NTT

トピックモデルの応用：時系列データ

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

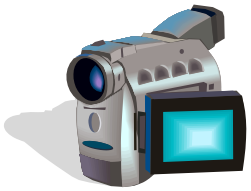
このスライドの“トピック”

- 購買データや科学論文など、時間変化をそもそも内包するデータは多数存在します

- 従って、時系列（時間変化）データ内のトピックの解析も多数試みがなされています

時系列データは数多く存在します

動画像



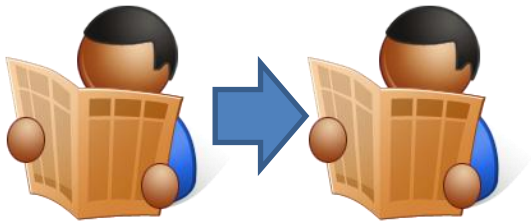
脳波・生体信号

購買履歴・市場インデックス

音楽・音響信号

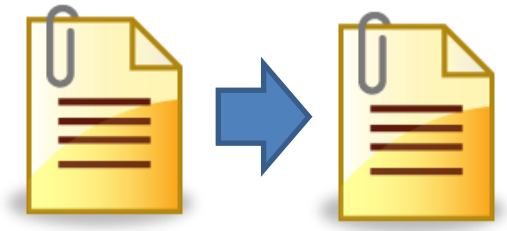
新聞・ニュース記事

01/15 01/16



科学論文・特許

2012 2013



☺ トピックモデルの有用性が明白になったため、これら時系列データへトピックモデルを応用した研究が多数発表されています

時系列データモデリングのキモ：どこに時間依存性を入れるか？

- マルコフ性：前の時刻に依存して現在の時刻の状態が変化する

- 多くの時系列データでは、モデルのどの部分にマルコフ性のアイデアを導入するか、がポイントとなります

- これはトピックモデルの時系列データモデルでも同様です

Dynamic Topic Model
[Blei & Lafferty, 2006]

Blei and Lafferty,
"Dynamic Topic Models",
in Proc. ICML, 2006.

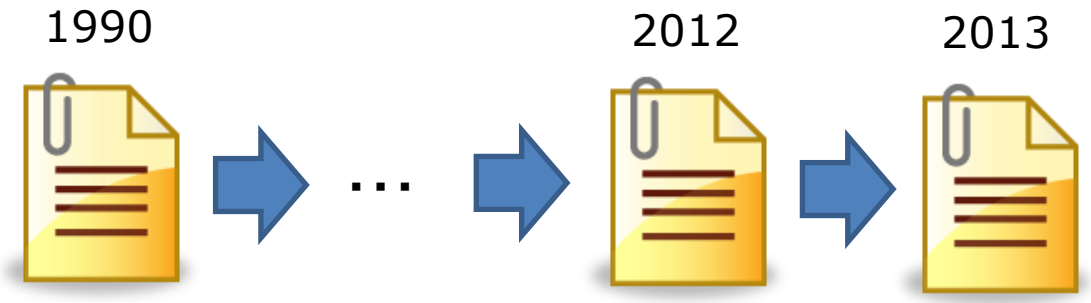
トピックモデルの大前提の仮定: exchangeability

- 簡単にいうと:「各文書 d, 各単語 wのインデックスはただのシンボルで順番や名前には意味が無い」

- これのおかげで各種モデル推論が簡単になっています

文書コレクションの時系列データを考えると、これは問題です

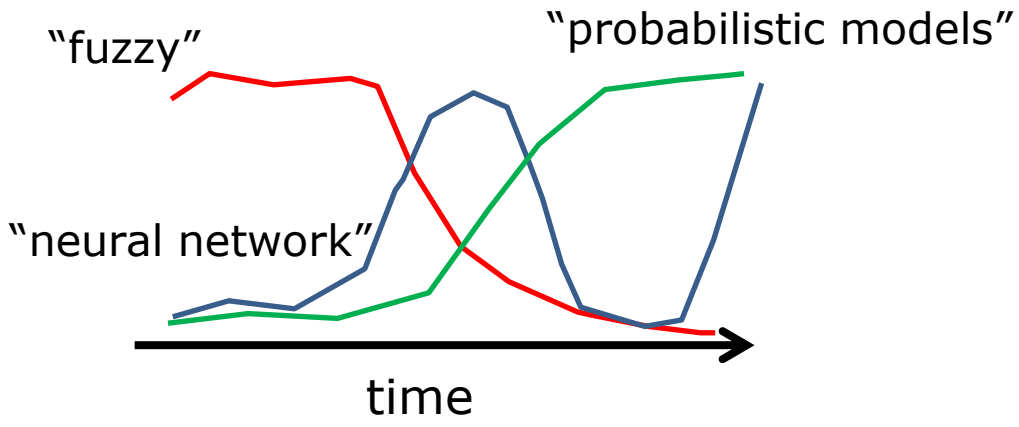
- 新聞記事は昨日までの報道の流れを汲んでいます
- 論文は先行研究の作った技術トレンドにのっています
- すなわち、文書 d は一般にはexchangeableではありません！



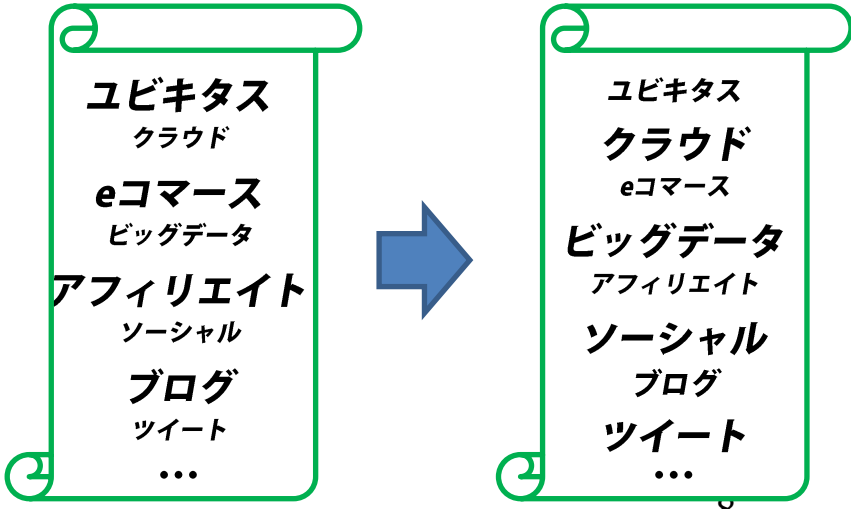
注目する時間依存性：トピック

- 1: 話題(topic)には流行り廃りがあります
- 2: トピックの中での言葉づかいも変化します
- これら2種類の「トピックの変化」を解析するモデルを考えたい

トピックの流行り廃り



トピックの中での言葉づかい

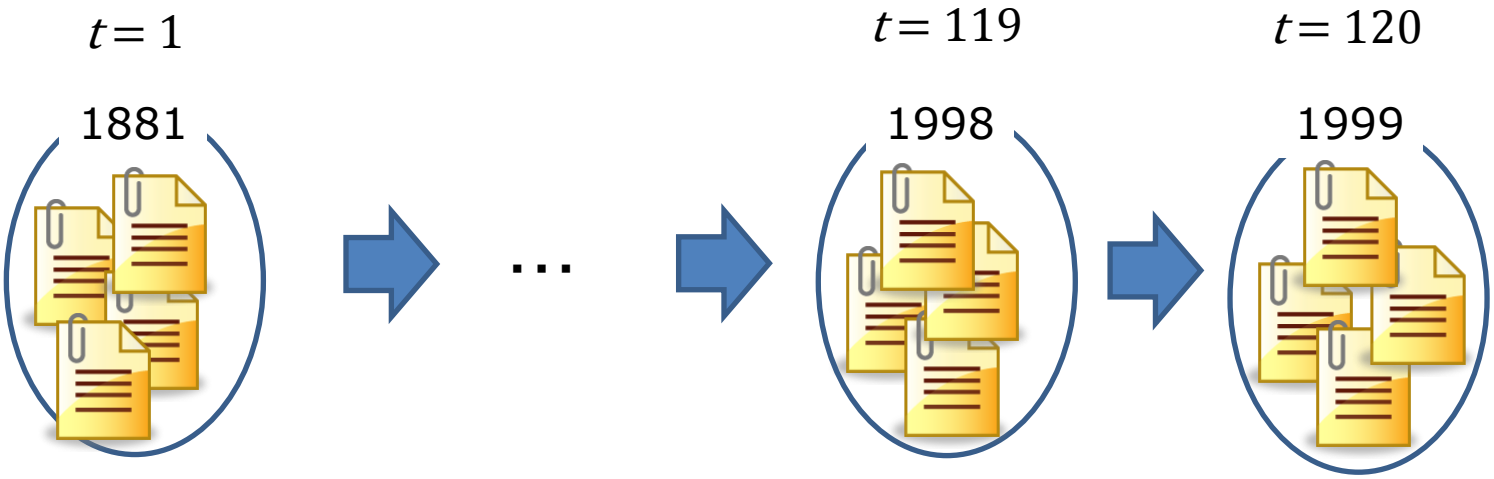


提案法: Dynamic Topic Models

- ☺ 非常に有名な時系列トピックモデルです
- 科学誌ScienceのOCRデータを用いて、科学論文の時系列トピック解析を行います
- topic proportionとtopic-word proportionに時間マルコフ性を入れたものです
- 推論は非常に難しいです

