2G1-2

データの階層構造を利用した潜在変数モデルの自動生成

Generating Probabilistic Latent Variable Models by Exploiting Hierarchical Stractural Information

石畠 正和　　岩田 具治
Masakazu Ishihata　Tomoharu Iwata

NTT コミュニケーション科学基礎研究所
NTT Comunication Science Laboratories

Probabilistic latent variable models have been successfully used to capture intrinsic characteristics of a wide variety of data sets. However, it is nontrivial to design an appropriate model for a given data set because it requires domain knowledge. In this paper, we propose a method to automatically generate a probabilistic latent variable model for a target dataset, which exploits its hierarchical structural information. In experiments, we show that our method can generate correct models by using artificial datasets. We also show that generated models by using real data sets capture those intrinsic characteristics.

1. はじめに

確率モデルは欠損やノイズを含むデータを解析する方法として広く利用されており、中でも確率的潜在変数モデルはデータに潜む構造や特徴を取り出せるため、クラスタリングや分類、次元削減などに利用されている。しかし、対象データに対し、どのような潜在変数が必要で、変数間にどのような依存関係を定めるかは自明ではなく、データに適した潜在変数モデルを設計することは容易ではない。そのため潜在変数モデルの設計はデータの性質の理解や試行錯誤を伴う高度な作業とされており、この自動化が達成されればデータ解析をよりスムーズに行うことができる。更にこれまで人手でモデルを設計することが困難であった複雑な構造を持つデータに対しても、データの構造を反映したモデルを用いた解析が可能になると期待される。

本稿ではデータに適した潜在変数モデルを自動生成するため、データの持つ階層情報を利用する。階層情報とはデータの持つ入れ子構造のことである。例えば文書データでは、各文書は複数の章からなり、各章は複数の文、各文は複数の文字からなる。また購買履歴では、データは複数のユーザの購買履歴の集合であり、各ユーザの購買履歴は複数回の買い物からなり、各買い物は複数の商品を含む。仮に文書やユーザをクラスタリングするために潜在変数モデルを用いる場合、文書やユーザの階層に対して潜在変数が必要であることは明らかだが、他の階層においてどのように潜在変数を用意するかは自明ではない。

本稿ではこの階層情報を持つデータに対する潜在変数モデルの自動生成法を提案する。まず階層構造を順序木で表現し、順序木を用いた一般的な潜在変数モデルを提案する。提案モデルは各階層における潜在変数の有無や依存関係をモデルパラメータにより調整可能である。提案法はこのパラメータを、モデル選択基準として利用できる変分自由エネルギーを最大化するよう最適化することでデータに適した潜在変数モデルの自動生成を実現する。結果として、提案法は Multinomial Mixture Model (MMM), Hidden Markov Model (HMM). Latent Dirichlet Allocation (LDA) など既存のよく知られたモデルを包含する。本稿では提案法を人工データに適用し、データの生成に利用したモデルを生成できることを確認する。また特徴の異なる 2 つの実データに対して提案法を適用し、特徴を反映したモデルが生成されることを確認する。

連絡先: ishihata.masakazu@lab.ntt.co.jp

2. 関連研究

複数のモデルからデータに適したモデルを選択するための基準としてモデル選択基準の研究が古くからなされている [7]。モデル選択基準を用いてモデルの自動生成を行うには、モデル候補を自動的に生成する枠組みが別途必要である。提案法はデータの持つ階層情報を元にモデル候補を生成し、モデル選択基準を用いて最良なモデルを探索する。

グラフィカルモデルの構造学習は確率モデルの自動生成の一種である。構造学習では観測変数間の条件付き独立性を推定するため、多くの場合、潜在変数を考慮しない。これに対して提案法は、観測変数は潜在変数にのみ依存すると仮定し、潜在変数の依存関係を推定することを目的とする。

潜在変数の階層構造を推定する手法がいくつか提案されている。潜在変数モデルの一種であるトピックモデルは、データを単語集合の集まりと捉え、各単語は対応するトピックと呼ばれる潜在変数から生成されると仮定する。[2, 9] はトピックの階層構造を抽出するモデルである。また [5] はデータを表す行列に対し、繰り返し行列分解を適用することで潜在変数の階層構造を学習する。これらの手法はデータの持つ階層情報を陽に利用しないが、提案法は階層構造を積極的に利用することで潜在変数の階層構造を推定する。

