混合ユニグラムモデルにおける確率分布関数及びギブズ・サンプリングの可視化教材

白田　由香利[†]　　橋本　隆子[‡]

†学習院大学経済学部経営学科　〒177-0855 東京都豊島区目白 1-5-1
‡千葉商科大学商経学部　〒272-8512 千葉県市川市国府台 1-3-1

E-mail:　†yukari.shirota(AT)gakushuin.ac.jp, ‡takako(AT)cuc.ac.jp

あらまし　本稿では，混合ユニグラムモデルにおける確率分布関数及びギブス・サンプリングの可視化教材ツールを作ったので報告する．トピック抽出の分野で潜在的ディリクレ配分 Latent Dirichlet Allocation (LDA)モデルによるギブズ・サンプリングアルゴリズムによるモンテカルロ法は広く使われているが，その数学的プロセスの意味を理解することは容易ではない．我々の行った可視化では，説明を簡単化するために，トピックモデルに代わり，その簡略化モデルである混合ユニモデルを用いた．この可視化により，「１つの文書だけを除いて，それ以外の文書の確率変数を全部固定すると，その１つの文書が，どのトピックに属すべきか分かる」というギブス・サンプリングの特長が容易に理解可能となる．

キーワード　ギブズ・サンプリング，ベイズ推論，可視化，混合ユニグラムモデル，潜在的ディリクレ配分モデル

1. 始めに

本稿では，混合ユニグラムモデルにおける確率分布関数及びギブス・サンプリングの可視化について報告する．マルコフ連鎖モンテカルロ法(MCMC と以下略す)とは，与えられた確率分布を不変分布にするようにデザインされた操作で，システムの状態を確率的に変化させていくことで，その分布からのサンプルを作り出すアルゴリズムである[1]．物理学では，動的なモンテカルロ法と呼ぶ．

MCMC の応用分野にテキストマイニングのトピック抽出がある．大量の文書からトピックを抽出するために，トピックモデルが広く使われているが，トピックモデルでは，文書を出現単語の多重集合(バッグ，bag)で表し，単語の並びに関する情報は廃し，文書を BOW(bag-of-words)で表現する[2]．トピックモデルは文書のための確率モデルである．そして，トピック分布にディリクレ事前分布を仮定し，ベイズ推論を行う手法を，潜在的ディリクレ配分 Latent Dirichlet Allocation (LDA)モデルと呼ぶ[3]．

この数学的考え方を可視化によって容易に理解できるようにすることが本研究の目的である．本稿では，トピックモデルを単純化した混合ユニグラムモデルを使う．トピックモデル，あるいは，混合ユニグラムモデルをベイズ推定する際，ギブズ・サンプリングという MCMC のアルゴリズムが広く使われている[4, 5]．ギブズ・サンプリングでは，近似するのではない．(例えば，変分ベイズ法は近似である)．真の事後分布からサンプリングできるので，原理的には，無限個の事例をサンプリングすることにより真の事後分布を求めることが可能となる[2]．

我々は，日本銀行金融政策決定会議議事録などの文書に対して，LDA モデル上でギブズ・サンプリングによってベイズ推論を行う，という手法を用いて，多くのトピック抽出を行ってきた[6-9]．このツールは，CRAN が公開している R の LDA パッケージ[1]を使っている．一般にベイズ推論に言えることは，ツールは提供されているので使うことは容易である．しかし，その数学的プロセスを理解することが難しい，ということである．その問題を解決しないことには，真にツールを使いこなしているとは言えないであろう．そこで，我々はトピックモデルの動作を可視化することで，従来理解できなかったユーザも，その数学的プロセスを理解できるようになることを目指して可視化ツールの作成を行った．

本可視化によって，「１つの文書だけを除いて，それ以外の文書の確率変数を全部固定すると，その１つの文書が，(混合ユニグラムモデルでは)どのトピックに属すべきか分かる」というギブス・サンプリングの特

[1] https://cran.r-project.org/web/packages/lda/index.html

長が容易に理解可能となる．

次節では，混合ユニモデルを説明する．第 3 節では，ギブズ・サンプリングのアルゴリズムの本質を説明する．第 4 節で，我々が開発した可視化ツールを説明する．最終節はまとめである．

2. 混合ユニグラムモデル

本節では，混合ユニグラムモデルを説明する．岩田は，3 種類の文書の確率モデルとして，ユニグラムモデル，混合ユニグラムモデル，トピックモデルの順に，各モデル上でのトピック抽出アルゴリズムを説明している[2]．本節では，そのモデル間の違いをグラフィカルモデル[3]により説明する(図 1 参照)．各モデル等の詳細については[2]を参照して頂きたい．