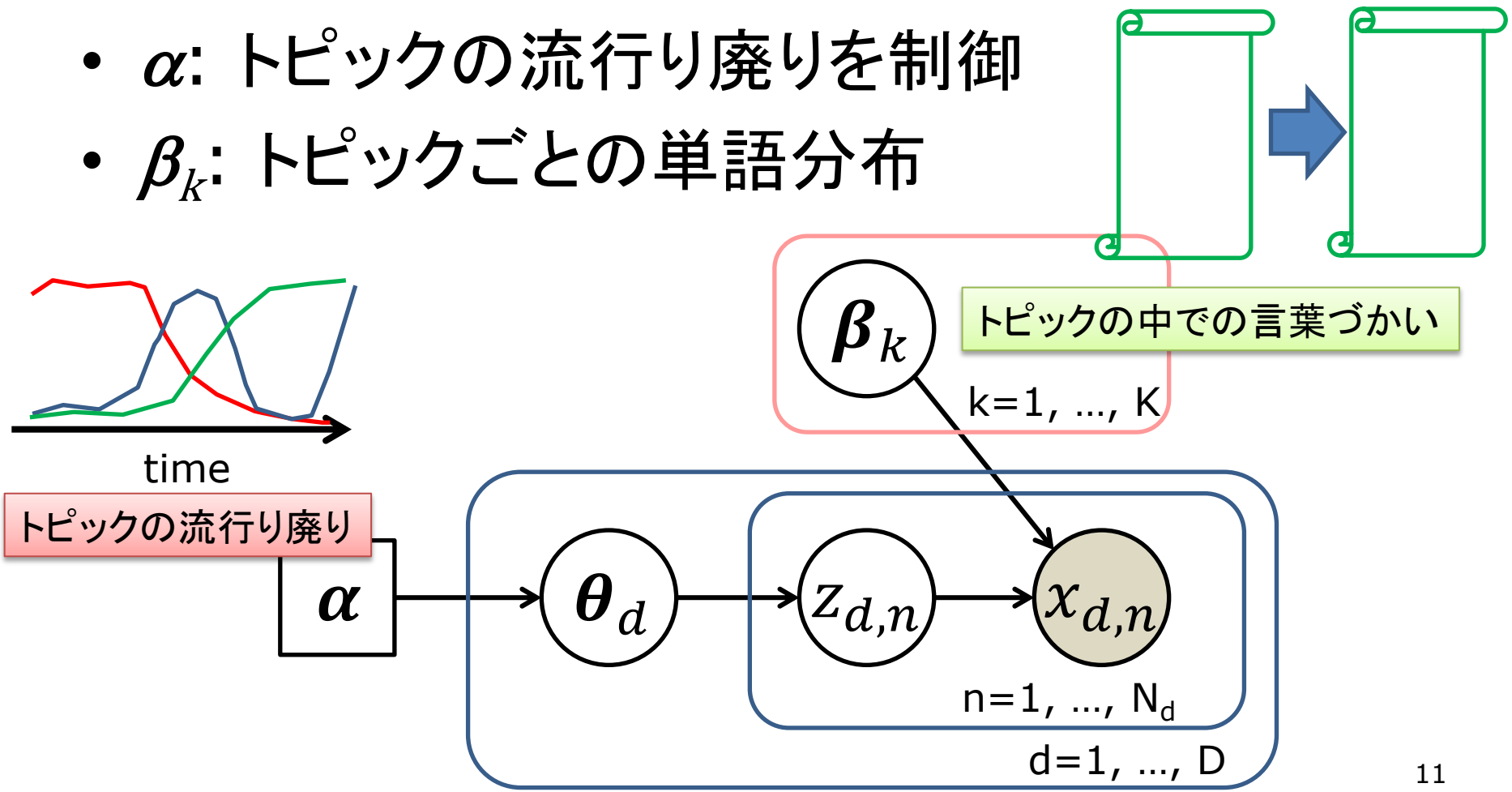
対象データ: Science誌

- 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)を利用して、発行年度ごとの文書時系列データを作成



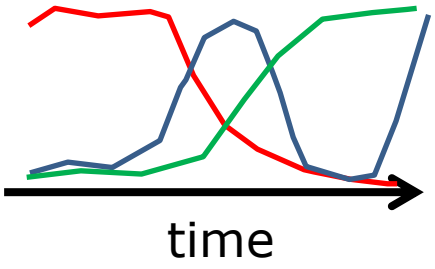
提案法のアイデア

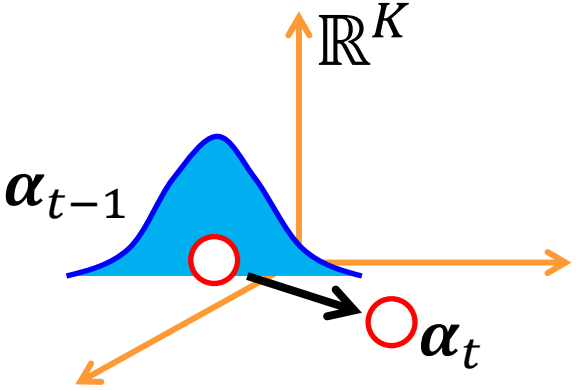
- 以下の2点を時間発展させます
- α: トピックの流行り廃りを制御
- β_k: トピックごとの単語分布

