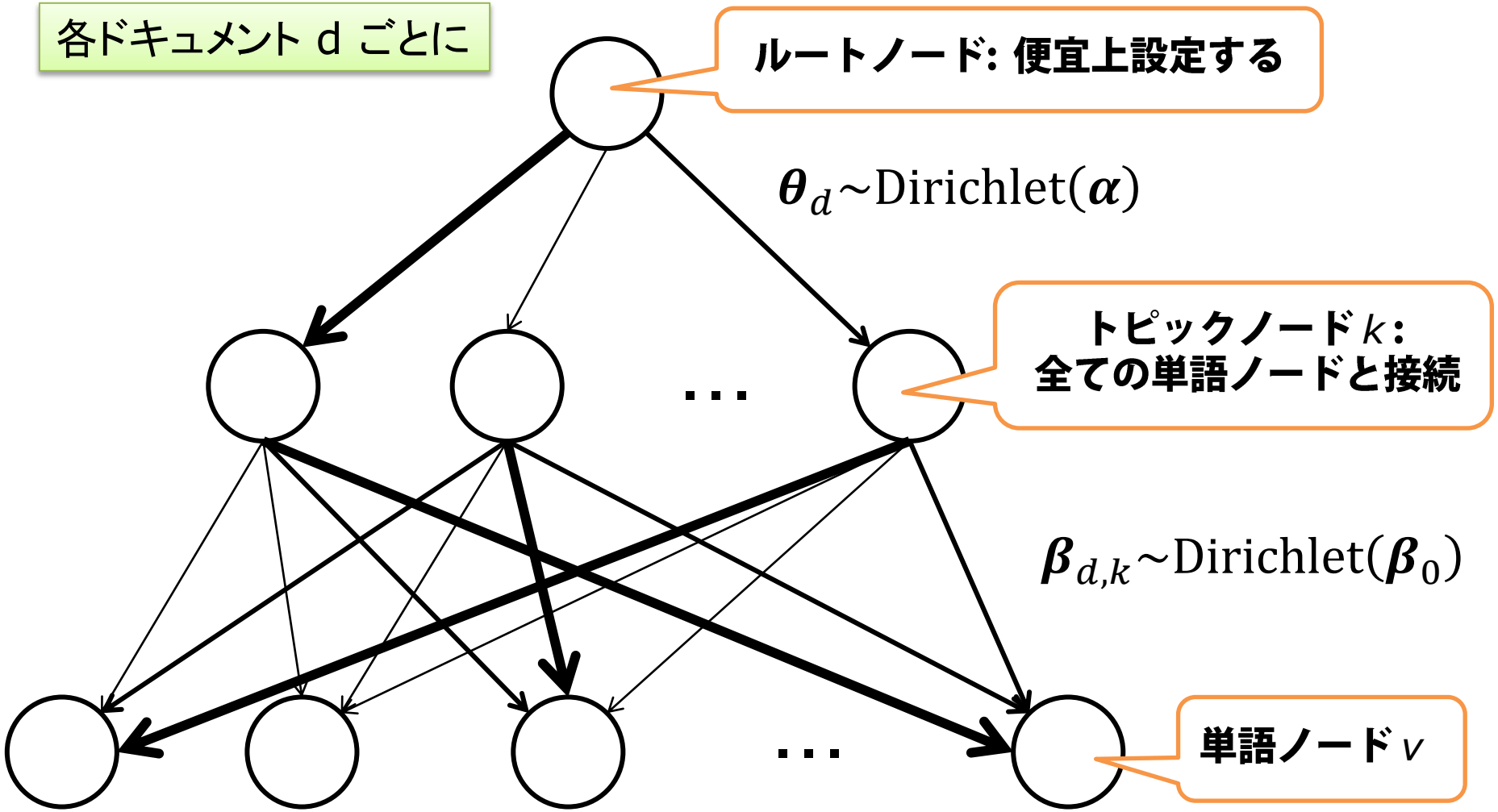
- トピック間の関係・相関を一般的に表現するモデルです
- トピック間の階層構造を基本として、パチンコ玉が落ちるように単語を生成します
- 複数トピックの共起なども表現できます

前提: 有向非巡回グラフ(DAG; Directed Acyclic Graph)

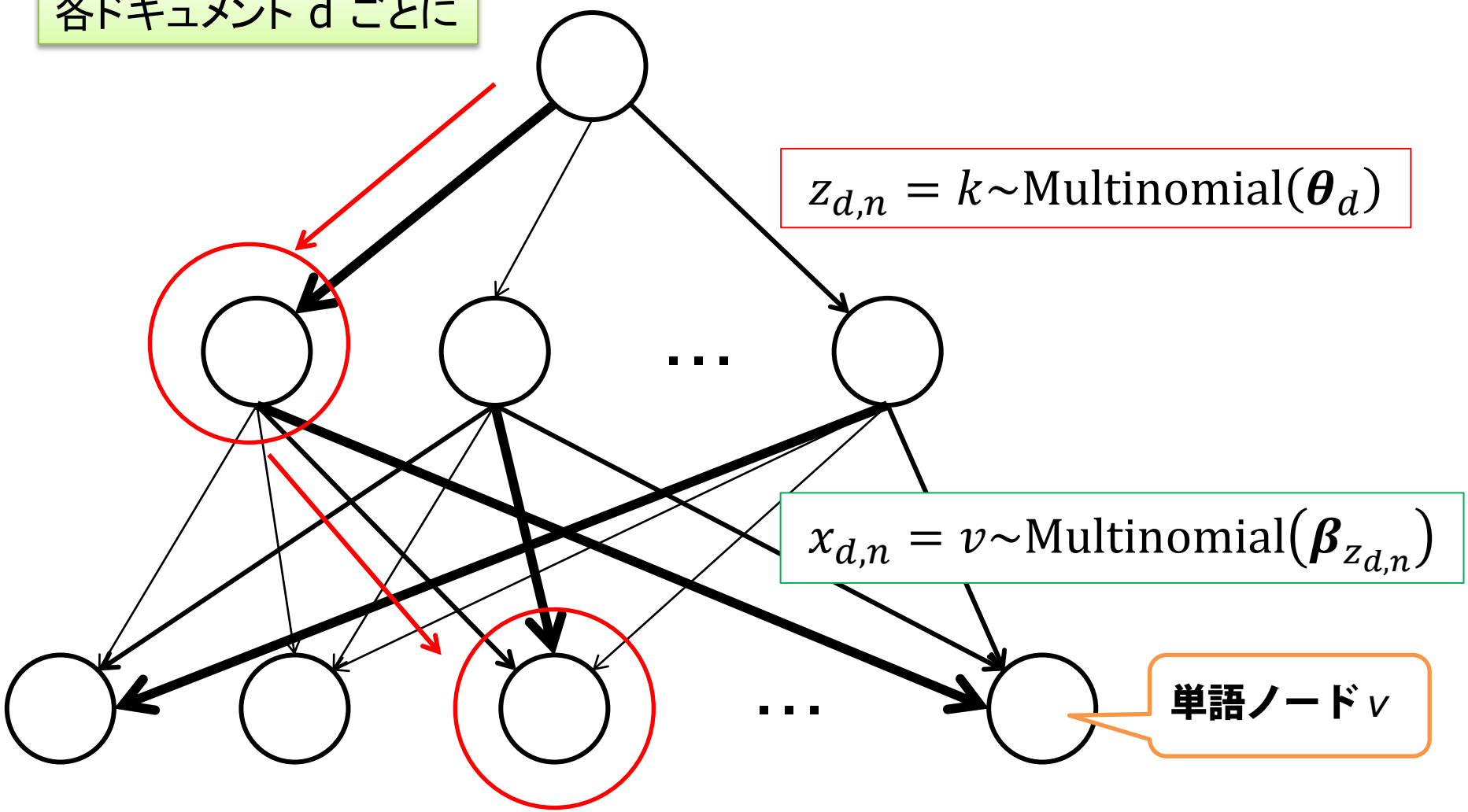
- 有向:ノード間のリンクは方向があります
- 非巡回:リンクをたどって、元のノードに戻ってくることがありません
- 木構造はDAGのさらに特殊な例です