直線的に描いたグラフィカルモデル[2]と比較して，図 1 のように，単語系とトピック系で分けてレイアウトする方が，直観的理解を得やすくなるであろう．

単語の分布に関しては，ユニグラムモデルは，単語の分布 ϕ が N や D の依存の枠の外に出ていることから分かるように，全ての文書の全ての単語に関する分布は一つしかない．混合ユニグラムモデルでは，トピックごとに異なる単語分布を持つ．トピックモデルも同様に，トピックごとに異なる単語分布をもつ．

トピック分布に関しては，そもそもユニグラムモデルには，トピックの概念は無い．混合ユニグラムでは，文書集合全体として一つのトピック分布をもつだけである(θ が枠外にある)．これに対し，トピックモデルでは，文書ごとに異なるトピック分布をもつことが可能である．よって，混合ユニグラムモデルでは，文書によるトピック分布の違いはないので，文書内の単語の分布によって，その文書のトピック分布が決まり，最も高い確率であるトピック番号が選ばれる．それに対し，トピックモデルでは，1 つの文書が複数のトピックを持つことが可能であり，同じ語でも異なるトピックのもとに生成されることがある．

このトピックモデルによる文書集合の生成における変数の依存関係を理解するには，岩田の解説図が参考になる[2]．理由は，グラフィカルモデルの変数間の依存関係を具体例を使って説明しているからである．しかし，この岩田による文書集合生成プロセスの可視化と，本論文の目指す可視化は目的が異なる．我々の目指すものは，生成プロセスの可視化ではなく，ギブス・サンプリングの数学プロセスの理解を目的とする可視化である．また，3 次元アニメーションを使って対話的に学習者が操作可能とすることで，理解を深めるという特長もある．

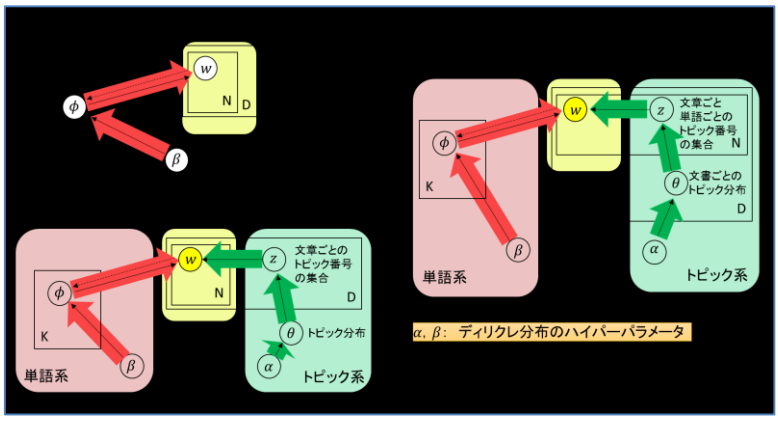

図 1：ユニグラムモデル，混合ユニグラムモデル，トピックモデルの比較

3. ギブズ・サンプリングの原理

本節では，ギブズ・サンプリングの原理，及びアルゴリズムの特長を説明する．

可視化する際には，その可視化において何を伝えたいのか方針を決めることが重要である．この可視化で表したいことは，「調べたい分布 P(x)が，条件付き確率に基づく置き換え操作で決まるマルコフ連鎖の不変分布になっている」という考え方である．置き換えにより状態が $x \to x'$ に変わるときの数学プロセスは以下のように表せる[1]．数式中の，色づけ及び，抜けている i 番目の要素の前後に間隔を入れるなどの数式レイアウトの工夫は，我々が行った．このほうが視覚的に理解しやすいと考えるからである．

$$\sum_{x_i}\{P(x_i'|x_1,\cdots,x_{i-1},\ x_{i+1},\cdots,x_N)\times \underline{P(x_1,\cdots,x_{i-1},x_i,x_{i+1},\cdots,x_N)}\}$$

$$= P(x_i'|x_1,\cdots,x_{i-1},\ x_{i+1},\cdots,x_N) \times P(x_1,\cdots,x_{i-1},\ x_{i+1},\cdots,x_N)$$

$$= \underline{P(x_1,\cdots,x_{i-1},x_i',x_{i+1},\cdots,x_N)}$$

ギブス・サンプリングの本質は，上式の 1 行目から 2 行目の変換で表される．周囲の文書の情報だけを用いて，つまり，当該の x_i の情報は使わずに，計算が行われる，ということである．