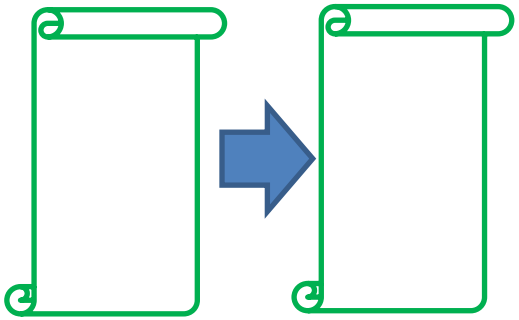


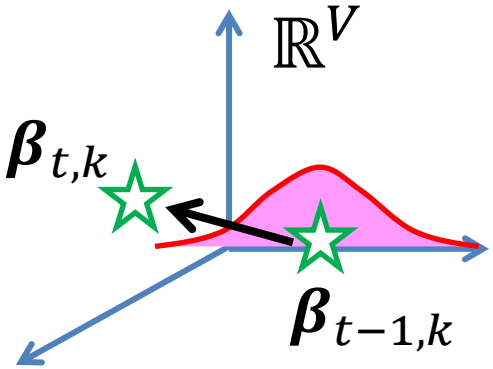
時間発展のモデル：正規ノイズでdrift

- 最も単純な時間発展モデルと言えます
- パラメータは前の時刻を中心に少しずつしか動かない、という想定です

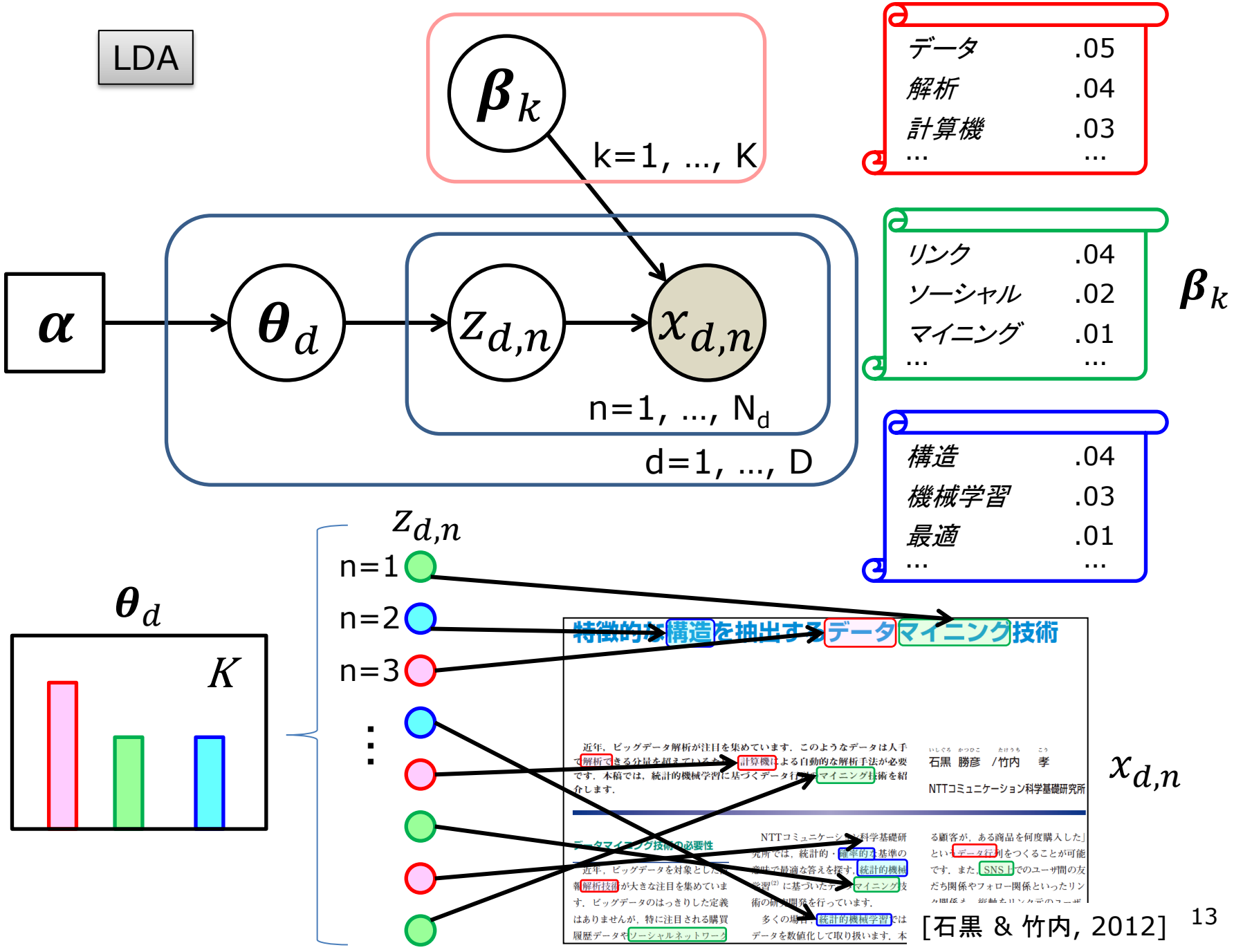






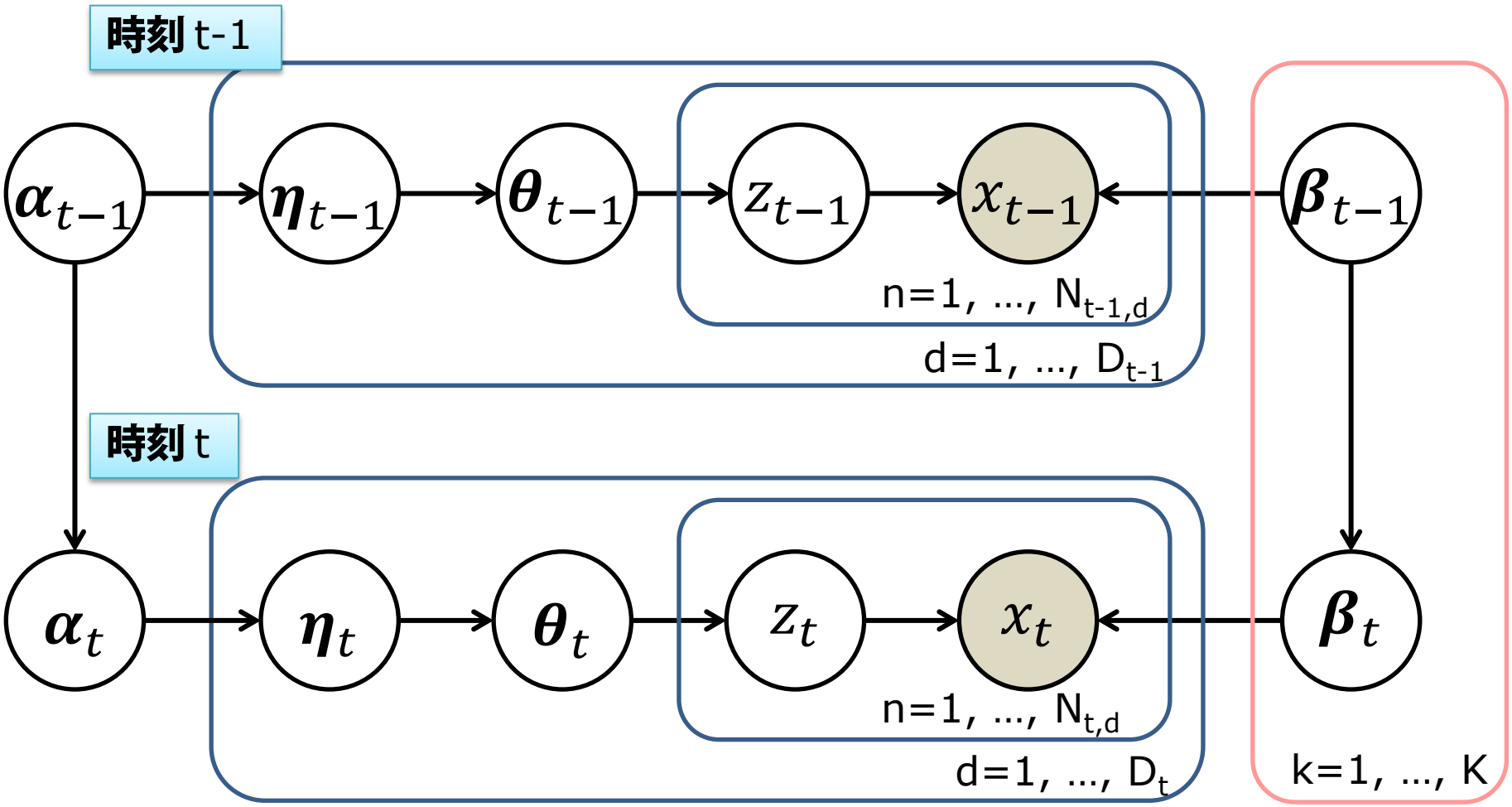


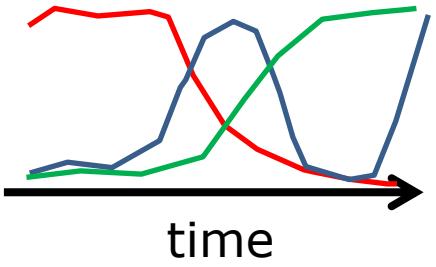
LDA

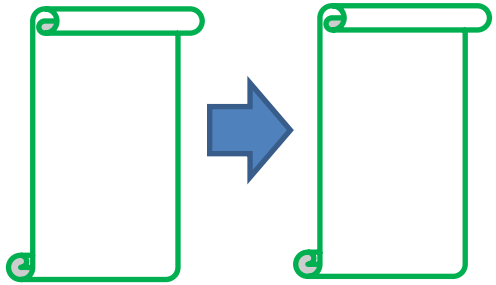


[石黒 & 竹内, 2012]

Dynamic Topic Model (添え字d,n,kは省略)







生成モデル

for 時間t = 1, 2, …, T

for theme (topic) k = 1, 2, …, K

topic-word proportion drift $\boldsymbol{\beta}_{t,k}|\boldsymbol{\beta}_{t-1,k}\sim\mathrm{N}\left(\boldsymbol{\beta}_{t-1,k},\sigma^2\boldsymbol{I}\right)$

topic proportion parameter drift $\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}\sim\mathrm{N}(\boldsymbol{\alpha}_{t-1},\delta^2\boldsymbol{I})$

for 文書d = 1, 2, …, D_t

topic proportion

for 単語n = 1, 2, …, $N_{t,d}$

topic-word assignment

word observation

for 時間t = 1, 2, …, T

$$\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}\sim\mathrm{N}(\boldsymbol{\alpha}_{t-1},\delta^2\boldsymbol{I}) \qquad \boldsymbol{\beta}_{t,k}|\boldsymbol{\beta}_{t-1,k}\sim\mathrm{N}(\boldsymbol{\beta}_{t-1,k},\sigma^2\boldsymbol{I})$$

for 文書d = 1, 2, …, D_t

topic proportion

$$\boldsymbol{\eta}_{t,d}|\boldsymbol{\alpha}_t\sim\mathrm{N}(\boldsymbol{\alpha}_t,a^2\boldsymbol{I})$$

$$\boldsymbol{\theta}_{t,d}|\boldsymbol{\eta}_{t,d}=\pi(\boldsymbol{\eta}_{t,d})$$

for 単語n = 1, 2, …, N_d

topic-word assignment

$$z_{t,d,n}|\boldsymbol{\theta}_{t,d}\sim\mathrm{Multinomial}(\boldsymbol{\theta}_{t,d})$$

word observation

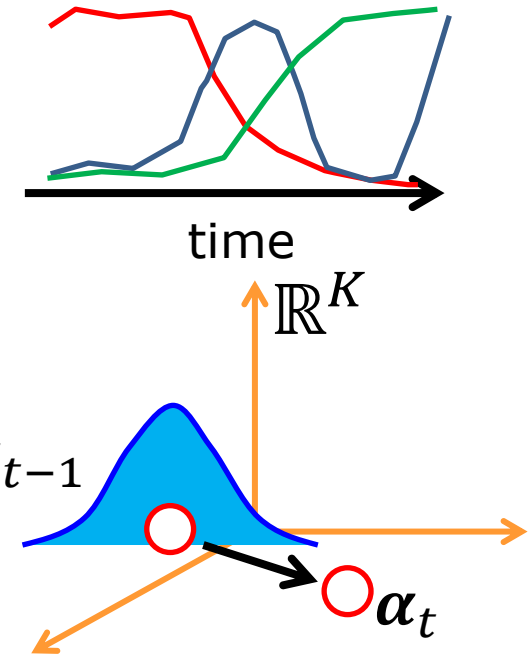
$$x_{d,n}|z_{d,n},\{\boldsymbol{\beta}_{t,k}\}\sim\mathrm{Multinomial}\left(\pi(\boldsymbol{\beta}_{t,z_{d,n}})\right)$$

πはsoft-max関数 $\pi(\boldsymbol{v})=\dfrac{\exp(v_k)}{\sum_l\exp(v_l)}$

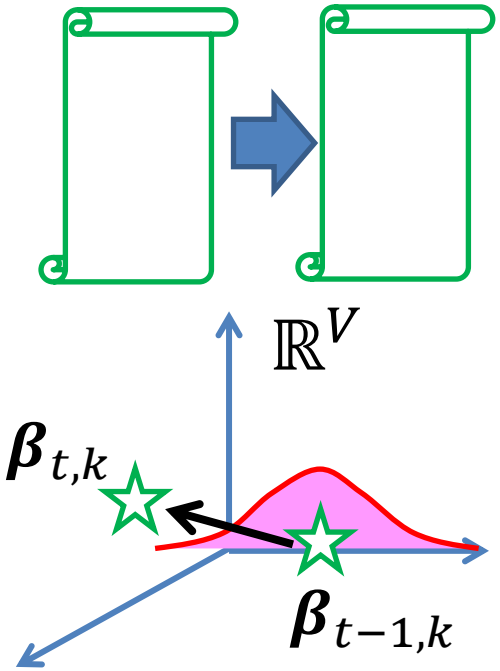
時間発展のモデル

- 正規分布を使って、1時刻のパラメータ遷移(drift)をモデル化します

$\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}\sim\mathrm{N}(\boldsymbol{\alpha}_{t-1},\delta^2\boldsymbol{I})$