DAGによるLDA解釈

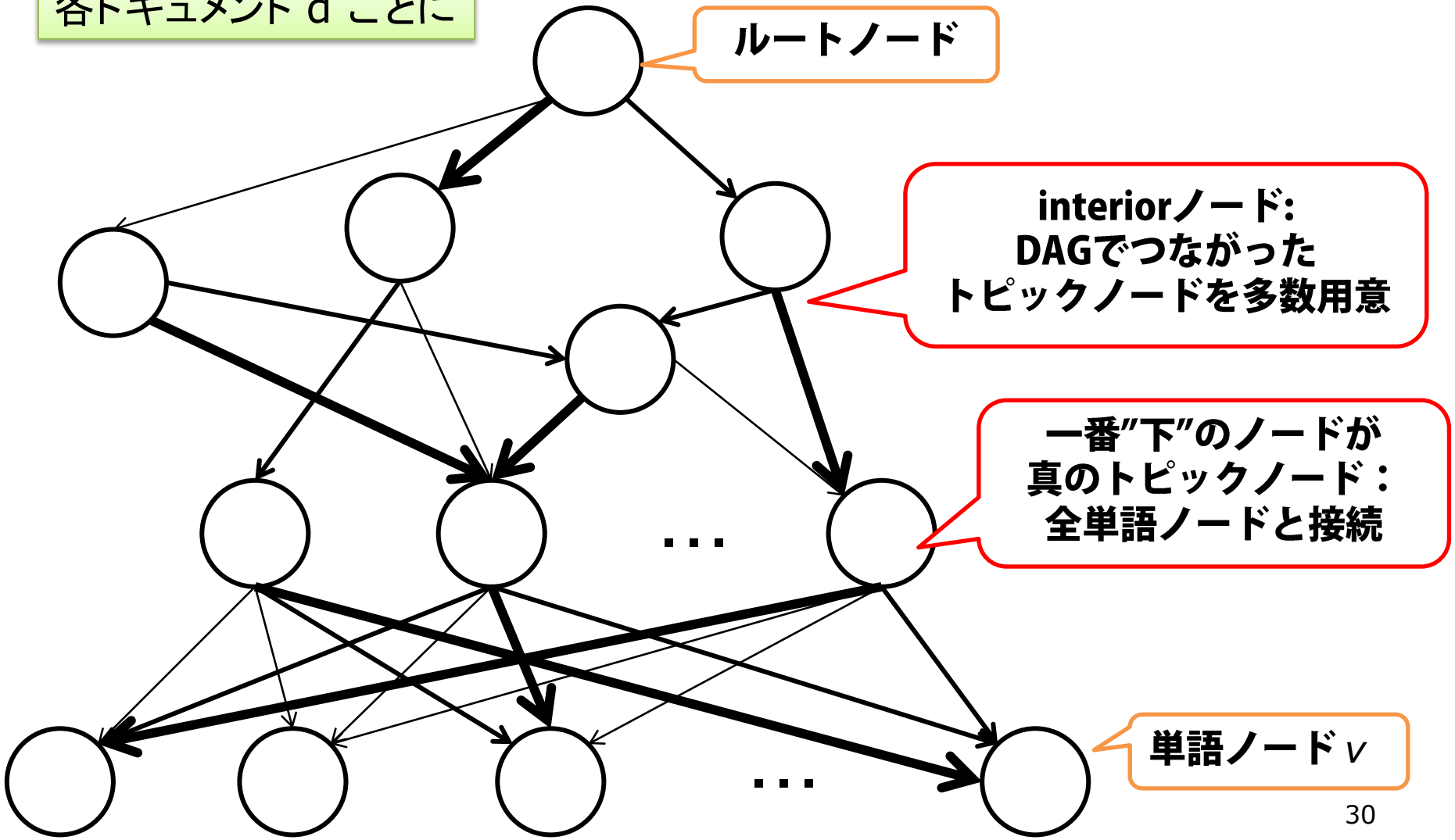


DAGによるLDA解釈



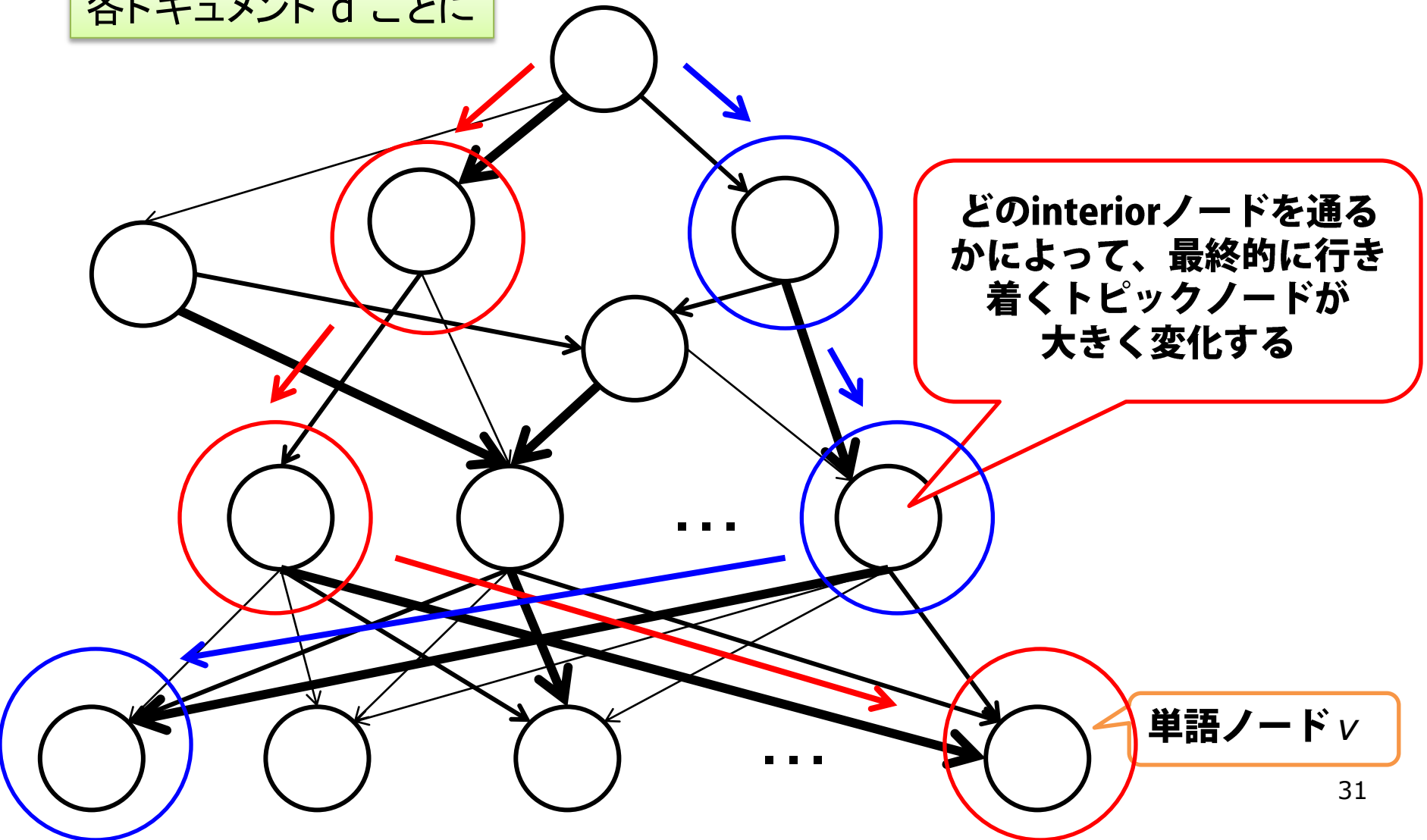
提案法のアイデア:
トピックノードをDAGで増やす

各ドキュメント d ごとに



“パチンコ玉”を落としてトピック、単語を生成します

各ドキュメント d ごとに

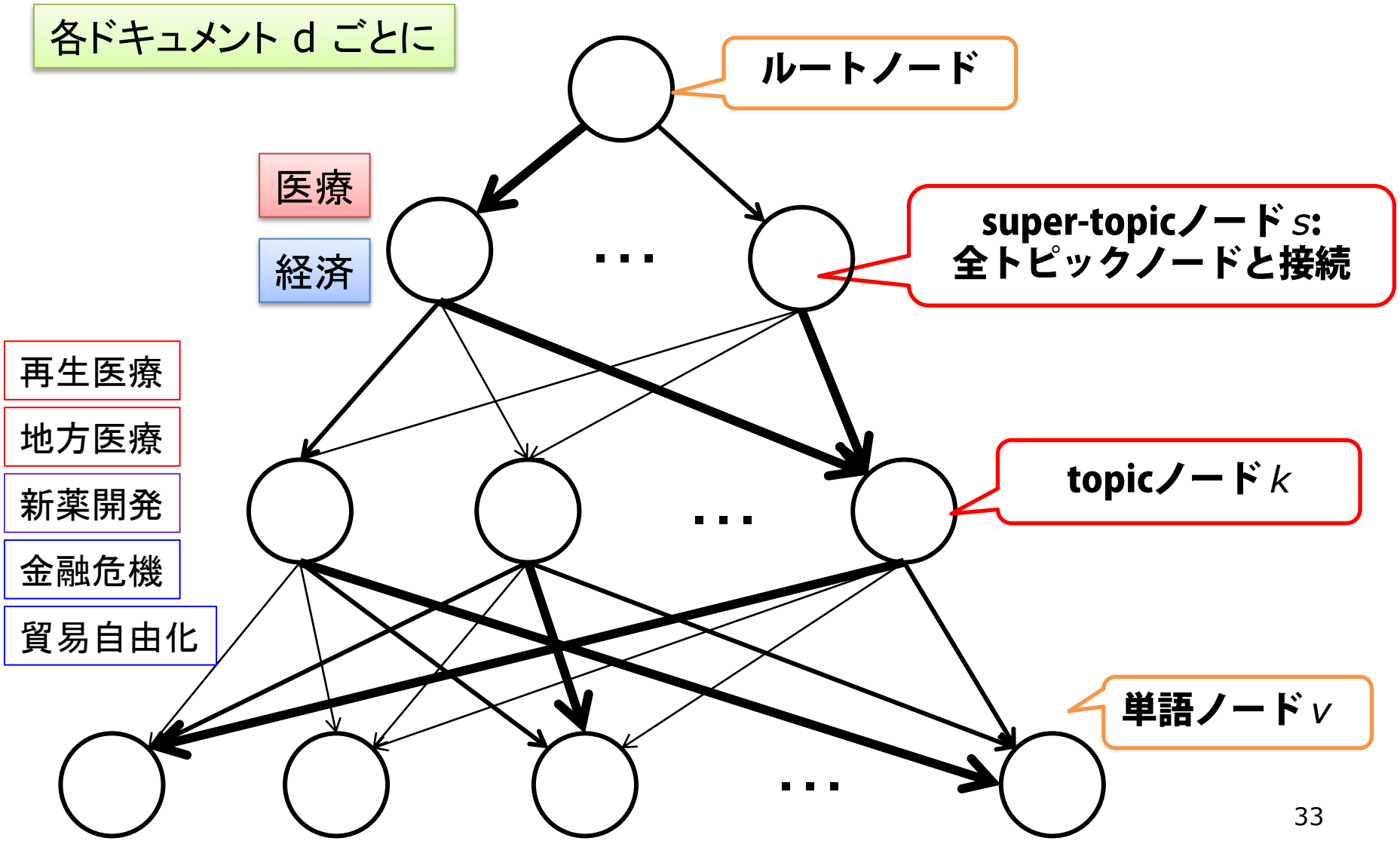


PAMの特長：表現力の高さと統一的な記述

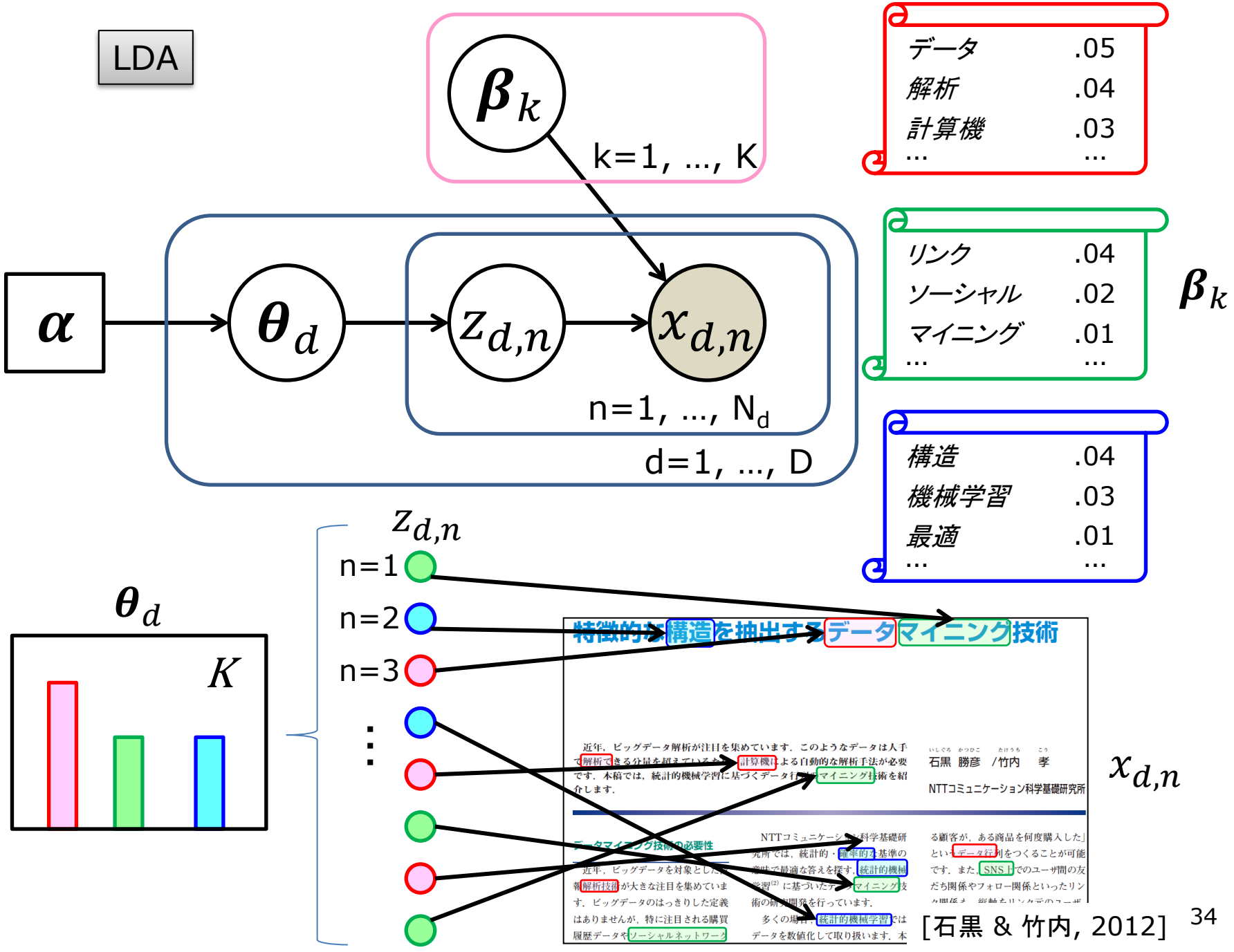
- 独立なトピックに対して、DAGで表現できるノード間の関係・構造をすべて持ち込めます
- interiorノード間の遷移確率は任意の確率分布でOK（Dir-Multが一番楽です）

- interiorノードをどのように入れても、完全に統一的な記述で表記可能です (これについては原論文を参照)

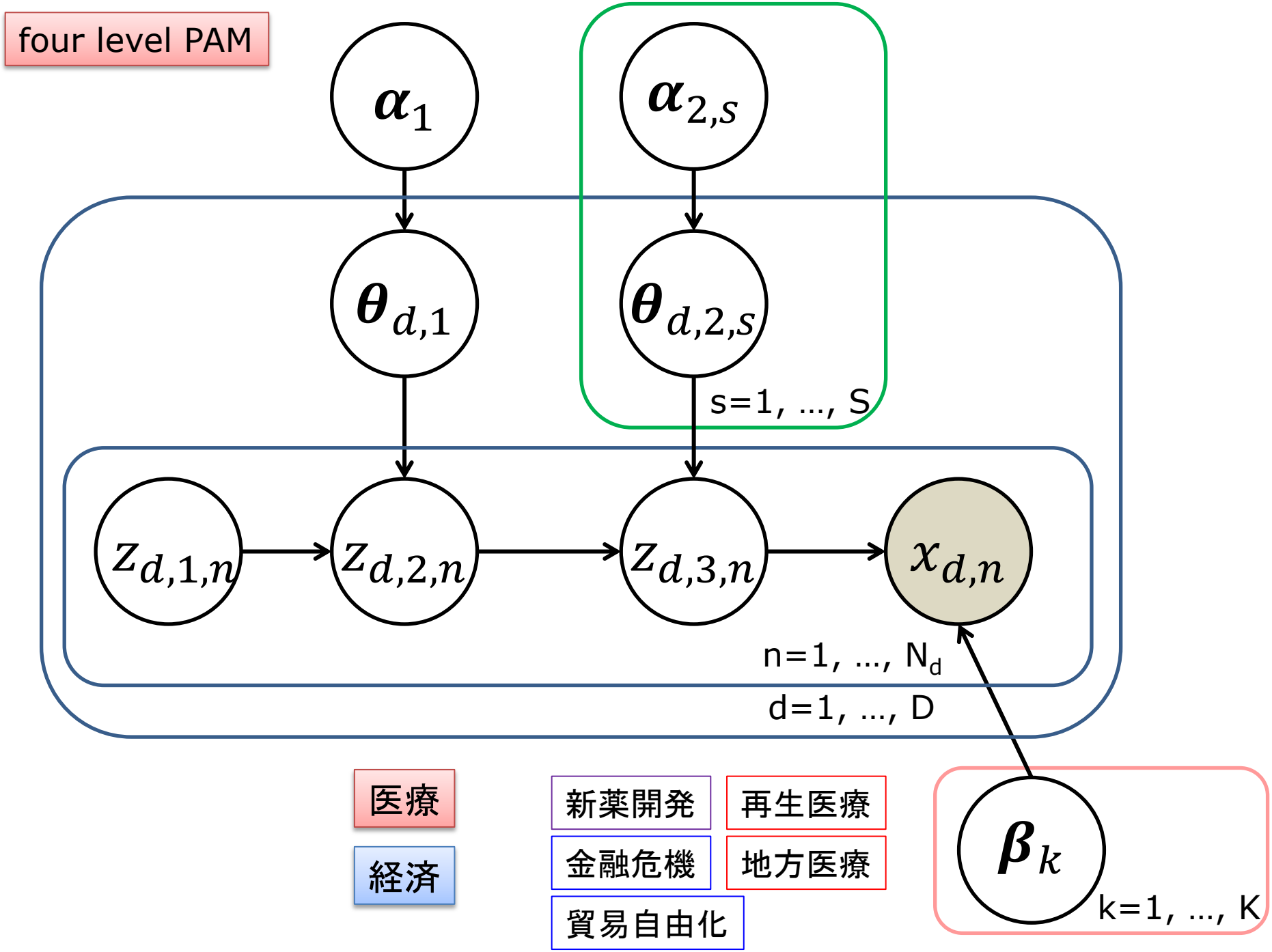
PAMの一形態: four-level PAM



LDA



[石黒 & 竹内, 2012]


four level PAM
α_1
$\alpha_{2,s}$
$\theta_{d,1}$
$\theta_{d,2,s}$
s=1, …, S
$z_{d,1,n}$
$z_{d,2,n}$
$z_{d,3,n}$
$x_{d,n}$
n=1, …, N_d
d=1, …, D
医療
経済
新薬開発
再生医療
金融危機
地方医療
貿易自由化
β_k
k=1, …, K

生成モデル

for 文書d = 1, 2, …, D_t

super-topic proportion $\boldsymbol{\theta}_{d,1}|\boldsymbol{\alpha}_1 \sim \mathrm{Dir}(\boldsymbol{\alpha}_1)$

for superトピック s = 1, 2, …, S

super-topic - topic proportion

$\boldsymbol{\theta}_{d,2,s}|\boldsymbol{\alpha}_{2,s} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{2,s})$

for 単語n = 1, 2, …, N_d

for トピック k = 1, 2, …, K

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

for 文書d = 1, 2, …, D_t

super-topic proportion $\boldsymbol{\theta}_{d,1}$

for superトピック s = 1, 2, …, S

super-topic - topic proportion $\boldsymbol{\theta}_{d,2,s}$

for 単語n = 1, 2, …, N_d

root node (const.) $\boldsymbol{z}_{d,1,n}$

super-topic - word assignment

$$z_{d,2,n}|\boldsymbol{\theta}_{d,1}\sim\mathrm{Mult}(\boldsymbol{\theta}_{d,1})$$

topic-word assignment

$$z_{d,3,n}|z_{d,2,n},\boldsymbol{\theta}_{d,2,s}\sim\mathrm{Mult}(\boldsymbol{\theta}_{d,2,z_{d,2,n}})$$