数学の概念を教える際に，どのような数式表現を選択するかということも重要な要素である．同じ内容でも数式表現によって分かり易いものとそうでないものがあるが(例えば，中心極限定理など定理の書き方はテキストごとに様々に異なっている)，伊庭による上記の数式表現[1]はギブズ．サンプリングの特長を

顕著に分かり易く表している．

我々は，可視化ツールによって「定常状態では，置き換え操作により，状態は移動するが，状態が分布する様子は全体として変わらない」ことをアニメーションにより表現したかった．十分な回数置き換え操作を行った後，状態は振動を起こす．振動の周期などはケースごとに違うと考えられる．ギブズ・サンプリングにおいて多くの場合，置き換え作業が十分行われたか否かは経験的に判断されている．定常状態になったことが可視化によって判別可能かというテーマについても今後研究していきたいと考える．

ギブズ･サンプリングで使う条件付き確率は一般的に以下のように表されるが，

$$P(x_i | x_1, \cdots, x_{i-1}, \quad x_{i+1}, \cdots, x_N)$$

混合ユニグラムモデルでは，$P(z_d = k \,|W, z_{\backslash d}, \alpha, \beta)$ となる．記号$z_{\backslash d}$ は，z からz_d のみを除いたトピック番号の集合を表す．

4. ギブズ・サンプリングの可視化教材

本節では，試作したギブズ・サンプリングの可視化教材について説明する．

本可視化ツールでは，文書モデルとして，トピックモデルを簡単化した混合ユニグラムモデルを使っている．これはギブズ・サンプリングの特長を分かり易く可視化表現するためには，複雑なトピックモデルよりも，混合ユニグラムモデルのほうが適していると考えたからである．トピックモデルでは，可視化表現が複雑になってしまい，本質が見えにくくなるからである．

可視化ツールは Wolfram 社の Mathematica を使ってプログラムした．Mathematica のプログラムは，CDF 形式に変換することで，フリーソフトウェアの Wolfram CDF player[2]で操作可能であり，WEB 公開する場合に利便性が高い．混合ユニグラムモデルによるギブズ・サンプリングアルゴリズムは岩田の記したものに従って Mathematica でプログラムした[2].

図 2 に可視化教材を示した．図 2 に示したケースでは，文書数 5，トピック数は 7，単語数は 6 としてある．最も強調したい点は，ギブズ・サンプリングで各文書に順番に置き換えをしている点である．そのため，同心円状に各文書を配置し，順番を示す矢印記号が，次の順番の文書を指し示すようにデザインした．垂直軸方向に白い球が配置されているが，白い球の高さでその文書のトピック番号を示してある．置き換え直後の文書の球は，オレンジ色にしてある．

円の中心から伸びている半径の直線上には，その文書のトピック確率分布を示した．

円の中心から外側に向かい，トピック#1,2,3,...,7 と番号付けしてあるので，図 2 では，トピック#3 が最大確率となっている．そして，そのトピック番号が#3 となっている．そして，そのトピック番号は，オレンジ色の球の高さとして表現される．

トピック分布 θ は，図 2 中左下のヒストグラムで表した．横軸がトピック番号である．その右横のグラフに DOCUMENT ID ごとのトピック番号の確率分布を示した．ここでは当該の文書に関する確率分布は，分かり易いように折れ線でつないで表示した．

下段，右端のグラフは TOPIC ID～WORD ID の確率分布を示している．ギブズ・サンプリングのアルゴリズムにおける単語分布の計算では，まず，当該の順番の文書のもつ単語を，全体のカウント対象からはずす．そして，その文書のトピック番号が決まると，その文書は，そのトピック番号の単語分布に加算する．つまり，当該文書は，自分の単語分布に一番類似する単語分布パターンをもつトピックを探し，そのトピックのメンバーになる．

詳細には，他の乗数として$(D_{k\backslash d} + \alpha)$があり，割り当てられた文書数が多いトピックの方が，確率が大きくなるという調整が施される．アルゴリズムの詳細は[2]を参照して頂きたい．

本可視化ツールでは，ギブズ・サンプリングの置き換え操作は，最上部のスライダーを動かすことで行う．このスライダー操作は逆向きの動きも可能である．

本可視化ツールをデータ工学関連の研究者，学生に見てもらったところ，アルゴリズムの理解が容易になる，と好評であった．従来のギブズ・サンプリングの可視化としては，MacKey の 2 変数の確率分布が与えられて，2 変数を交互に更新しあう図[10]や，Bishop や伊庭らによる相関のあるガウス分布に従う２つの変数を交互に更新するギブズ・サンプリングの図解などが存在する[3, 5]．しかし，いずれも 2 変数で交互に置き換えをするだけであるので，実際の LDA モデルを使ったギブズ・サンプリングを理解するためには，十分ではなかった．