$\boldsymbol{\beta}_{t,k}|\boldsymbol{\beta}_{t-1,k}\sim\mathrm{N}(\boldsymbol{\beta}_{t-1,k},\sigma^2\boldsymbol{I})$



トピックモデルへの適合

- 正規分布からは実数ベクトルが生成されるため、そのままでは多項分布(Multinomial)に使えません
- Soft-max関数を利用して変換します

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

$$\boldsymbol{\eta}_{t,d}|\boldsymbol{\alpha}_t \sim \mathrm{N}(\boldsymbol{\alpha}_t, a^2\boldsymbol{I})$$ 時刻t, 文書dのtopic proportion

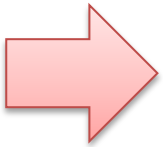
$$\boldsymbol{\theta}_{t,d}|\boldsymbol{\eta}_{t,d} = \pi(\boldsymbol{\eta}_{t,d})$$ Soft-max

$$z_{t,d,n}|\boldsymbol{\theta}_{t,d} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_{t,d})$$ topic-word assign.

$$x_{d,n}|z_{d,n}, \{\boldsymbol{\beta}_{t,k}\} \sim \mathrm{Multinomial}\left(\pi(\boldsymbol{\beta}_{t,z_{d,n}})\right)$$

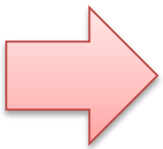
隠れ変数・パラメータの推定：非常に難しくなります

- 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**が不可能になるため、変分ベイズ法が候補になります**

- 原因2: 時刻 t が前時刻 t-1 に依存するため、時間依存性を考慮した推定が必要になります ☹



時間発展するパラメータは、時刻依存性を考慮して変分ベイズ法を構築する必要があります

時間発展パラメータの推定: 状態空間モデル解釈[北川, 2005]

- 連続なパラメータの時間変化を追いかける定番の手法です
- DTMの時間発展部分も状態空間モデルとして解釈できます

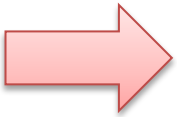
	一般の状態空間モデル	DTM(k, d, zなどを省略)
状態モデル	$y_t \mid y_{t-1} \sim f(y_{t-1}, \theta)$	$\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1} \sim \mathrm{N}(\boldsymbol{\beta}_{t-1}, \sigma^2 \boldsymbol{I})$
観測モデル	$x_t \mid y_t \sim g(x_t, \varphi)$	$x_{t,n} \mid \boldsymbol{\beta}_t \sim \mathrm{Mult}\left(\pi(\boldsymbol{\beta}_t)\right)$

変分近似によるKalman filter
[Kalman, 1960]

- 状態モデル、観測モデルの双方が正規分布の場合, Kalman Filterを用いてexactな解が計算できます
- 変分事後分布として、観測モデルに正規分布を“強引に”仮定して推論します

$$\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}\sim\mathrm{N}(\boldsymbol{\beta}_{t-1},\sigma^2\boldsymbol{I})$$

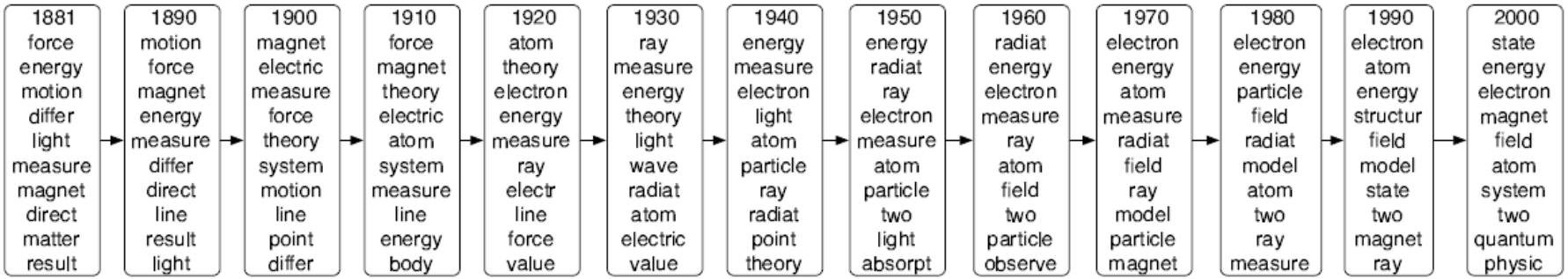
$$x_{t,n}|\boldsymbol{\beta}_t\sim\mathrm{Mult}\big(\pi(\boldsymbol{\beta}_t)\big)$$



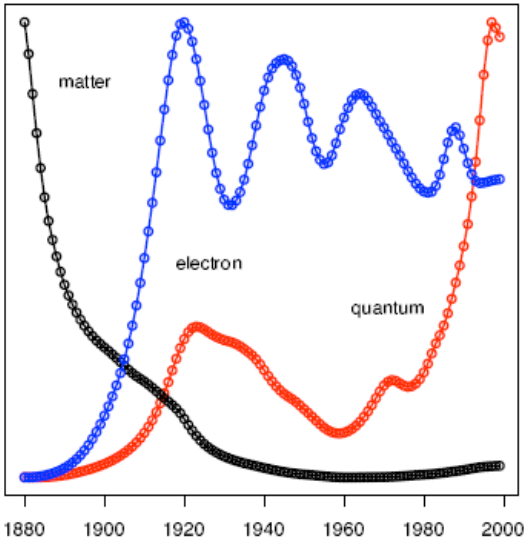
$$\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}\sim\mathrm{N}(\boldsymbol{\beta}_{t-1},\sigma^2\boldsymbol{I})$$

$$\widehat{\boldsymbol{\beta}}_t|\boldsymbol{\beta}_t\sim\mathrm{N}(\boldsymbol{\beta}_t,\hat{v}_t\boldsymbol{I})$$

変分観測量

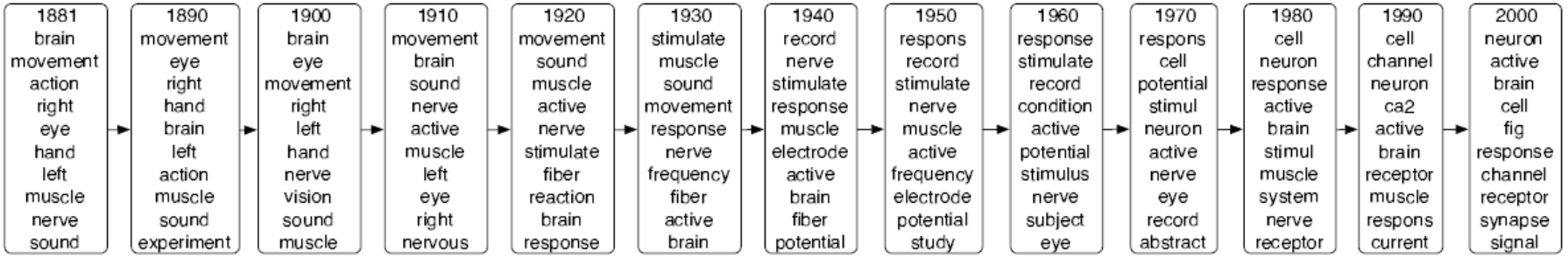


"Atomic Physics"

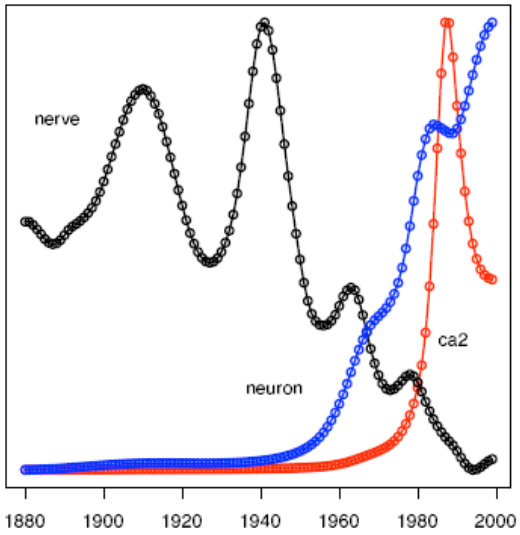


1881 On Matter as a form of Energy
1892 Non-Euclidean Geometry
1900 On Kathode Rays and Some Related Phenomena
1917 ``Keep Your Eye on the Ball"
1920 The Arrangement of Atoms in Some Common Metals
1933 Studies in Nuclear Physics
1943 Aristotle, Newton, Einstein. II
1950 Instrumentation for Radioactivity
1965 Lasers
1975 Particle Physics: Evidence for Magnetic Monopole Obtained
1985 Fermilab Tests its Antiproton Factory
1999 Quantum Computing with Electrons Floating on Liquid Helium

[Blei & Lafferty, 2006]