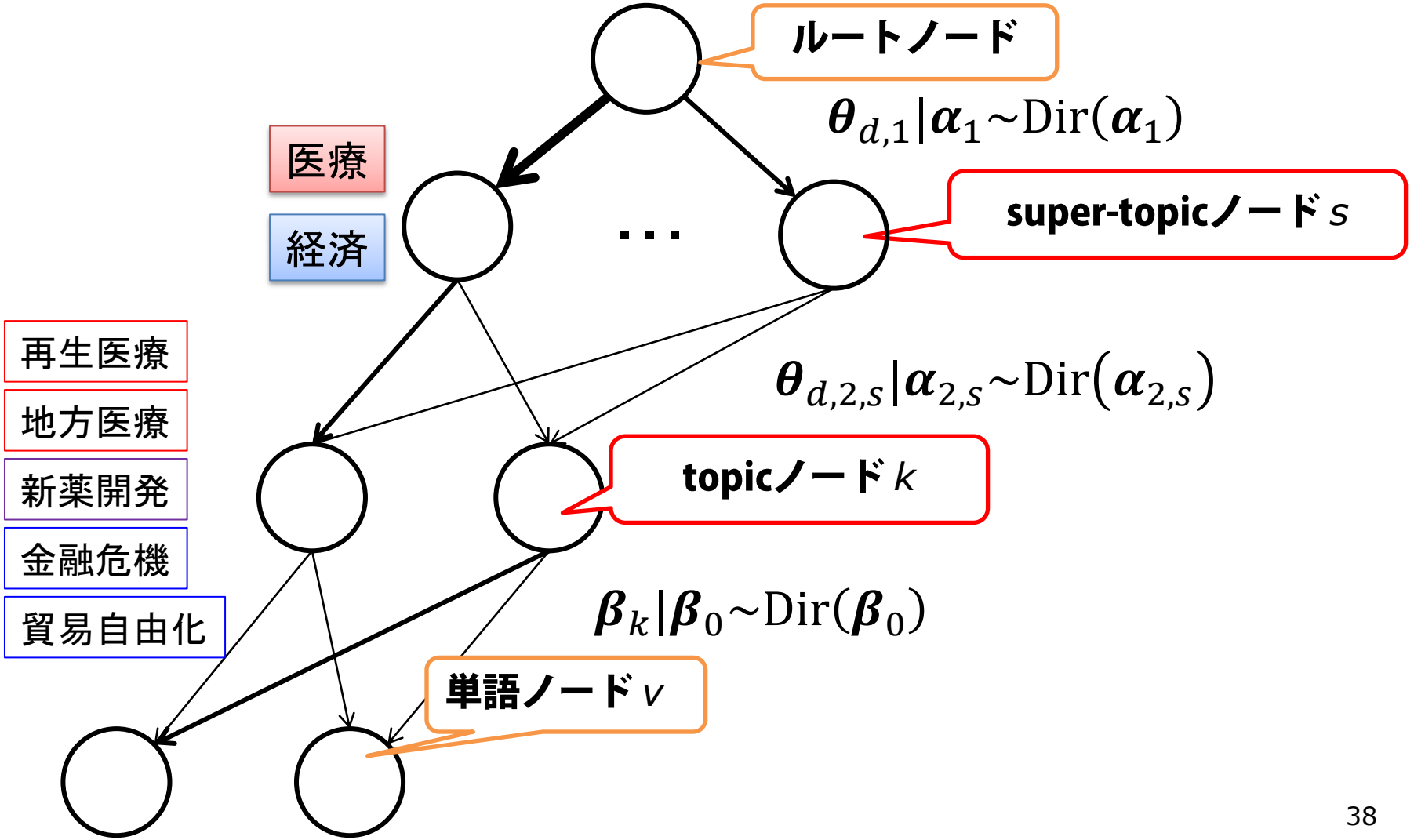
word observation

$$x_{d,n}|z_{d,3,n},\{\boldsymbol{\beta}_k\}\sim\mathrm{Mult}(\boldsymbol{\beta}_{z_{d,3,n}})$$

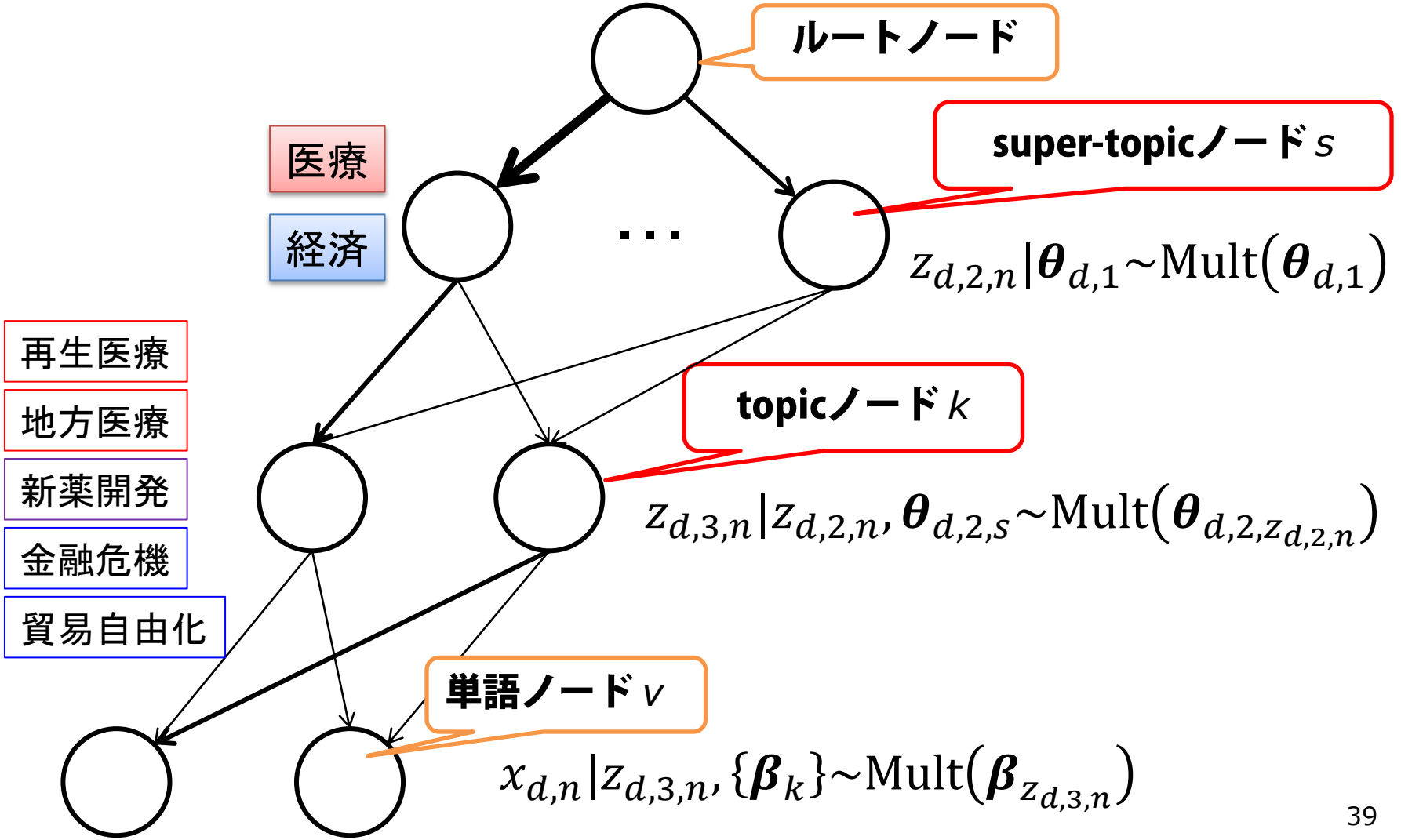
for トピック k = 1, 2, …, K

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

Dirichlet-Multinomialで統一されているので簡単です



Dirichlet-Multinomialで統一されているので簡単です



Dirichlet-Multinomialで統一されているので簡単です

- つまり、この調子で何段でも階層を重ねていくことができます

- 単純ながら、有効なモデル化戦略といえるでしょう

隠れ変数・パラメータの推定

- パラメータは積分消去可能
- 隠れ変数はGibbs samplingで最適化
  - モデルの階層性から、変分ベイズでは局所解が多すぎるようです
- ハイパーパラメータ α はモーメントマッチングで最適化します

- ☺ モデルの複雑さに反して、式の導出・計算結果は非常にシンプルです

隠れ変数のGibbsサンプリング

- 通常のLDA(Dirichlet-Multinomial)と同じ構造を持ちます

$$p\left(z_{d,2,n}=s, z_{d,3,n}=k \middle| x_{d,n}=w, \boldsymbol{X}_{\neg(d,n)}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta_0}\right) \propto$$

$$\frac{m_{ds}+\alpha_{1,s}}{\sum_{s'}\left(m_{ds'}+\alpha_{1,s'}\right)} \frac{m_{dsk}+\alpha_{2,s,k}}{\sum_{k'}\left(m_{dsk'}+\alpha_{2,s,k'}\right)} \frac{m_{kw}+\beta_{0,w}}{\sum_{w'}\left(m_{kw'}+\beta_{0,w'}\right)}$$

文書 dから superトピック sが 生成される確率

文書 dから superトピック sを経由して 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

ハイパーパラメータ α の最適化

- 普通は周辺化尤度最大化やhyper priorを仮定して推定します
- 論文ではモーメントマッチを利用しています
  - 申し訳ありませんが、講師には導出の根拠がわかりませんでした ☹

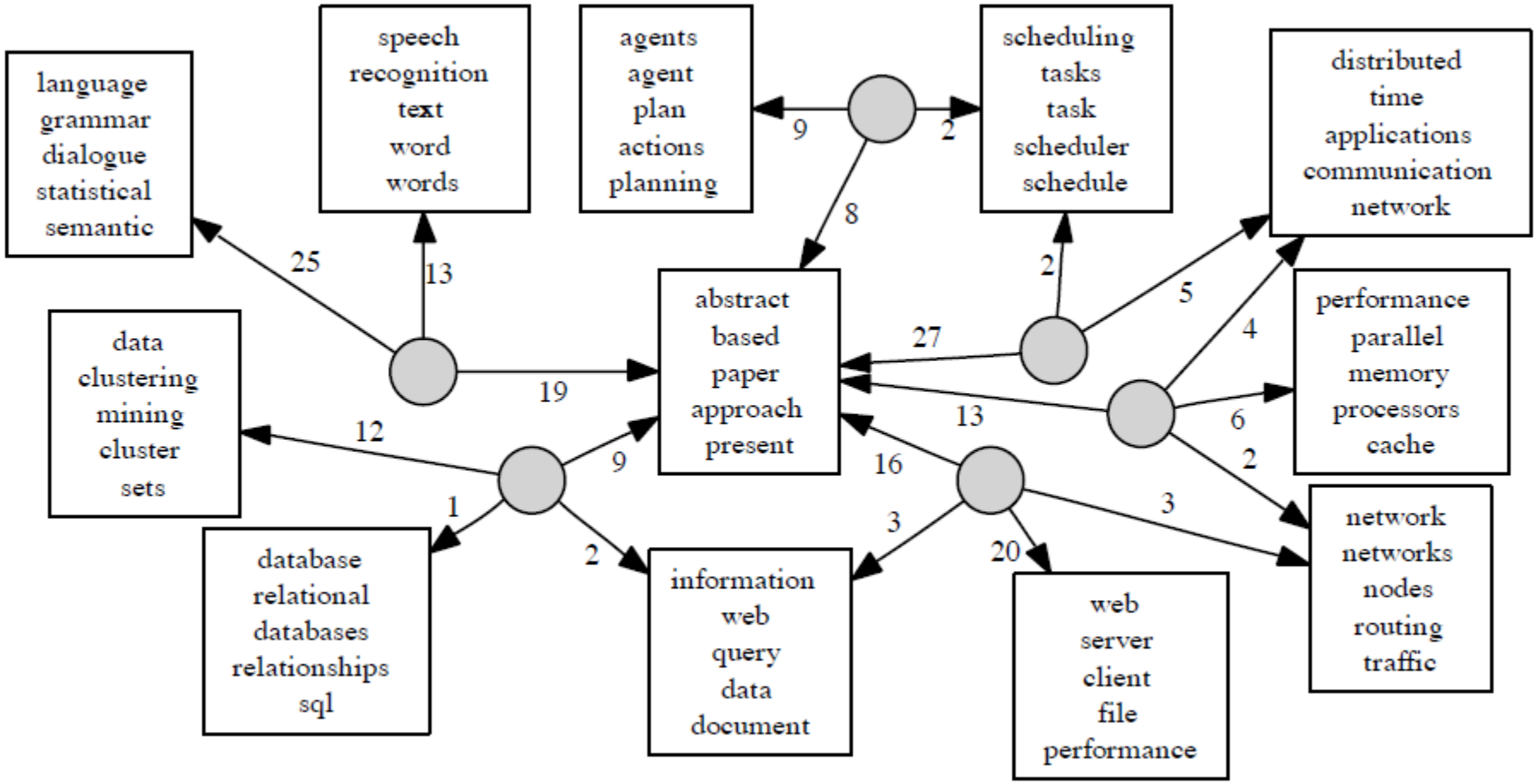
$$mean_{xy} = \frac{1}{N} \times \sum_d \frac{n_{xy}^{(d)}}{n_x^{(d)}};$$