今回，多数の文書及びトピック，単語の変数に関してギブズ・サンプリングの可視化を行ったので，ギブズ・サンプリングのアルゴリズムを学習する人たちにとって，アルゴリズムの見通しがよくなったと考える．これらの可視化ツールはギブズ・サンプリングの原理を学習する際，学習効果を高めるものであると考える．

[2] https://www.wolfram.com/cdf-player/

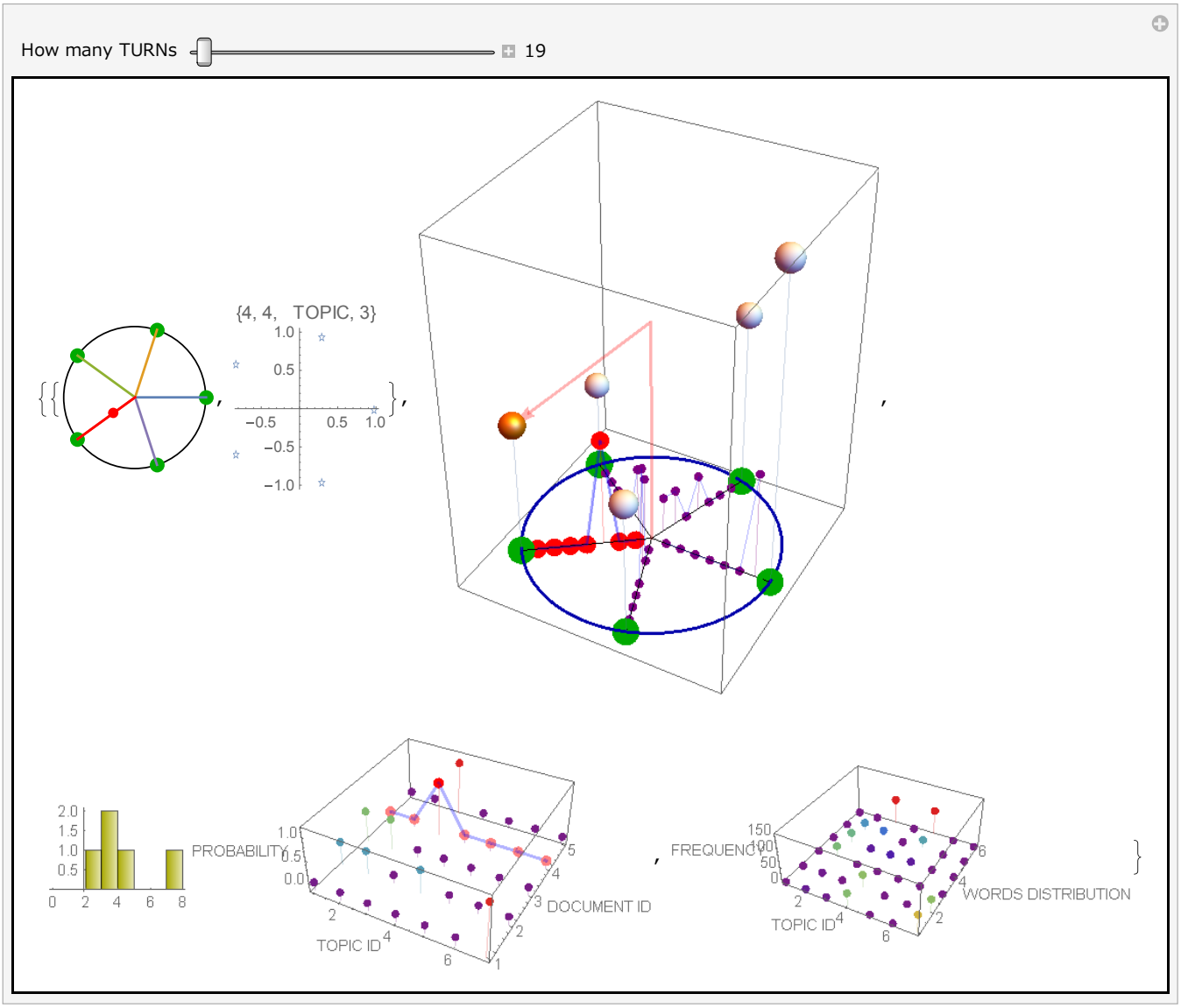

図 2： 混合ユニグラムモデルによるギブズ・サンプリングの可視化ツール

5. まとめ

混合ユニグラムモデルにおける確率分布関数及びギブス・サンプリングの可視化ツールについて報告した．トピック抽出の分野で LDA モデルによるギブズ・サンプリングアルゴリズムによるモンテカルロ法は広く使われているが，その数学的プロセスの意味を理解することは容易ではない．LDA モデルによるギブズ・サンプリングアルゴリズムを多くの学生に理解してもらうことを目的として本可視化ツールを作った．