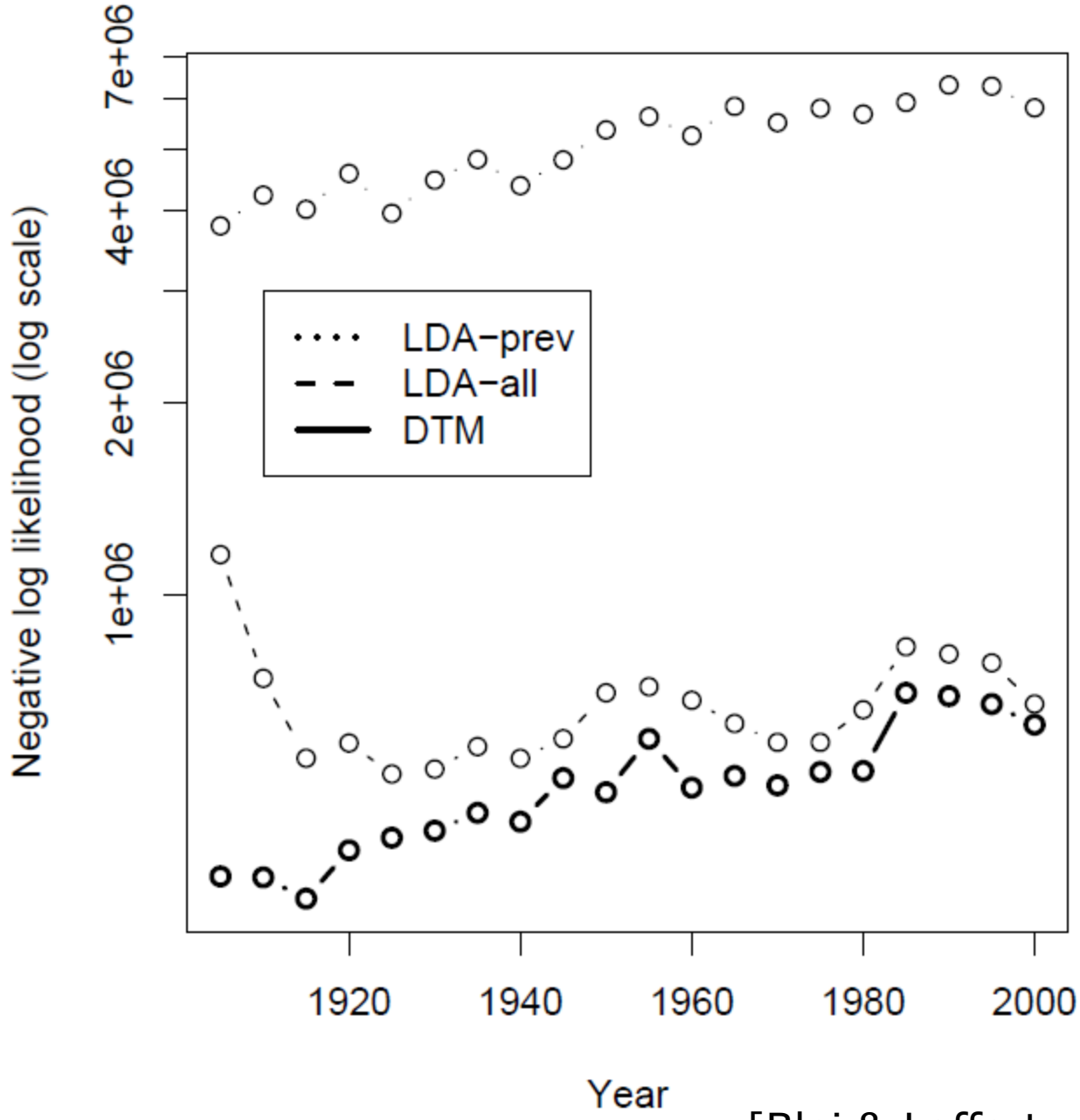


"Neuroscience"



1887 Mental Science
1900 Hemianopsia in Migraine
1912 A Defence of the ``New Phrenology''
1921 The Synchronal Flashing of Fireflies
1932 Myoesthesis and Imageless Thought
1943 Acetylcholine and the Physiology of the Nervous System
1952 Brain Waves and Unit Discharge in Cerebral Cortex
1963 Errorless Discrimination Learning in the Pigeon
1974 Temporal Summation of Light by a Vertebrate Visual Receptor
1983 Hysteresis in the Force-Calcium Relation in Muscle
1993 GABA-Activated Chloride Channels in Secretory Nerve Endings

[Blei & Lafferty, 2006]



[Blei & Lafferty, 2006]

まとめ: Dynamic Topic Models

- トピックごとの単語分布、トピックの割合の二つを時間発展させたトピックモデルです
- 正規分布によるdriftで時間遷移を表現します
- ☺ 非常に有名なので、時間モデルでは必ず押さえる必要がある論文です

Topic Tracking Model
[Iwata, 2009]

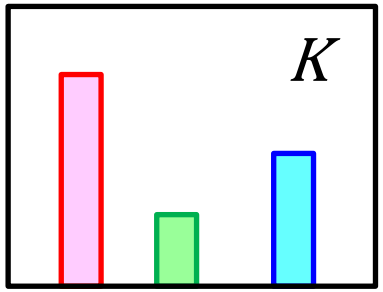
Iwata et al,
"Topic Tracking Model for Analyzing Consumer Purchase Behavior",
in Proc. IJCAI, 2009.

購買履歴データへの トピックモデル応用

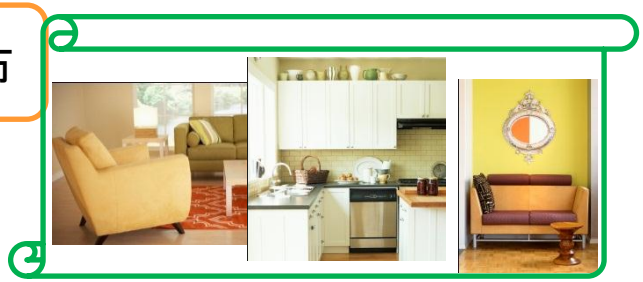
- PLSIなどのように、潜在変数モデルを使った購買履歴データのモデリングは多数存在します (e.g. [Jin, 2004])