$$var_{xy} = \frac{1}{N} \times \sum_d (\frac{n_{xy}^{(d)}}{n_x^{(d)}} - mean_{xy})^2;$$

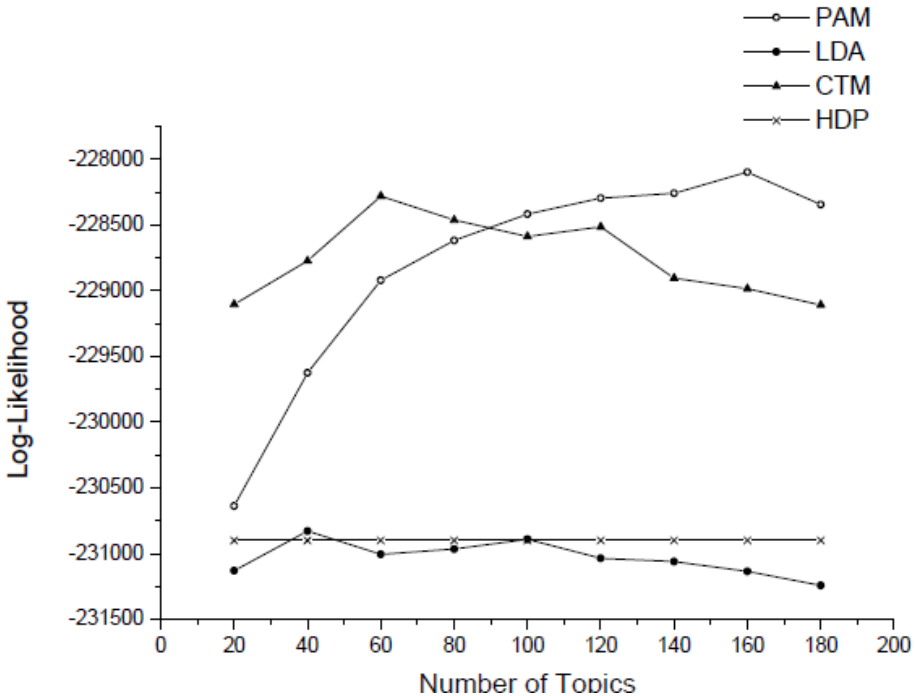
$$m_{xy} = \frac{mean_{xy} \times (1 - mean_{xy})}{var_{xy}} - 1;$$

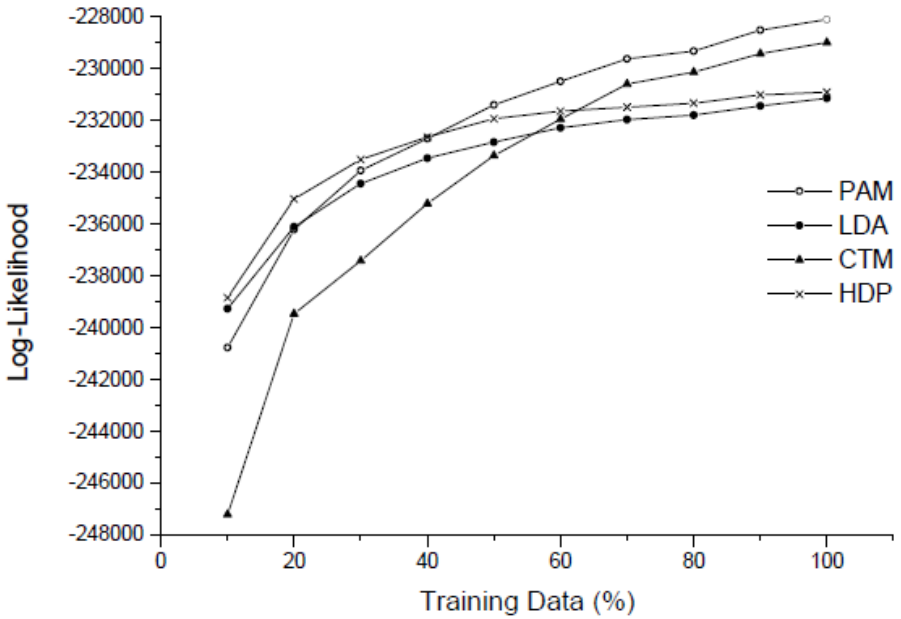
$$\alpha_{xy} \propto mean_{xy};$$

$$\sum_y \alpha_{xy} = \frac{1}{5} \times exp(\frac{\sum_y log(m_{xy})}{s_2 - 1}).$$



[Li and McCaullum, 2006]





class	# docs	LDA	PAM
graphics	243	83.95	86.83
os	239	81.59	84.10
pc	245	83.67	88.16
mac	239	86.61	89.54
windows.x	243	88.07	92.20
total	1209	84.70	87.34

Table 3. Document classification accuracy (%)

[Li and McCallum, 2006]

まとめ: Pachinko Allocation Model

- トピック間の関係構造をモデル化する、非常に柔軟かつ一般的なモデルです
- DAGによって、トピック間を任意の順番で連結してtopic pathを作ります
- 複雑なモデルですが、Gibbs samplingでシンプルかつ直観的な推論過程を実現

原論文を読むときの注意

- 論文では、一般のPAMについて生成モデルを説明しています

- また、interior nodeはすべて
  - θ ~ Dir の確率のもとで
  - z ~ Multi をサンプリングする

  として表記が統一されています

Multi-grained Topic model for aspect rating

[Titov & McDonald, 2008]

Titov and McDonald,
"Modeling Online Reviews with Multi-grained Topic Models",
in Proc. WWW, 2008.

レビュー記事のトピックモデリング

- レビュー記事はトピックモデル解析の典型的な対象です

amazon.co.jp

25 人中、23人の方が、[redacted]ています。

★★★★★ オススメです。 2012/4/10

By [redacted]

製品の品質、OSのユーザビリティともに大変良くできています。
スペックも一番安いので全く問題ありません。
配線が電源コード1本だけになるのも素晴らしい。
仕事で使ったりゲームをするのでなければWINDOWSより断然オススメできます。
ただ付属のMagic MouseはAdobeのPhotoshopやIllustratorのソフトなどをするのであればトラックパッド付属を選び、マウスは別途購入をお勧めします。

1コメント | このレビューは参考になりましたか？ はい いいえ

41 人中、36人の方が、「このレビューが参考になった」と投票しています。

★★★★★ Best of Mac 2011/8/22

By [redacted] トップ1000レビュアー VINE™ メンバー

Amazon.co.jpで購入済み

iMacはRev.Aから10台近く使っています。
アルミiMacは2台目、非常にきれいなモニター、
使いやすいワイヤレスキーボード、マジックマウスに満足しています。

マジックマウスの電池がやたら減る事も大したことじゃないし、
初心者にもハードユーザーにもお勧め。

Hotels.com

"立地は最高です"

2012/12/07 [redacted]

総合的な評価

空港から列車に乗って、駅から降りたら目の前にホテルがあるっていうのはうれしいです。朝食はビュッフェ形式でした。種類はまあまあといったところ。ただ、ヨーロッパ旅行をしていると飽きるかもです。せっかく世界遺産の広場『グランプラス』が徒歩5分くらいのところにあるので、世界遺産の広場を見ながらカフェで朝食っていうのがいいかも。朝7時くらいだとほんとに誰もいないし、広場を独占できる贅沢を味わえます。それも30分くらいのもので、7時30分には団体がきて雰囲気が台無しに。せっかく立地がいいホテルなので、早朝の広場を体感してみてください。

"駅前は最高です"

2012/10/13 [redacted]

総合的な評価

ブリュッセル中央駅出口目の前。グランプラスまで徒歩5分とかからない最高の立地です。中央駅を利用すればアントワープ、ゲント、ブルージュetc・・まで乗り換えなしで移動できるのが嬉しいです。ホテル近辺にコンビニがありませんが、駅構内に、コンビニ、カルフール、スタバなどがあり、とっても便利ですよ。ホテルは、SPGプラチナでの滞在でしたが、広いお部屋はUPGされました。基本的アメニティ(歯ブラシなし)にくわえ、バスローブ、使い捨て白いスリッパ、お水2本、ネスプレッソ4杯分、それとダンドアのスペキュロスクッキーをいただけました。 水周りは、ハンドシャワー有、ウォシュレットなし。ラウンジはないので、朝食はとりませんでしが、外へ行けば美味しいワッフル屋さんがあるので問題なしでした！！ それと、至る所に設置のある「レンタサイクル」ですが、支払いがクレカのみなのですが、日本のクレカでは使用出来ませんでした。(観光局の方に確認済み) 石畳は非常に歩きにくいため、ゴム底の楽な靴でないと、大変なことになります。総合的に、古いですが、とてもすごしやすいホテルでした。 日本語は通じませんが。。

Aspect rating

総合評価

aspects

立地

食事

観光地へのアクセス

アメニティ

Hotels.com

"立地は最高です"