データの階層構造を反映したトピックモデルもいくつか提案されている [4, 6]。これらの手法は全階層に潜在変数を導入し、同じ階層にある潜在変数の依存関係を考慮しない。提案法は同じ階層内の潜在変数の依存関係も考慮し、データを説明するのに不要である階層や依存関係を取り除くことが可能である。

3. 提案モデル

3.1 階層情報の順序木表現

本稿では観測データのもつ階層情報は順序木で表現されるとする。つまりデータ $\boldsymbol{D}$ は観測列 $\boldsymbol{x} \equiv (x_n)_{n=1}^N$ と階層情報を表す順序木 $\boldsymbol{T}$ の組として与えられるとする。 順序木 $\boldsymbol{T}$ は 3 つ組 (N, par, sib) で定義され、$\mathrm{N} = \{0, \ldots, N\}$ は $\boldsymbol{T}$ の節点集合、写像 $par : \mathrm{N} \to \mathrm{N}$ と $sib : \mathrm{N} \to \mathrm{N}$ はそれぞれ $\boldsymbol{T}$ 中の親子関係と順序関係を定義する。つまり $par(n)$ と $sib(n)$ はそれぞれ節点 n の親と兄である。$D_{\boldsymbol{T}}$ と d_n をそれぞれ $\boldsymbol{T}$ と n の深さとする。また N_d $(1 \leq d \leq D_{\boldsymbol{T}})$ を $d_n = d$ なる節点集合とする。各節点 n は対応する観測変数 x_n を持ち、同じ深

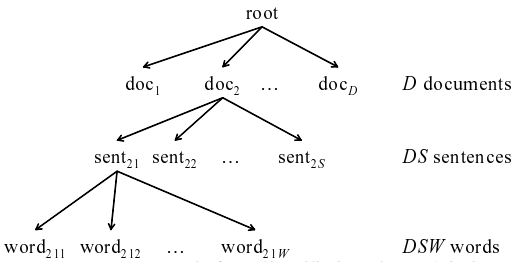


図 1: 文書の階層構造を表す順序木

さにある節点の観測変数は同じ値域を持つとする。本稿では x_n $(n \in \mathrm{N}_d)$ は離散値 $\{1, \ldots, V_d\}$ を取るとし、 $V_d = 0$ により $n \in \mathrm{N}_d$ が観測を持たないこと表す。

例えば、データとして文書集合が与えられたとする。データは D 文書からなり、各文書は S 文から、各分は W 語からなるとする。このとき、このデータの階層情報は図 1 の順序木で表現される。

3.2 モデル定義

階層情報を持つデータ $\boldsymbol{D} = (\boldsymbol{x}, \boldsymbol{T})$ に対する潜在変数モデルを定義する。提案モデル $\boldsymbol{M}$ は順序木 $\boldsymbol{T}$, 仮定 $\boldsymbol{A}$ そしてモデルパラメータ $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_d)_{d=1}^{D_{\boldsymbol{T}}}$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_d)_{d=1}^{D_{\boldsymbol{T}}}$ の 4 つ組で定義されるとする。各節点 n は (離散) 潜在変数 $z_n \in \{1, \ldots, K\}$ を持ち、観測変数 x_n は z_n に依存して生成されると仮定する。一方、z_n は n の親 $par(n)$ と兄 $sib(n)$ の潜在変数 $z_{par(n)}, z_{sib(n)}$ に依存する。以後、表記を簡潔にするため $l = par(n), m = sib(n)$ とする。深さ d の潜在変数 z_n の依存関係は仮定変数 A_d により表 1 に示すように定義されるとする。I-det, P-det はそれぞれ z_n が親の ID, 親の値を決定的に取ることを意味する。N-dep は z_n は他の潜在変数に依存しないことを意味する。P-dep, S-dep はそれぞれ z_n が親の潜在変数、兄の潜在変数にのみ依存することを意味し、B-dep はその両方に依存することを意味する。潜在変数 $\boldsymbol{z} = (z_n)_{n=1}^{N}$ と観測変数 $\boldsymbol{x} = (x_n)_{n=1}^{N}$ の生成仮定を以下とする。