- 当然、トピックモデルによる購買履歴データモデリングを考慮することもできます

購買履歴データの
ナイーブなトピック化



興味トピック k ごとのアイテム分布



ユーザ d の
興味トピック分布

$\boldsymbol{\beta}_k$

k=1, …, K

$\boldsymbol{\alpha}$

$\boldsymbol{\theta}_d$

$z_{d,n}$

$x_{d,n}$

n=1, …, N_d

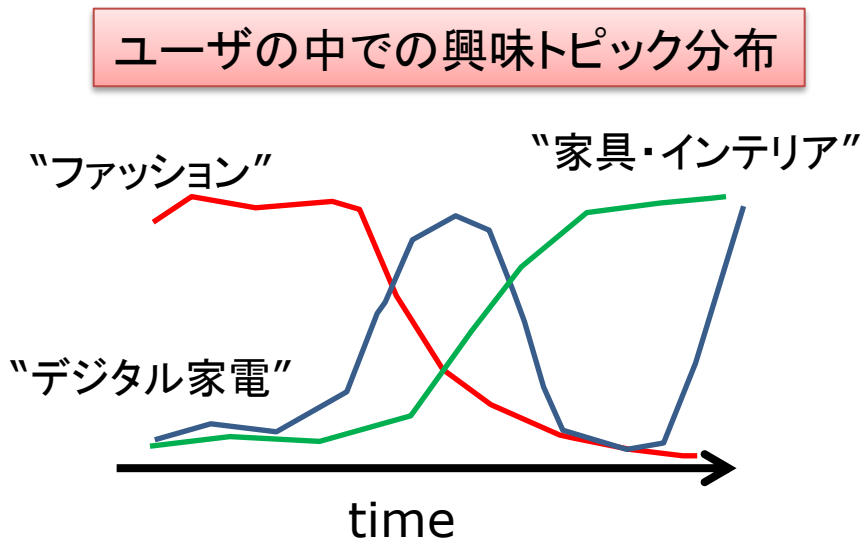
d=1, …, D

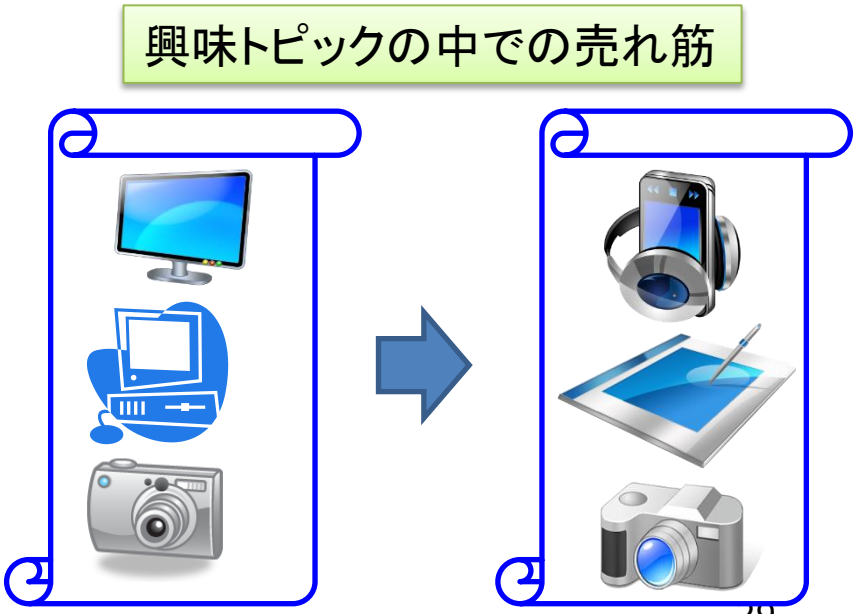
購買行動 n の潜在興味トピック

購買行動 n

注目する時間依存性：ユーザ＋トピック

- 1: ユーザの興味は少しづつ変わります
- 2: 興味トピックの中でのアイテムの売れ筋も変化します



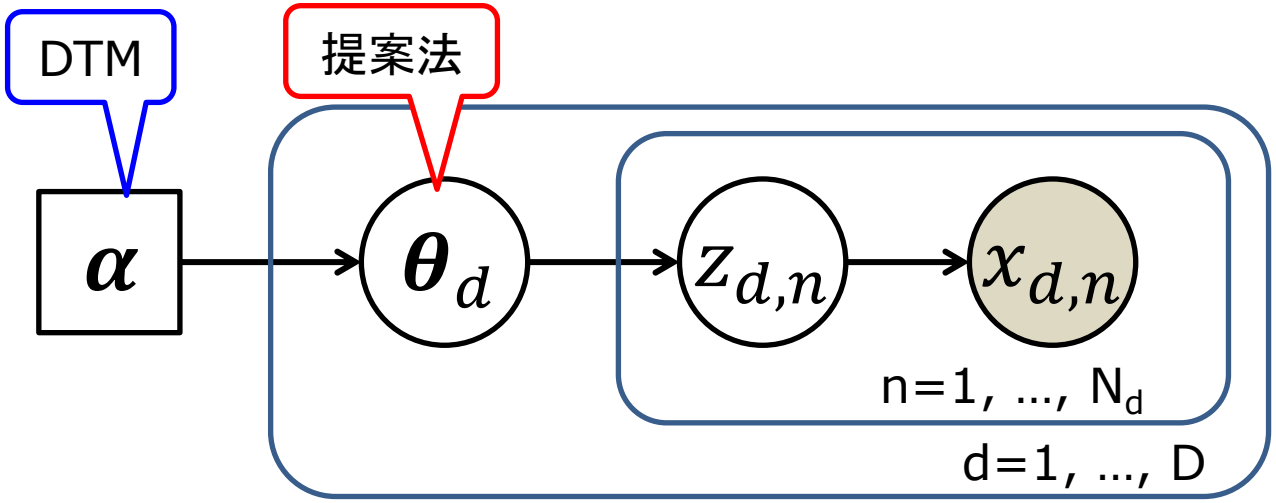


提案法: Topic Tracking Model

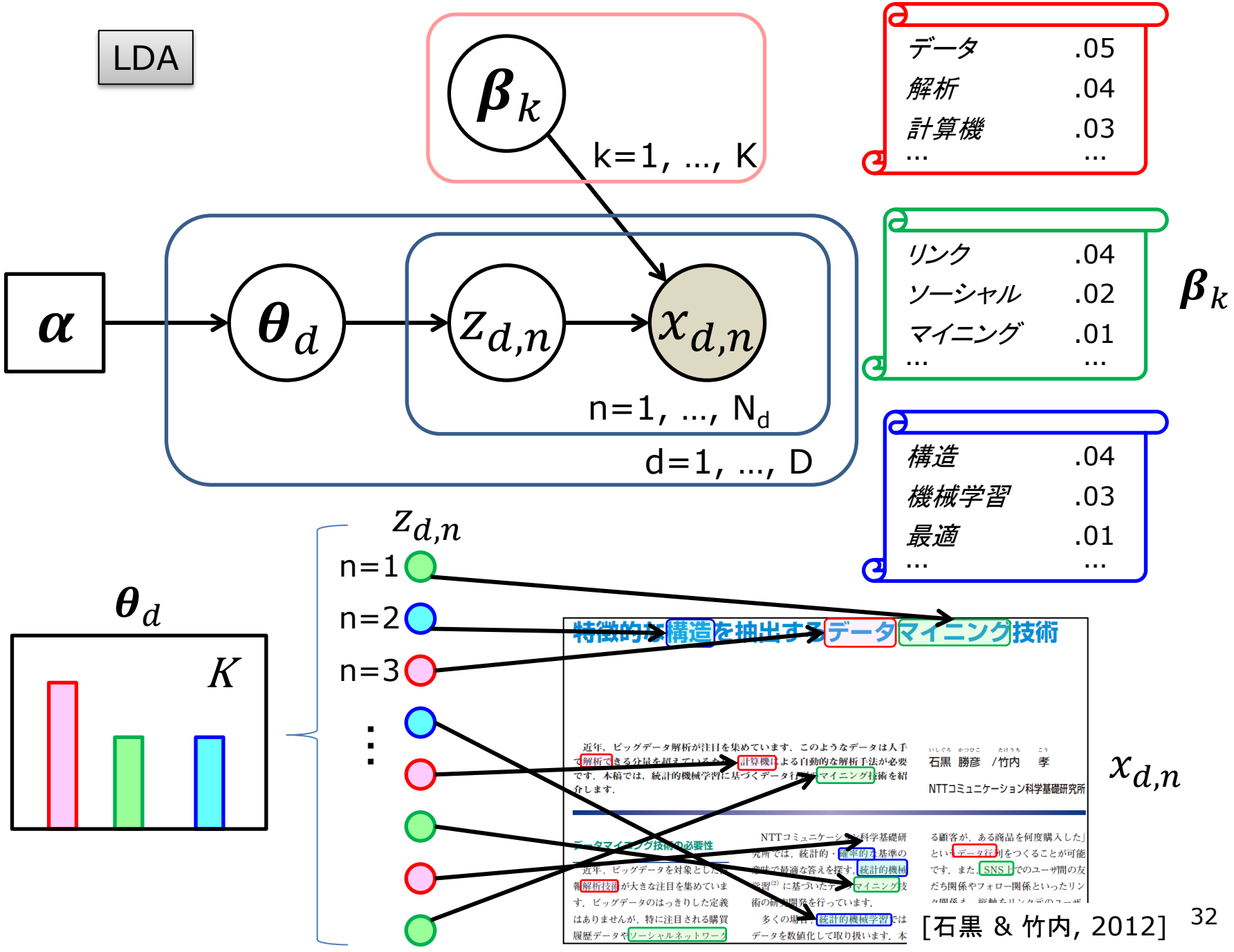
- Dynamic Topic Model(DTM)とはまた違う時系列トピックモデルです
- 文書（ユーザ）ごとのトピック分布と、トピックの単語（アイテム）分布が時間遷移します
- 推論はDTMに比べて少し簡単になるように工夫されています

提案法のアイデア: DTMとモデリングの観点が違います

- DTM: 各年度での話題の隆盛が知りたい → トピック分布制御パラメータ α を時間依存
- 提案法:ユーザの興味の変化が知りたい → 各ユーザ(文書)のトピック分布 θ を時間依存させる

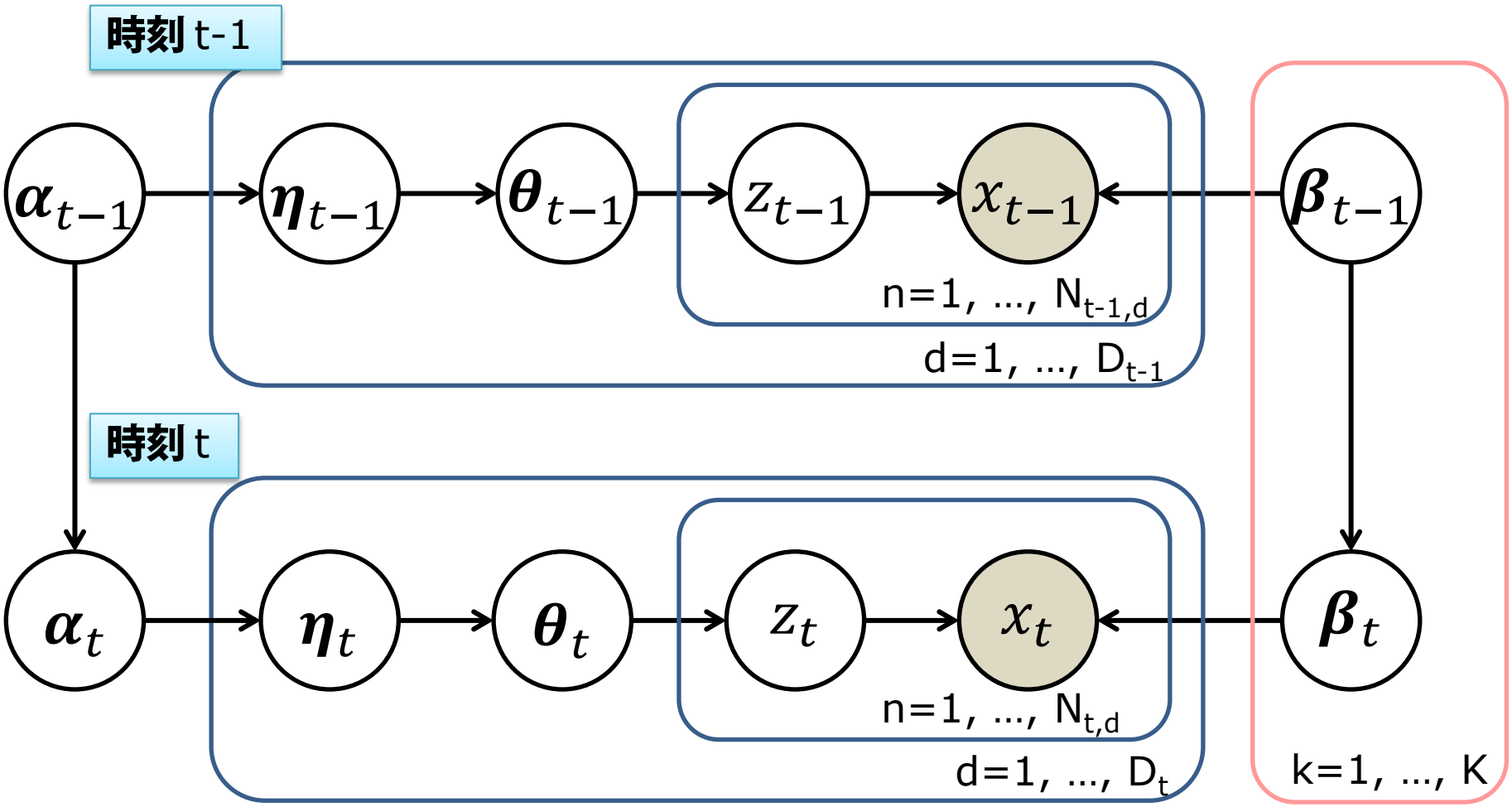


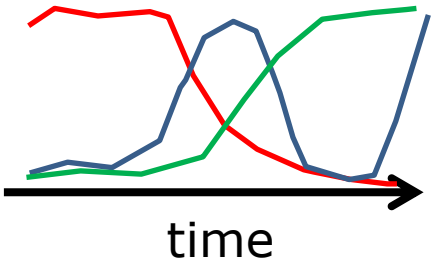
LDA

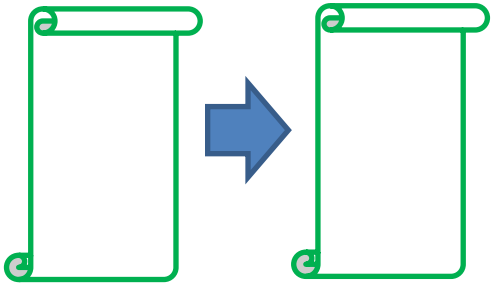


[石黒 & 竹内, 2012]