2012/12/07 [redacted]

総合的な評価

空港から列車に乗って、駅から降りたら目の前にホテルがあるっていうのはうれしいです。朝食はビュッフェ形式でした。種類はまあまあといったところ。ただ、ヨーロッパ旅行をしていると飽きるかもです。せっかく「世界遺産の広場"グランプラス"」が徒歩5分くらいのところにあるので、世界遺産の広場を見ながらカフェで朝食っていうのがいいかも。朝7時くらいだとほんとに誰もいないし、広場を独占できる贅沢を味わえます。それも30分くらいのもので、7時30分には団体がきて雰囲気が台無しに。せっかく立地がいいホテルなので、早朝の広場を体感してみてください。

"駅前は最高です"

2012/10/13 [redacted]

総合的な評価

ブリュッセル中央駅出口目の前。グランプラスまで徒歩5分とかからない最高の立地です。中央駅を利用すればアントワープ、ゲント、ブルージュetc・・まで乗り換えなしで移動できるのが嬉しいです。ホテル近辺にコンビニがありませんが、駅構内に、コンビニ、カルフール、スタバなどがあり、とっても便利ですよ。ホテルは、SPGプラチナでの滞在でしたが、広いお部屋はUPGされました。基本的アメニティ(歯ブラシなし)にくわえ、バスローブ、使い捨て白いスリッパ、お水2本、ネスプレッソ4杯分、それとダンドアのスペキュロスクッキーをいただけました。 水周りは、ハンドシャワー有、ウォシュレットなし。ラウンジはないので、朝食はとりませんでしが、外へ行けば美味しいワッフル屋さんがあるので問題なしでした！！それと、至る所に設置のある「レンタサイクル」ですが、支払いがクレカのみなのですが、日本のクレカでは使用出来ませんでした。(観光局の方に確認済み) 石畳は非常に歩きにくいため、ゴム底の楽な靴でないと、大変なことになります。総合的に、古いですが、とてもすごしやすいホテルでした。 日本語は通じませんが。。

統計モデルによるAspect分析

- 統計モデルによる客観的・自動的なaspect分析が可能となると、より詳細なサービス評価・レビュー解析が可能になります

総合評価：4

ratable aspects

- 値段：A
- 性能：A
- アフターサービス：C
- 使いやすさ：B

amazon.co.jp

25 人中、23人の方 ています。

☆☆☆☆☆ オススメです。 2012/1/10

By [redacted]

製品の品質、OSのユーザビリティともに大変良くできています。
スペックも一番安いので全く問題ありません。
配線が電源コード1本だけになるのも素晴らしい。
仕事で使ったりゲームをするのでなければWINDOWSより断然オススメできます。
ただ付属のMagic MouseはAdobeのPhotoshopやIllustratorのソフトなどをお使いの場合、
トラックパッド付属を選び、マウスは別途購入をお勧めします。

1コメント | このレビューは参考になりましたか？ はい いいえ

41 人中、36人の方が、「このレビューが参考になった」と投票しています。

☆☆☆☆☆ Best of Mac 2011/8/22

By [redacted] トップ1000レビュアー VINE™ メンバー

Amazon.co.jpで購入済み

iMacはRev.Aから10台近く使っています。
アルミiMacは2台目、非常にきれいなモニター、
使いやすいワイヤレスキーボード、マジックマウスに満足しています。

マジックマウスの電池がやたら減る事も大したことじゃないし、
初心者にもハードユーザーにもお勧め。

提案法: Multi-grained Topic Models

- 主にレビュー記事を対象に、文章を全体的なトピックとaspectに関するトピックに分解するモデルです
- トピックの相関ではなく、トピックの種類を増やした構造をもつモデルです
- 推論はモデルの複雑さに反してシンプルです