1. For each depth $d = 1, \ldots, D_{\boldsymbol{T}}$
   (a) Draw topic distributions $\boldsymbol{\theta}_{d,i,j} \sim \mathrm{Dir}(\boldsymbol{\alpha}_d)$
   (b) Draw symbol distributions $\boldsymbol{\phi}_{d,k} \sim \mathrm{Dir}(\boldsymbol{\beta}_d)$
2. For each depth $d = 1, \ldots, D_{\boldsymbol{T}}$, for each node $n \in \mathrm{N}_d$
   (a) Choose a topic z_n by

   case A_d
   when I-det: $z_n := n$
   when P-det: $z_n := z_l$
   when N-dep: $z_n \sim \mathrm{Cat}(\boldsymbol{\theta}_{d,0,0})$
   when P-dep: $z_n \sim \mathrm{Cat}(\boldsymbol{\theta}_{d,z_l,0})$
   when S-dep: $z_n \sim \mathrm{Cat}(\boldsymbol{\theta}_{d,0,z_m})$
   when B-dep: $z_n \sim \mathrm{Cat}(\boldsymbol{\theta}_{d,z_l,z_m})$

   (b) Draw a symbol $x_n \sim \mathrm{Cat}(\boldsymbol{\phi}_{d,z_n})$

ここで $\boldsymbol{\theta}_{d,i,j} \equiv (\theta_{d,i,j,k})_{k=1}^{K}$ と $\boldsymbol{\phi}_{d,k} \equiv (\phi_{d,k,v})_{v=1}^{V_d}$ はカテゴリカル分布のパラメータであり、$\theta_{d,i,j,k}$ は深さ d において $z_l = i$, $z_m = j$ のときに $z_n = k$ である確率、$\phi_{d,k,v}$ は $z_n = k$ のときに $x_n = v$ である確率である。また $\boldsymbol{\alpha}_d \equiv (\alpha_{d,k})_{k=1}^{K}$ と $\boldsymbol{\beta}_d \equiv (\beta_{d,v})_{v=1}^{V_d}$ はディリクレ分布のパラメータであり、$\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ の事前分布のパラメータである。

A_d	Explanation	Dependency
I-det	Index-deterministic	$z_n := n$
P-det	Parent-deterministic	$z_n := z_l$
N-dep	Non-dependent	$z_n \perp\!\!\!\perp z_l$, $z_n \perp\!\!\!\perp z_m$
P-dep	Parent-dependent	$z_n \not\perp\!\!\!\perp z_l$, $z_n \perp\!\!\!\perp z_m$
S-dep	Sibling-dependent	$z_n \perp\!\!\!\perp z_l$, $z_n \not\perp\!\!\!\perp z_m$
B-dep	Both-dependent	$z_n \not\perp\!\!\!\perp z_l$, $z_n \not\perp\!\!\!\perp z_m$

表 1: 仮定変数 A_d と依存関係 ($p = par(n)$, $s = sib(n)$)。

	$\boldsymbol{A} = (A_1, A_2, A_3)$	Corresponding Model
(1)	N-dep, P-det, P-det	dMMM
(2)	I-det, P-det, P-dep	dLDA
(3)	I-det, I-det, S-dep	wHMM
(4)	I-det, S-dep, P-dep	sHMM + wMMM
(5)	I-det, P-det, B-dep	dLDA + wHMM
(6)	I-det, B-dep, P-dep	dLDA + sHMM + wMMM

表 2: 提案モデルによる既存モデルの表現例。ここで頭文字の d, s, w はそれぞれ文書レベル、文レベル、単語レベルであることを意味する。

提案モデルは様々な潜在変数モデルを表現可能である。例えば、階層情報として図 1 の順序木が与えられたとする。このとき仮定 $\boldsymbol{A}$ を調整することで表 2 に示すように、Multinomial Mixture Model(MMM) や Hidden Markov Model (HMM), Latent Dirichlet Allocation (LDA) [3], そしてそれらを合わせたモデルなどが表現できる。