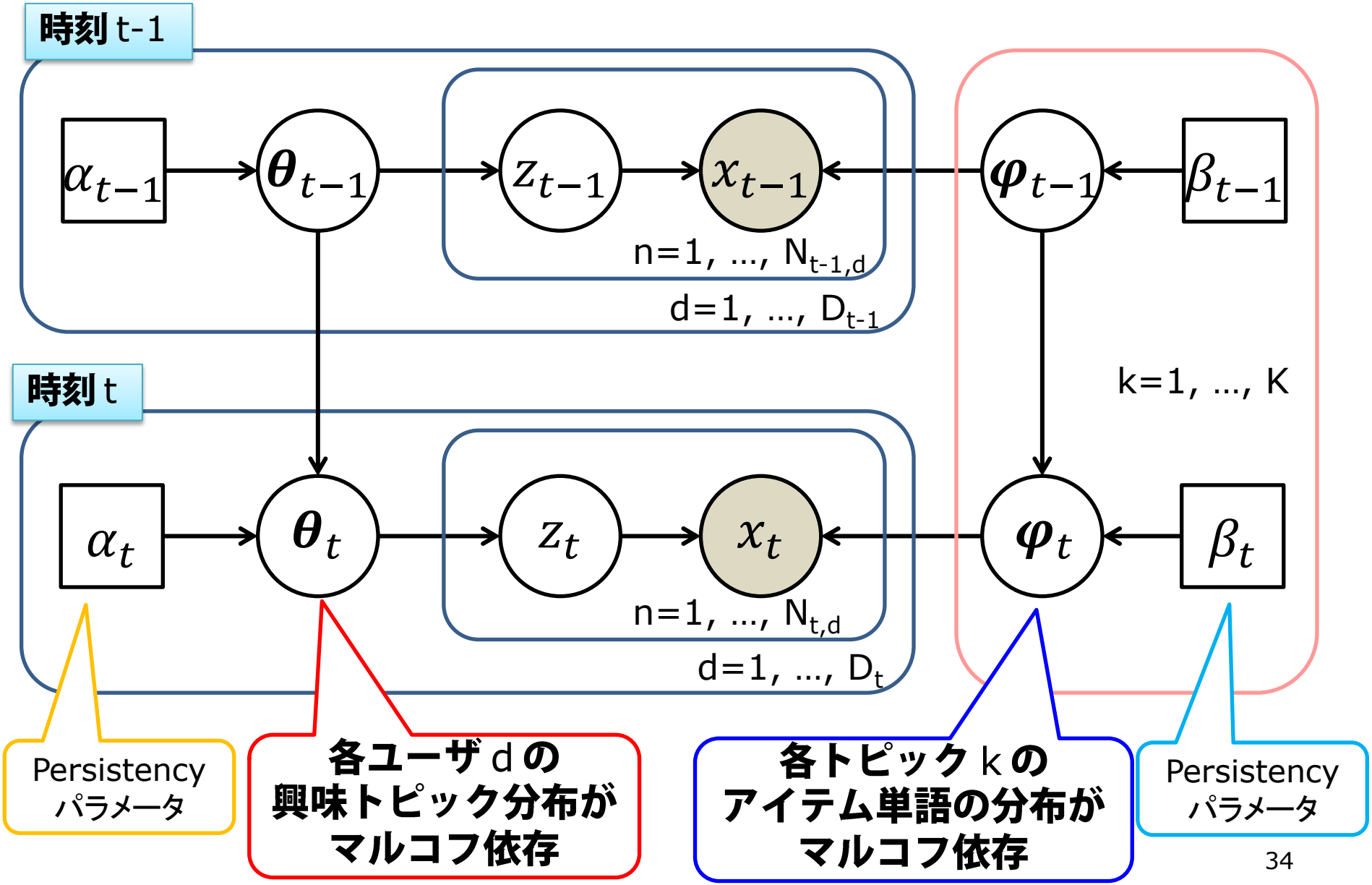
Dynamic Topic Model (添え字d,n,kは省略)







Topic Tracing Model （添え字d,n,kは省略）



生成モデル

for 時間t = 1, 2, …, T

for 興味topic k = 1, 2, …, K

topic-item word proportion parameter $\boldsymbol{\beta}_{t,k}$

for ユーザ d = 1, 2, …, D_t

topic proportion parameter t $\boldsymbol{\alpha}_{t,d}$

topic proportion

for 購買行動 n = 1, 2, …, $N_{t,d}$

topic-item word assignment

item word observation

for 時間t = 1, 2, …, T

for 興味topic k = 1, 2, …, K

topic-item word proportion evolution

$$\boldsymbol{\varphi}_{t,k}|\hat{\boldsymbol{\varphi}}_{t-1,k},\beta_{t,k}\sim\mathrm{Dir}\left(\beta_{t,k}\hat{\boldsymbol{\varphi}}_{t-1,k}\right)$$

for ユーザ d = 1, 2, …, D_t

topic proportion evolution

$$\boldsymbol{\theta}_{t,d}|\hat{\boldsymbol{\theta}}_{t-1,d},\alpha_{t,d}\sim\mathrm{Dir}\left(\alpha_{t,d}\hat{\boldsymbol{\theta}}_{t-1,d}\right)$$

for 購買行動 n = 1, 2, …, N_d

topic-item word assignment

$$z_{t,d,n}|\boldsymbol{\theta}_{t,d}\sim\mathrm{Mult}\left(\boldsymbol{\theta}_{t,d}\right)$$

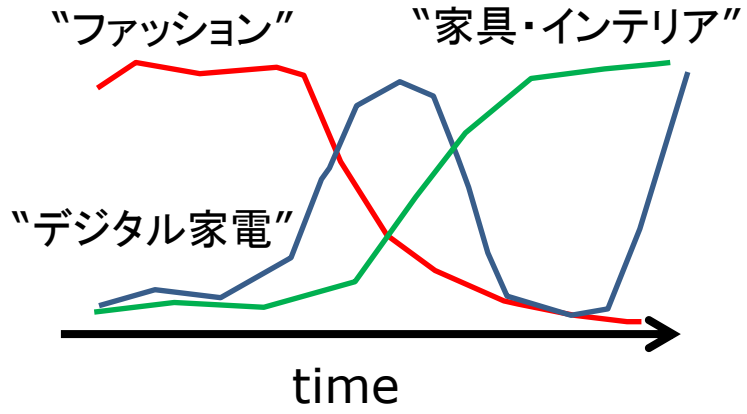
item word observation

$$x_{t,d,n}|z_{t,d,n},\left\{\boldsymbol{\varphi}_{t,k}\right\}\sim\mathrm{Mult}\left(\boldsymbol{\varphi}_{t,z_{t,d,n}}\right)$$

$\hat{\cdot}$ は”事後分布での期待値”を表す

興味トピック分布のモデル

- DTMと違い、ディリクレ分布を利用して時間発展をモデル化しています
- ユーザ、時間ごとに、興味トピックの持続度(persistency)もモデル化します



平均 $\hat{\theta}_{t-1,d}$ のディリクレ分布

$$\boldsymbol{\theta}_{t,d}|\hat{\boldsymbol{\theta}}_{t-1,d},\alpha_{t,d}\sim\mathrm{Dir}\left(\alpha_{t,d}\hat{\boldsymbol{\theta}}_{t-1,d}\right)$$

$\hat{\boldsymbol{\theta}}_{t-1,d}$ **: 前時刻の事後分布期待値**

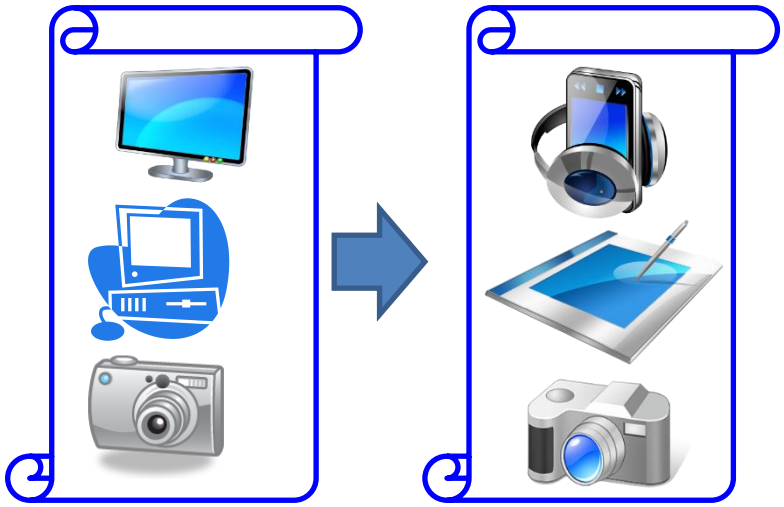
$\alpha_{t,d}$**: 持続度パラメータ**

α大 ＝ θ_tの分散小 → 小さな時間変化

α小 ＝ θ_tの分散大 → 大きな時間変化

トピック-アイテム（単語）分布のモデル

- 興味トピックと同様です



平均 $\hat{\varphi}_{t-1,d}$ のディリクレ分布

$$\boldsymbol{\varphi}_{t,k}|\hat{\boldsymbol{\varphi}}_{t-1,k},\beta_{t,k}\sim\mathrm{Dir}\left(\beta_{t,k}\hat{\boldsymbol{\varphi}}_{t-1,k}\right)$$