可視化の表現を分かり易くするために，トピックモデルに代わり，その簡略化モデルである混合ユニモデルを用いた．この可視化により，「１つの文書だけを除いて，それ以外の文書の確率変数を全部固定すると，その１つの文書が，(混合ユニグラムモデルでは)どのトピックに属すべきか分かる」というギブス・サンプリングの特長が容易に理解できることを目的としている．ギブズ・サンプリングを必要として学んでいる学生の数は少ないが，本ツールを見てもらい，ヒアリングを実施したところ，「分かり易い」と好評であった．

こうした統計に係る定理やアルゴリズムの説明に可視化ツールは有効である[11-15]．今後とも，機械学習で普及しているアルゴリズムの可視化教材の作成を続けていきたい．

謝辞

本論文の作成にあたり、常日頃から有意義な助言を頂いております学習院大学計算機センター久保山哲二教授に感謝します．

参 考 文 献

[1] 伊庭幸人, *ベイズ統計と統計物理*: 岩波書店, 2003.

[2] 岩田具治, *トピックモデル*: 講談社サイエンティフィク, 2015.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*: Springer, 2006.

[4] T. L. Griffiths, and M. Steyvers, “"Finding scientific topics,” *Proceedings of the National Academy of Sciences,* vol. 101 (Suppl. 1), pp. 5228–5235, 2004.

[5] 伊庭幸人, 種村正美, 大森裕浩, 和合肇, 佐藤整尚, and 高橋明彦, *計算統計 II マルコフ連鎖モンテカルロ法とその周辺*: 岩波書店, 2005.

[6] Y. Shirota, T. Hashimoto, and S. Suzuki, “Extraction of the Financial Policy Topics by Latent Dirichlet Allocation,” *Proc. of IEEE TENCON 2014*, PID=493, 2014.

[7] Y. Shirota, T. Hashimoto, and T. Sakura, "Topic Extraction Analysis for Monetary Policy Minutes of Japan in 2014," *Advances in Data Mining: Applications and Theoretical Aspects*, Lecture Notes in Computer Science P. Perner, ed., pp. 141-152: Springer International Publishing, 2015.

[8] Y. Shirota, T. Kuboyama, T. Hashimoto, S. Aramvith, and T. Chauksuvanit, *Study of Thailand People Reaction on SNS for the East Japan Great Earthquake - Comparion with Japanese People Reaction* -, No. 59, Occasional papers of Research Institute for Oriental Cultures Gakushuin University, 2015, ISSN 0919-6536.

[9] Y. Shirota, T. Hashimoto, and S. Tamaki, “ MONETARY POLICY TOPIC EXTRACTION BY USING LDA －JAPANESE MONETARY POLICY OF THE SECOND ABE CABINET TERM－,” *Proc. of IIAI International Congress on Advanced Applied Informatics 2015, 12-16 July, 2015, Okayama, Japan*, pp. 8-13, 2015.

[10] D. J. MacKay, *Information Theory, Inference, and Learning Algorithms*: Cambridge university press, 2003.

[11] Y. Shirota, and S. Suzuki, “Visualization of the Central Limit Theorem and 95 Percent Confidence Intervals,” *Gakushuin Economics Papers,* Vol.50, No. 3, 4, pp. 125-136, 2014.

[12] Y. Shirota, T. Hashimoto, and S. Suzuki, “Knowledge Visualization of Reasoning for Financial Mathematics with Statistical Theorems,” *Proc. of the DNIS (Databases in Networked Information Systems) 2014, LNCS 8381, Springer, Heidelberg,*, pp. 132-143, 2014.

[13] YukariShirota, T. Hashimoto, and S. Iitaka, *感じて理解する数学入門 "Introduction to Financial Mathematics" (e-Book written in Japanese)* O'Reilley JAPAN, 2012.

[14] Y. Shirota, and B. Chakraborty, “Visual Explanation of Mathematics in Latent Semantic Analysis,” *Proc. of IIAI International Congress on Advanced Applied Informatics 2015, 12-16 July, 2015, Okayama, Japan*, pp. 423-428, 2015.

[15] Y. Shirota, Y. Takahasi, N. Tanaka, and M. Morita, “Visually Do Statistics for Business Persons Visual Materials from Regression to Black-Sholes Model,” *Proc. of VINCI 2015, ACM, 24-26 August, 2015, Tokyo*, pp. 170-173, 2015.