提案法のアイデア:
2種類のトピックを仮定する

グローバルトピック: レビュー対象の全体的な特徴

ホテル記事ならば・・・”ロンドンのホテル” “ビジネスホテル”など

ロンドン	.05
地下鉄	.04
五輪	.03
…	…

機能的	.04
お手頃	.02
ネット接続	.01
…	…

ローカルトピック = ratable aspect

ホテル記事ならば・・・”アクセス” “値段” “食事” “立地” “部屋の広さ”など

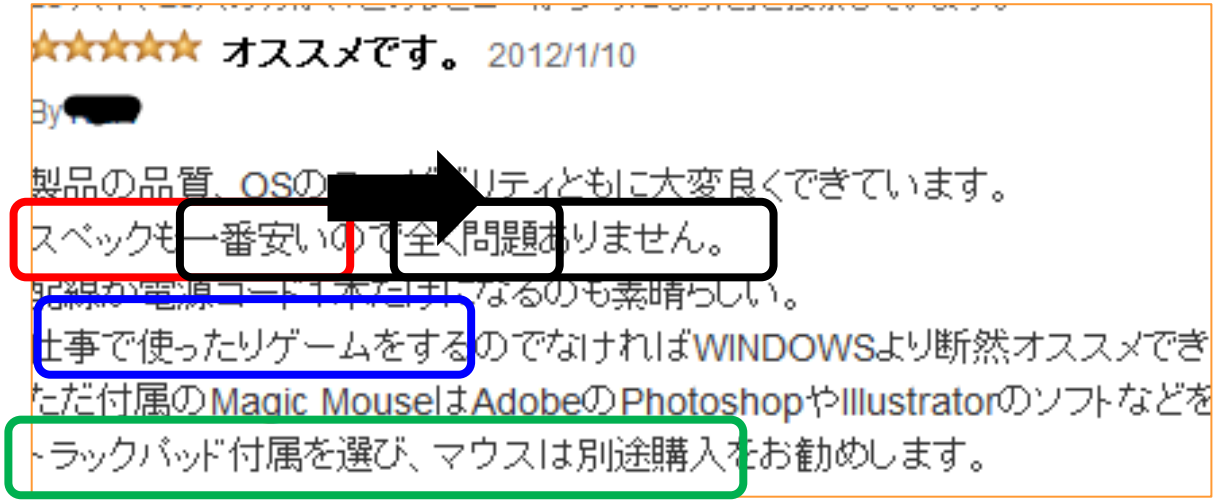
駅前	.04
徒歩圏内	.03
バス	.01
…	…

リーズナブル	.03
見合った	.03
納得できる	.02
…	…

おいしい	.07
ルームサービス	.03
味が濃い	.02
…	…

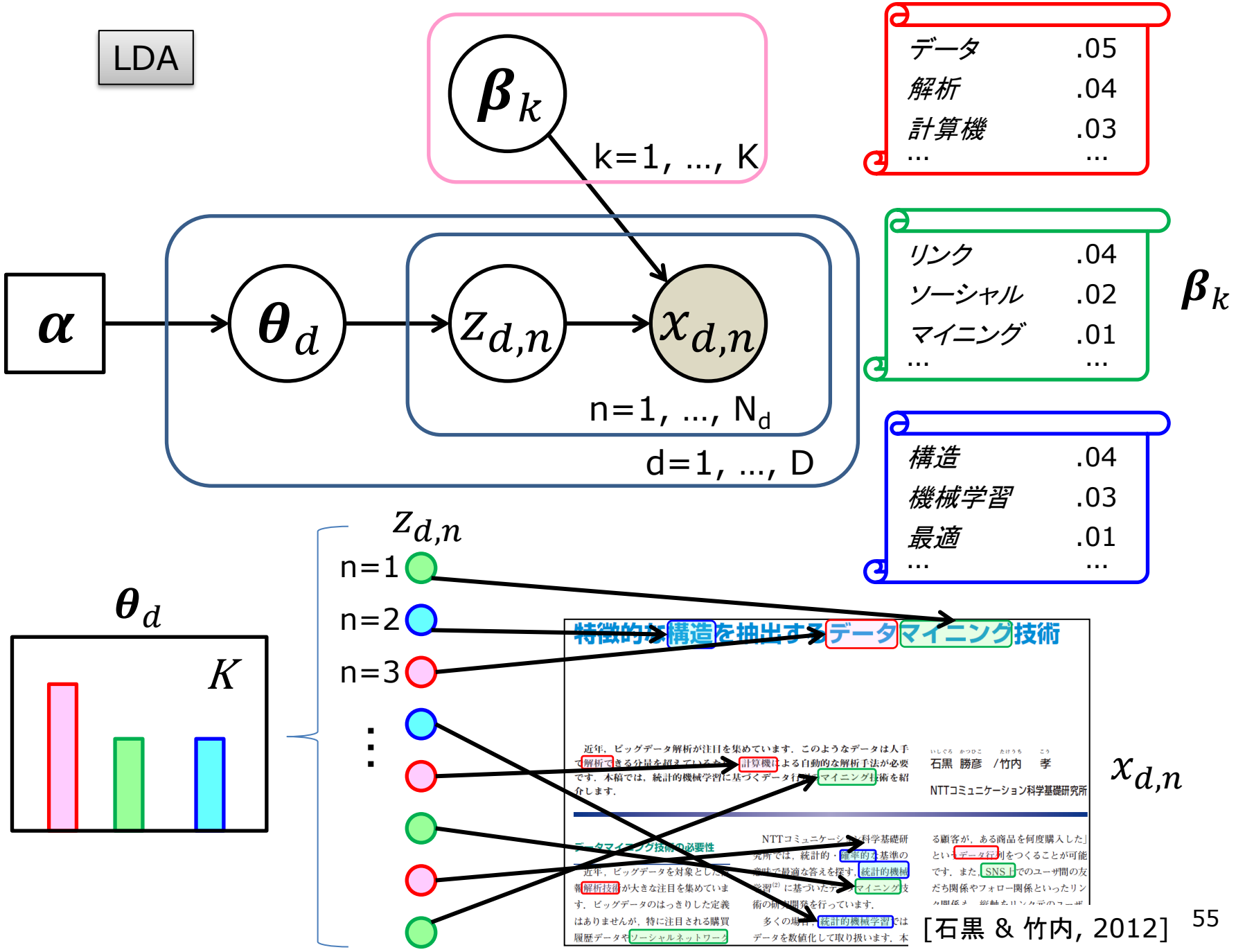
提案法のアイデア: Sliding window

- local topicは文章の一部でしか出てこないので、sliding windowで局所的にモデル化



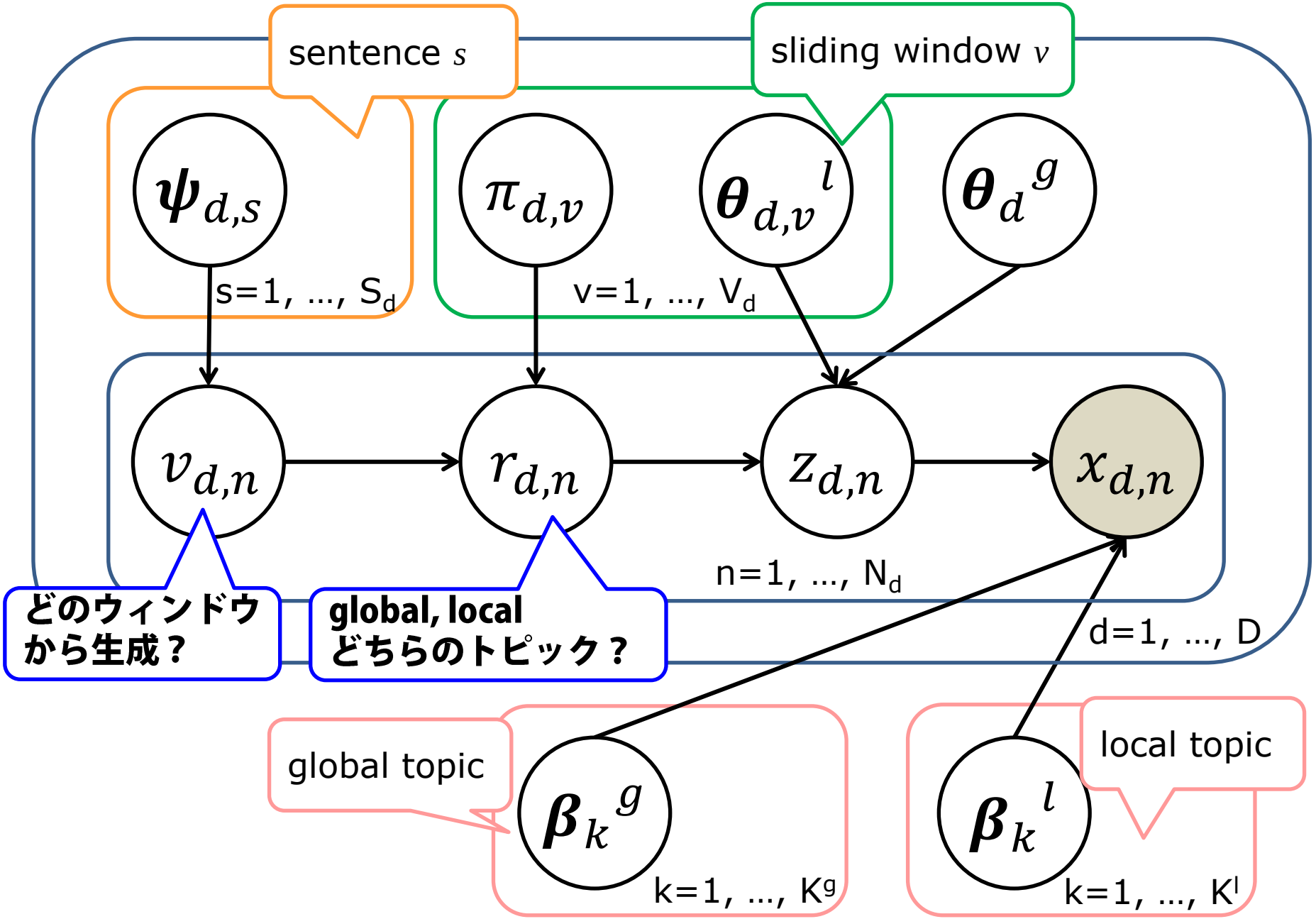
local topic 1 (値段) の割合	**小**	**大**	**中**
local topic 2 (用途)の割合	**大**	**中**	**小**
local topic 3 (付属品)の割合	**小**	**小**	**大**

LDA



[石黒 & 竹内, 2012]

Multi-grained Topic Models (ハイパーパラメータを省略)



for 文書d = 1, 2, …, D_t

for sentence s = 1, 2, …, S_d

window proportion $\boldsymbol{\psi}_{d,s}|\boldsymbol{\gamma}\sim\mathrm{Dir}(\boldsymbol{\gamma})$

for sliding window v = 1, 2, …, V_d

global-local proportion $\pi_{d,v}|\boldsymbol{\alpha}^{mix}\sim\mathrm{Beta}(\boldsymbol{\alpha}^{mix})$

local-topic proportion $\boldsymbol{\theta}_{d,v}{}^{l}|\boldsymbol{\alpha}^{l}\sim\mathrm{Dir}(\boldsymbol{\alpha}^{l})$

global-topic proportion $\boldsymbol{\theta}_{d}{}^{g}|\boldsymbol{\alpha}^{g}\sim\mathrm{Dir}(\boldsymbol{\alpha}^{g})$

for Localトピック k = 1, 2, …, K^l

local topic-word proportion $\boldsymbol{\beta}_{k}{}^{l}|\boldsymbol{\beta}_{0}{}^{l}\sim\mathrm{Dir}(\boldsymbol{\beta}_{0}{}^{l})$

for Globalトピック k = 1, 2, …, K^g

global topic-word proportion $\boldsymbol{\beta}_{k}{}^{g}|\boldsymbol{\beta}_{0}{}^{g}\sim\mathrm{Dir}(\boldsymbol{\beta}_{0}{}^{g})$

for 文書d = 1, 2, …, D_t

for sentence s = 1, 2, …, S_d window proportion $\boldsymbol{\psi}_{d,s}$

global-topic proportion $\boldsymbol{\theta}_d{}^{g}$

for sliding window v = 1, 2, …, V_d global-local proportion $\boldsymbol{\pi}_{d,v}$

local-topic proportion $\boldsymbol{\theta}_{d,v}{}^{l}$

for 単語n = 1, 2, …, N_d in sentence s

window-word assignment

$$v_{d,n}|\boldsymbol{\psi}_{d,s} \sim \mathrm{Mult}(\boldsymbol{\psi}_{d,s})$$

global/local-word assignment

$$r_{d,n}|\pi_{d,v}, v_{d,n} \sim \mathrm{Bernoulli}(\pi_{d,v_{d,n}})$$

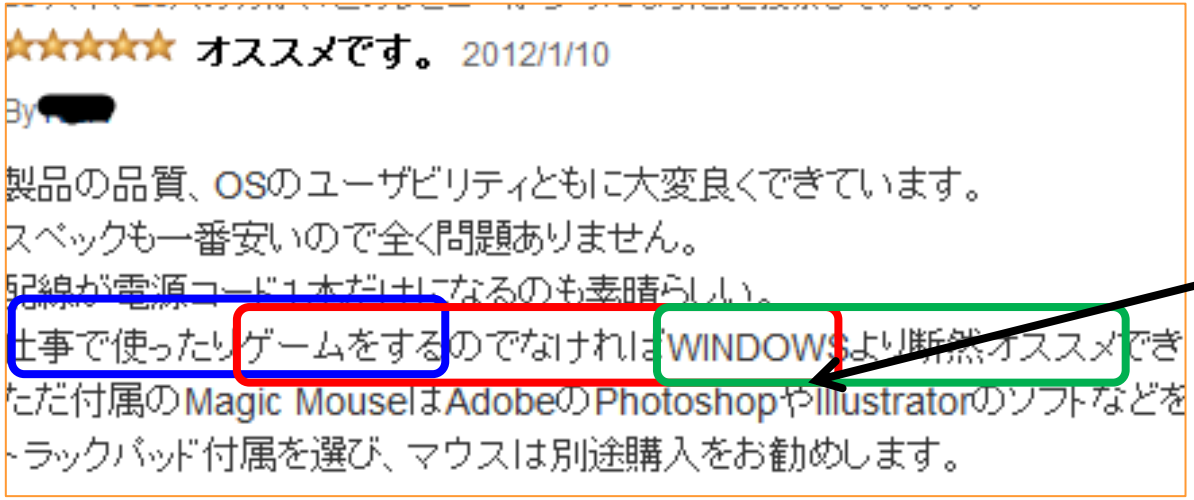
topic-word assignment

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

word observation

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

sliding windowもDirichlet-Multinomialで簡単です



sentence sの window選択確率

$$\psi_{d,s}|\gamma \sim \mathrm{Dir}(\gamma)$$

グローバルトピック

ロンドン	.05
地下鉄	.04
五輪	.03
…	…

window vのglobal/local割合

$$\pi_{d,v}|\alpha^{mix} \sim \mathrm{Beta}(\alpha^{mix})$$

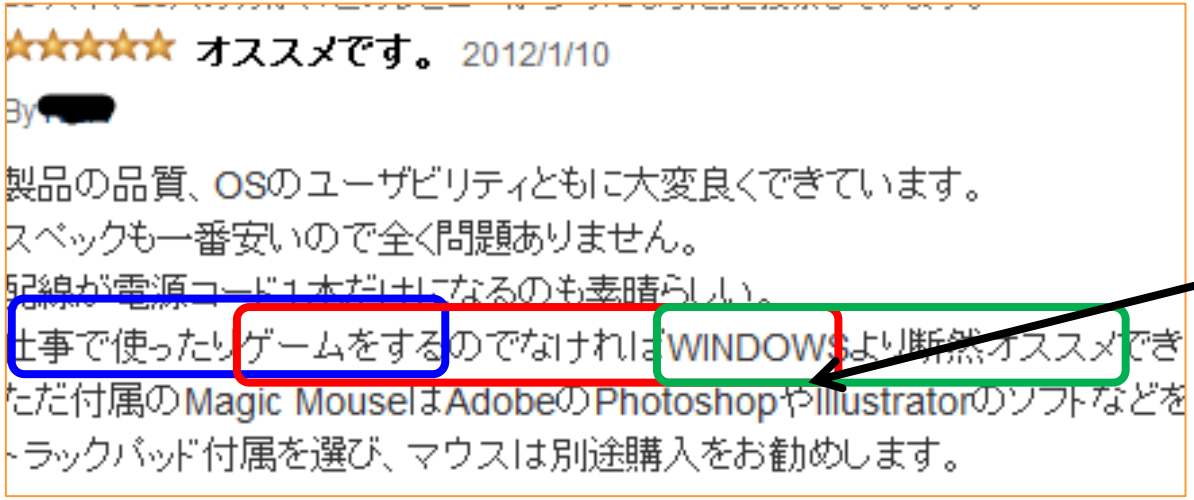
ローカルトピック

駅前	.04
徒歩圏内	.03
バス	.01
…	…

選んだトピックの種類の中で実際のトピックを決定

$$\theta_d{}^g|\alpha^g \sim \mathrm{Dir}(\alpha^g) \quad \theta_{d,v}{}^l|\alpha^l \sim \mathrm{Dir}(\alpha^l)$$