$\hat{\varphi}_{t-1,d}$：前時刻の事後分布期待値

$\beta_{t,d}$：持続度パラメータ

β大 ＝ ϕ_tの分散小→ 小さな時間変化

β小 ＝ ϕ_tの分散大 → 大きな時間変化

長期時間依存モデル

- 1時刻前に依存するだけでなく、数ステップ前までに依存する形への拡張も簡単です

1ステップ前からの依存関係モデル

$$\boldsymbol{\varphi}_{t,k}|\widehat{\boldsymbol{\varphi}}_{t-1,k},\beta_{t,k}\sim\mathrm{Dir}\left(\beta_{t,k}\widehat{\boldsymbol{\varphi}}_{t-1,k}\right)$$

$$\boldsymbol{\theta}_{t,d}|\widehat{\boldsymbol{\theta}}_{t-1,d},\alpha_{t,d}\sim\mathrm{Dir}\left(\alpha_{t,d}\widehat{\boldsymbol{\theta}}_{t-1,d}\right)$$

Lステップ前からの依存関係モデル

$$\boldsymbol{\varphi}_{t,k}|\widehat{\boldsymbol{\varphi}}_{t-1,k},\beta_{t,k}\sim\mathrm{Dir}\left(\sum_{l=1}^{L}\beta_{t,k,l}\widehat{\boldsymbol{\varphi}}_{t-l,k}\right)$$

$$\boldsymbol{\theta}_{t,d}|\widehat{\boldsymbol{\theta}}_{t-1,d},\alpha_{t,d}\sim\mathrm{Dir}\left(\sum_{l=1}^{L}\alpha_{t,d,l}\widehat{\boldsymbol{\theta}}_{t-l,d}\right)$$

隠れ変数・パラメータの推定

- ☺ 非常に簡単な逐次推定アルゴリズムが導出できます
  - 正規分布やsoft-maxがないため！！
  - LDAのGibbs, VB (EM) を導出したことがある方にとっては自明な解が得られます
- ☹ ただし、DTMのように系列としての最適解は得られません

Table 1: Average N-best accuracies (%) over time. The digit in the bracket is the standard deviation.

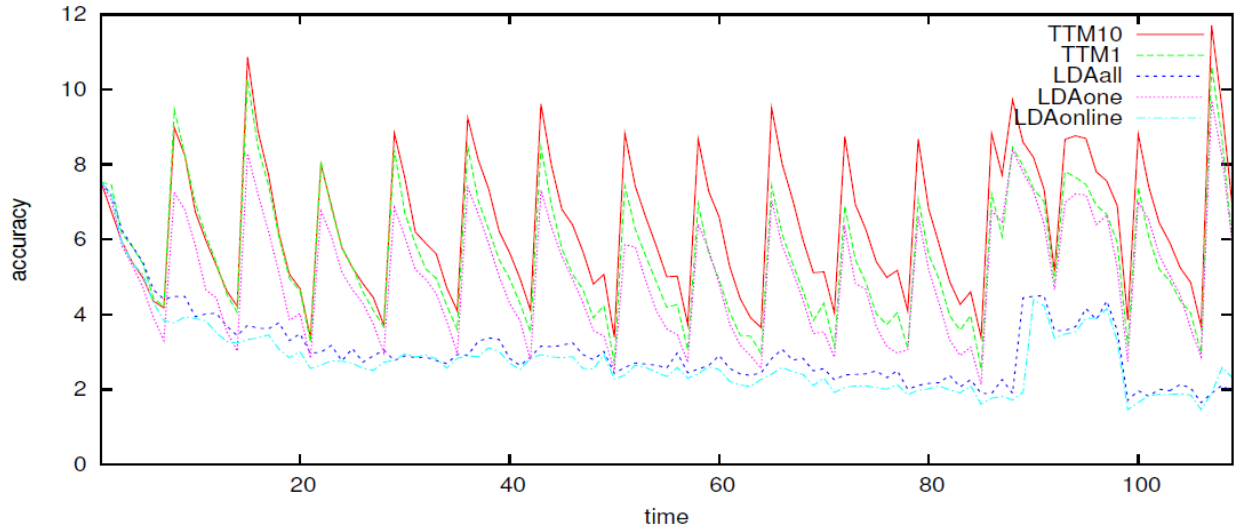
(a) movie

N	LDAall	LDAonline	LDAone	TTM1	TTM10
1	1.21 (0.61)	1.08 (0.54)	1.91 (0.78)	2.22 (0.91)	**2.46** (0.92)
2	2.18 (0.79)	2.00 (0.78)	3.52 (1.22)	3.99 (1.33)	**4.47** (1.36)
3	3.06 (1.04)	2.81 (1.02)	5.04 (1.64)	5.60 (1.75)	**6.35** (1.85)
4	3.90 (1.27)	3.56 (1.24)	6.24 (1.90)	6.82 (2.01)	**7.82** (2.15)
5	4.70 (1.51)	4.26 (1.44)	7.37 (2.20)	7.92 (2.26)	**9.20** (2.42)

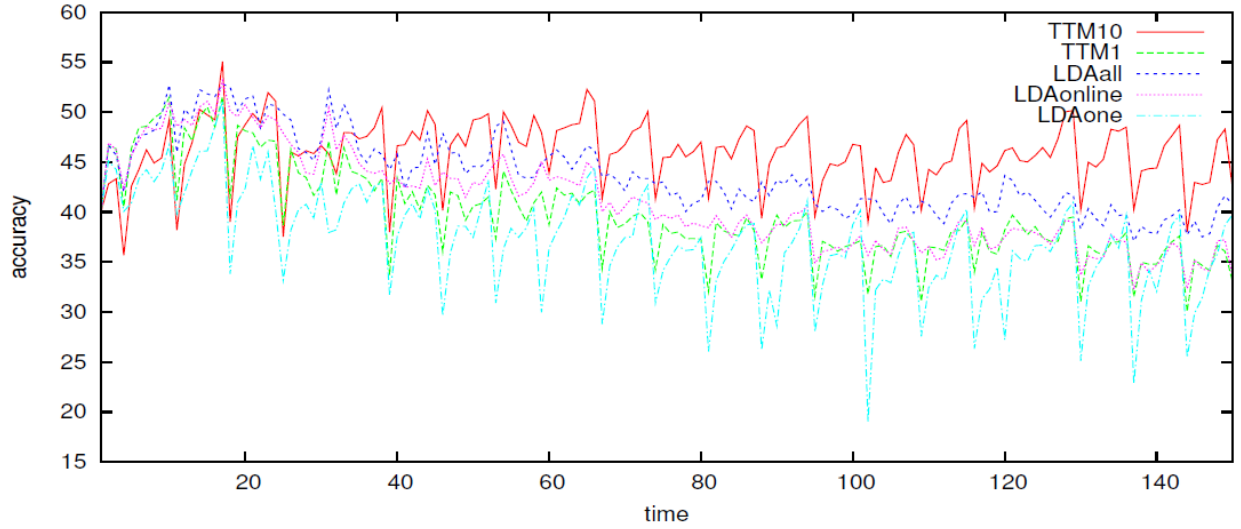
(b) cartoon

N	LDAall	LDAonline	LDAone	TTM1	TTM10
1	27.0 (3.3)	26.0 (3.5)	24.8 (4.5)	26.8 (4.2)	**30.5** (3.4)
2	37.3 (3.6)	35.1 (4.2)	32.4 (4.9)	34.2 (4.5)	**39.9** (3.5)
3	43.7 (3.9)	41.1 (4.8)	37.2 (5.3)	39.8 (4.6)	**45.9** (3.3)
4	48.5 (4.0)	45.8 (5.1)	40.9 (5.3)	44.5 (4.6)	**50.6** (3.2)
5	52.4 (4.2)	49.6 (5.4)	44.1 (5.4)	48.5 (4.6)	**54.4** (3.0)

[Iwata, 2009]



(a) movie



(b) cartoon

[Iwata, 2009]

Figure 3: Three-best accuracies (%) for each day.

まとめ: Topic Tracking model

- ユーザ(文書)ごとのトピック分布、トピックの単語分布を時間発展させたトピックモデル

- Dirichletで時間遷移を表現したことで、非常に簡単に解を導出できます

その他の時系列データ応用

- Wang and McCallum, “Topics over Time: A Non-Markov Continuous-Time model of Topical Trends”, in Proc. KDD, 2006.
- Iwata et al., “Sequential Modeling of Topic Dynamics with Multiple Timescales”, ACM Trans. on Knowledge Discovery from Data. Vol. 5(4). pp. 19:1-19:27, 2012.
- Pruteanu-Malinici, et al., “Hierarchical Bayesian Modeling of Topics in Time-Stamped Documents”, IEEE Trans. PAMI, Vol. 32(6), pp.996-1011, 2010.

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, 2006], Blei and Lafferty, “Dynamic Topic Models”, in Proc. ICML, 2006.
- [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.
- [北川, 2005] 北川, “時系列解析入門”, 岩波書店, 2005.
- [Kalman, 1960] Kalman, “A New Approach to Linear Filtering and Prediction Problems”, Journal of Basic Engineering, 1960.
- [Iwata, 2009] Iwata et al, “Topic Tracking Model for Analyzing Consumer Purchase Behavior”, Proc. in IJCAI, 2009.
- [Jin, 2004] Jin et al, “Web Usage Mining based on Probabilistic Latent Semantic Anlysis”, Proc. in KDD, 2004.