3.3 モデル生成

階層情報付きデータ $\boldsymbol{D} = (\boldsymbol{x}, \boldsymbol{T})$ が与えられたとき、提案モデル $\boldsymbol{M} = (\boldsymbol{T}, \boldsymbol{A}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ の仮定 $\boldsymbol{A}$ を調整することで $\boldsymbol{D}$ に適した潜在変数モデルを生成する。ここで仮定 $\boldsymbol{A}$ がどれだけデータ $\boldsymbol{D}$ に合っているかを測る尺度として対数周辺尤度 $\mathcal{L}[\boldsymbol{M}] \equiv \ln p(\boldsymbol{x} \mid \boldsymbol{M})$ が考えられる [7]。しかし $\mathcal{L}[\boldsymbol{M}]$ を計算するには $\boldsymbol{z}$ に対する全割り当てを考える必要があり、指数的な計算時間を要する。そこで本稿では $\mathcal{L}[\boldsymbol{M}]$ の下限値である変分自由エネルギー $\mathcal{F}[\boldsymbol{A}]$ をモデルの選択基準とする。変分自由エネルギーの定義と計算法については次章で述べる。しかし変分自由エネルギー $\mathcal{F}[\boldsymbol{A}]$ が計算できても、変分自由エネルギーを最大化する $\boldsymbol{A}$ を直接計算することは困難である。そこで本稿では以下の局所探索により $\boldsymbol{A}$ を決定する。

1. 初期仮定 $\boldsymbol{A}$ を $\forall d(A_d = \text{P-det})$ とし、初期仮定候補を $C = \{\boldsymbol{A}\}$ とする。
2. 仮定候補 C 中の全 $\boldsymbol{A}$ に対してスコア $\mathcal{F}[\boldsymbol{A}]$ を計算する。
3. スコアの最大値が更新されなければ終了、更新されれば C 中の最もスコアの高い仮定 w 個の隣接仮定を新たな C とし、2. へ戻る。
4. 最もスコアの高い $\boldsymbol{A}$ を最終結果として出力する。

4. 変分自由エネルギー

変分自由エネルギー $\mathcal{F}[\boldsymbol{A}]$ の定義と計算法を述べる。Jensen の不等式より以下の対数周辺尤度 $\mathcal{L}[\boldsymbol{M}]$ の下限を得る。

$$\begin{aligned}\mathcal{L}[\boldsymbol{M}] &\geq \mathrm{E}_q[\ln p(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{\phi})] + \mathrm{E}_q[\ln p(\boldsymbol{z} \mid \boldsymbol{A}, \boldsymbol{\theta})] \\ &\quad + \mathrm{E}_q[\ln p(\boldsymbol{\theta} \mid \boldsymbol{\alpha})] + \mathrm{E}_q[\ln p(\boldsymbol{\phi} \mid \boldsymbol{\beta})] - \mathrm{H}[q] \\ &\equiv \mathcal{F}[q, \boldsymbol{M}]\end{aligned}$$

ここで q は $q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) \equiv q(\boldsymbol{z})\, q(\boldsymbol{\theta})\, q(\boldsymbol{\phi})$ を満たす近似分布であり、$\mathrm{H}[q]$ はそのエントロピーである。この近似分布 q を下式

で繰り返し更新することで、下限 $\mathcal{F}[q, \boldsymbol{M}]$ を最大化できる。

$$q(\boldsymbol{z}) \propto \exp\big(\mathrm{E}_{q(\phi)}[\ln p(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{\phi})] + \mathrm{E}_{q(\theta)}[\ln p(\boldsymbol{z} \mid \boldsymbol{\theta})]\big)\,,$$
$$q(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \exp\big(\mathrm{E}_{q(\boldsymbol{z})}[\ln p(\boldsymbol{z} \mid \boldsymbol{\theta})]\big)\,,$$
$$q(\boldsymbol{\phi}) \propto p(\phi \mid \boldsymbol{\beta}) \exp\big(\mathrm{E}_{q(\boldsymbol{z})}[\ln p(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{\phi})]\big)\,.$$

更にモデルパラメータ $\boldsymbol{\alpha}, \boldsymbol{\beta}$ も [8] により推定可能である。本稿では推定された q, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ を用いて計算された下限 $\mathcal{F}[q, \boldsymbol{M}]$ を $\mathcal{F}[\boldsymbol{A}]$ と表し、変分自由エネルギーと呼ぶ。

次に具体的な更新式を示す。$q(\boldsymbol{\theta})$, $q(\boldsymbol{\phi})$ を以下とする。

$$q(\boldsymbol{\theta}) = \prod_{d=1}^{D}\prod_{i=1}^{K}\prod_{j=1}^{K} \mathrm{Dir}(\boldsymbol{\theta}_{d,i,j}; \boldsymbol{a}_{d,i,j})\,,$$
$$q(\boldsymbol{\phi}) = \prod_{d=1}^{D}\prod_{k=1}^{K} \mathrm{Dir}(\boldsymbol{\phi}_{d,k}; \boldsymbol{b}_{d,k})\,,$$