sliding windowもDirichlet-Multinomialで簡単です



sentence s内の単語 nが生成されるwindow v

$$v_{d,n}|\boldsymbol{\psi}_{d,s}\sim\mathrm{Mult}(\boldsymbol{\psi}_{d,s})$$

グローバルトピック

ロンドン	.05
地下鉄	.04
五輪	.03
…	…

ローカルトピック

駅前	.04
徒歩圏内	.03
バス	.01
…	…

window vのglobal/local割合でトピックの種類を選択

$$r_{d,n}|\pi_{d,v},v_{d,n}\sim\mathrm{Bernoulli}(\pi_{d,v_{d,n}})$$

選んだトピックの種類の中で実際のトピック kを決定

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

隠れ変数・パラメータの推定

- パラメータは積分消去可能
- 隠れ変数はGibbs samplingで最適化

- ☺ PAMと同様に、モデルの複雑さに反して、式の導出・計算結果は非常にシンプルです

隠れ変数のGibbsサンプリング

- 通常のLDA(Dirichlet-Multinomial)とほとんど同じ式を導出できます

$$p\left(v_{d,n}=v, r_{d,n}=g, z_{d,n}=k \middle| x_{d,n}=w, \{\boldsymbol{X},\boldsymbol{V},\boldsymbol{R},\boldsymbol{Z}\}_{\neg(d,n)}\right)$$

$$\propto \frac{m_{dsv}+\gamma_v}{\sum_{v'}(m_{\mathrm{ds}v'}+\gamma_{v'})} \frac{m_{dvg}+{\alpha_g}^{mix}}{\sum_{r'\in\{g,l\}}(m_{dvr'}+{\alpha_{r'}}^{mix})}$$

sentence sの単語のうち
window vから生成された単語の割合

文書 dのwindow vから
globalトピックが出てくる割合

$$\times \frac{{m_{dk}}^{g}+{\alpha_k}^{g}}{\sum_{k'}({m_{dk'}}^{g}+{\alpha_{k'}}^{g})} \frac{{m_{kv}}^{g}+{\beta_{0,w}}^{g}}{\sum_{w'}\left({m_{kw'}}^{g}+{\beta_{0,w'}}^{g}\right)}$$

文書 dの中でglobalトピック kが使われる割合

トピック kから単語 wが生成される割合

隠れ変数のGibbsサンプリング

- 通常のLDA(Dirichlet-Multinomial)とほとんど同じ式を導出できます

$$p\left(v_{d,n}=v, r_{d,n}=l, z_{d,n}=k \middle| x_{d,n}=w, \{\boldsymbol{X},\boldsymbol{V},\boldsymbol{R},\boldsymbol{Z}\}_{\neg(d,n)}\right)$$

$$\propto \frac{m_{dsv}+\gamma_v}{\sum_{v'}(m_{\mathrm{d}sv'}+\gamma_{v'})} \frac{m_{dvl}+{\alpha_l}^{mix}}{\sum_{r'\in\{g,l\}}(m_{dvr'}+{\alpha_{r'}}^{mix})}$$

sentence sの単語のうち
window vから生成された単語の割合

文書 dのwindow vから
localトピックが出てくる割合

$$\times \frac{{m_{dk}}^{l}+{\alpha_k}^{l}}{\sum_{k'}({m_{dk'}}^{l}+{\alpha_{k'}}^{l})} \frac{{m_{kv}}^{l}+{\beta_{0,w}}^{l}}{\sum_{w'}\left({m_{kw'}}^{l}+{\beta_{0,w'}}^{l}\right)}$$

文書 dの中でlocalトピック kが使われる割合

トピック kから単語 wが生成される割合

Table 2: Top words from MG-LDA and LDA topics for Mp3 players' reviews.

	label	top words
MG-LDA local (all topics)	sound quality	sound quality headphones volume bass earphones good settings ear rock excellent
	features	games features clock contacts calendar alarm notes game quiz feature extras solitaire
	connection with PC	usb pc windows port transfer computer mac software cable xp connection plug firewire
	tech. problems	reset noise backlight slow freeze turn remove playing icon creates hot cause disconnect
	appearance	case pocket silver screen plastic clip easily small blue black light white belt cover
	controls	button play track menu song buttons volume album tracks artist screen press select
	battery	battery hours life batteries charge aaa rechargeable time power lasts hour charged
	accessories	usb cable headphones adapter remote plug power charger included case firewire
	managing files	files software music computer transfer windows media cd pc drag drop file using
	radio/recording	radio fm voice recording record recorder audio mp3 microphone wma formats
MG-LDA global	iPod	ipod music apple songs use mini very just itunes like easy great time new buy really
	Creative Zen	zen creative micro touch xtra pad nomad waiting deleted labs nx sensitive 5gb eax
	Sony Walkman	sony walkman memory stick sonicstage players atrac3 mb atrac far software format
	video players	video screen videos device photos tv archos pictures camera movies dvd files view
	support	player product did just bought unit got buy work $ problem support time months
LDA (out of 40)	iPod	ipod music songs itunes mini apple battery use very computer easy time just song
	Creative	creative nomad zen xtra jukebox eax labs concert effects nx 60gb experience lyrics
	memory/battery	card memory cards sd flash batteries lyra battery aa slot compact extra mmc 32mb
	radio/recording	radio fm recording record device audio voice unit battery features usb recorder
	controls	button menu track play volume buttons player song tracks press mode screen settings
	opinion	points reviews review negative bad general none comments good please content aware
	-	player very use mp3 good sound battery great easy songs quality like just music

[Titov & McDonald, 2008]

Table 3: Top words from MG-LDA and LDA topics for hotel reviews.

	label	top words
MG-LDA local (all topics)	amenities	coffee microwave fridge tv ice room refrigerator machine kitchen maker iron dryer
	food and drink	food restaurant bar good dinner service breakfast ate eat drinks menu buffet meal
	noise/conditioning	air noise door room hear open night conditioning loud window noisy doors windows
	bathroom	shower water bathroom hot towels toilet tub bath sink pressure soap shampoo
	breakfast	breakfast coffee continental morning fruit fresh buffet included free hot juice
	spa	pool area hot tub indoor nice swimming outdoor fitness spa heated use kids
	parking	parking car park lot valet garage free street parked rental cars spaces space
	staff	staff friendly helpful very desk extremely help directions courteous concierge
	Internet	internet free access wireless use lobby high computer available speed business
	getting there	airport shuttle minutes bus took taxi train hour ride station cab driver line
	check in	early check morning arrived late hours pm ready day hour flight wait
	smells/stains	room smoking bathroom smoke carpet wall smell walls light ceiling dirty
	comfort	room bed beds bathroom comfortable large size tv king small double bedroom
	location	walk walking restaurants distance street away close location shopping shops
	pricing	$ night rate price paid worth pay cost charge extra day fee parking
MG-LDA global	beach resorts	beach ocean view hilton balcony resort ritz island head club pool oceanfront
	Las Vegas	vegas strip casino las rock hard station palace pool circus renaissance
LDA (out of 45)	beach resorts	beach great pool very place ocean stay view just nice stayed clean beautiful
	Las Vegas	vegas strip great casino $ good hotel food las rock room very pool nice
	smells/stains	room did smoking bed night stay got went like desk smoke non-smoking smell
	getting there	airport hotel shuttle bus very minutes flight hour free did taxi train car
	breakfast	breakfast coffee fruit room juice fresh eggs continental very toast morning
	location	hotel rooms very centre situated well location excellent city comfortable good
	pricing	card credit $ charged hotel night room charge money deposit stay pay cash did
	front desk	room hotel told desk did manager asked said service called stay rooms
	noise	room very hotel night noise did hear sleep bed door stay floor time just like
	opinion	hotel best stay hotels stayed reviews service great time really just say rooms
	cleanliness	hotel room dirty stay bathroom rooms like place carpet old very worst bed
	-	motel rooms nice hotel like place stay parking price $ santa stayed good

[Titov & McDonald, 2008]

Table 4: Multi-aspect ranking experiments with the PRanking algorithm for hotel reviews.

Unigram features only

Model	Overall	Check-in	Service	Value	Location	Rooms	Cleanliness
Baseline	1.118	1.126	1.208	1.272	0.742	1.356	1.002
PRank	0.774	0.831	0.799	0.793	0.707	0.798	0.715
PRank + LDA	0.735	0.786	0.762	0.749	0.677	0.746	0.690
PRank + MG-LDA	**0.706**	**0.748**	**0.731**	**0.725**	**0.635**	**0.719**	**0.676**

Unigram, bigram and trigram features

Model	Overall	Check-in	Service	Value	Location	Rooms	Cleanliness
PRank	0.689	0.735	0.725	0.710	0.627	0.700	0.637
PRank + LDA	0.682	0.728	0.717	0.705	0.620	0.684	0.637
PRank + MG-LDA	**0.669**	**0.717**	**0.700**	**0.696**	**0.607**	**0.672**	0.636

[Titov & McDonald, 2008]

まとめ: Multi-grained Topic Models

- 主にレビュー記事を対象に、文章を全体的なトピックとaspectに関するトピックに分解するモデルです
- Sliding windowを基準にしたトピックモデルになります
- 推論はモデルの複雑さに反してシンプルです

その他の構造トピックモデル

- Blei et al., “Hierarchical Topic Models and the Nested Chinese Restaurant Process”, in Advances in Neural Information Processing Systems 16 (Proc. NIPS), 2003.

- Titov and McDonald, “A Joint Model of Text and Aspect Ratings for Sentiment Summarization”, in Proc. ACL, 2008.

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 & Lafferty, 2007] 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.
- [Li & McCallum, 2006] Li and McCallum, “Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations”, in Proc. ICML, 2006.
- [Titov & McDonald, 2008] Titov and McDonald, “Modeling Online Reviews with Multi-grained Topic Models”, in Proc. WWW, 2008.