ここで $\boldsymbol{a}_{d,i,j} = (a_{d,i,j,k})_{k=1}^{K}$ と $\boldsymbol{b}_{d,k} = (b_{d,k,v})_{v=1}^{V_d}$ は近似分布のパラメータであり、更新式は以下となる。

$$a_{d,i,j,k} = \alpha_{d,i,j,k} + \mathrm{E}_{q(\boldsymbol{z})}[c_{d,i,j,k}]\,, \tag{1}$$
$$b_{d,k,v} = \beta_{d,k,v} + \mathrm{E}_{q(\boldsymbol{z})}[c_{d,k,v}]\,, \tag{2}$$
$$c_{d,i,j,k} \equiv |\{n \in \mathrm{N}_d \mid z_l = i, z_m = j, z_n = k\}|,$$
$$c_{d,k,v} \equiv |\{n \in \mathrm{N}_d \mid z_n = k, x_n = v\}|,$$

ここで $p = par(n)$, $s = sib(s)$ であり、$q(\boldsymbol{z})$ は以下である。

$$q(\boldsymbol{z}) \propto \prod_{d=1}^{D}\prod_{k=1}^{K}\prod_{v=1}^{V_d} {\phi^*_{d,k,v}}^{c_{d,k,v}} \prod_{d=1}^{D}\prod_{i=1}^{K}\prod_{j=1}^{K}\prod_{k=1}^{K} {\theta^*_{d,i,j,k}}^{c_{d,i,j,k}},$$
$$\phi^*_{d,k,v} \equiv \exp\left(\Psi(b_{d,k,v}) - \Psi\left(\sum_{l=1}^{V_d} b_{d,k,l}\right)\right),$$
$$\theta^*_{d,i,j,k} \equiv \exp\left(\Psi(a_{d,i,j,k}) - \Psi\left(\sum_{l=1}^{V_d} a_{d,i,j,l}\right)\right).$$

ここで $q(\boldsymbol{z}) = p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$ が成り立つ。これより式 (1), (2) 中の期待値は以下のように計算できる。

$$\mathrm{E}_{q(\boldsymbol{z})}[c_{d,i,j,k}(\boldsymbol{z})] \propto \sum_{n \in \mathrm{N}_d} p_{n,i,j,k} \tag{3}$$

$$\mathrm{E}_{q(\boldsymbol{z})}[c_{d,k,v}(\boldsymbol{z})] \propto \sum_{n \in \mathrm{N}_d \text{ s.t } x_n = v} \sum_{i=1}^{K}\sum_{j=1}^{K} p_{n,i,j,k} \tag{4}$$

$$p_{n,i,j,k} \equiv p(z_l = i, z_m = j, z_n = k, \boldsymbol{x} \mid \boldsymbol{\theta}^*, \boldsymbol{\phi}^*) \tag{5}$$

よって近似分布のパラメータ $\boldsymbol{a}, \boldsymbol{b}$ は、式 (3) (4) の期待値計算と、式 (1) (2) の更新を繰り返すことで推定できる。

最後に式 (5) の確率 $p_{n,i,j,k}$ の計算法を述べる。この確率を愚直に計算すると、無関係である潜在変数をすべて周辺化する必要があるため、指数的な時間を要する。本稿ではこれを動的計画法により効率的に計算する。節点 n の子孫を $\mathrm{Dec}(n)$ とし、 n の弟集合を $\mathrm{Sib}^-(n)$, n の兄集合を $\mathrm{Sib}^+(n)$ とする。更に以下の４種の節点集合を導入する。

$$\mathrm{I}(n) \equiv \{n\} \cup \mathrm{Dec}(n)\,, \qquad \mathrm{O}(n) \equiv \mathrm{N} \backslash \mathrm{Dec}(n)\,,$$
$$\mathrm{F}(n) \equiv \bigcup_{m \in \mathrm{Sib}^-(n)} \mathrm{I}(m)\,, \qquad \mathrm{B}(n) \equiv \bigcup_{m \in \mathrm{Sib}^+(n)} \mathrm{I}(m)\,.$$

定義より $\mathrm{N} = \mathrm{O}(p) \cup \mathrm{F}(s) \cup \mathrm{B}(n)$ である。ある節点集合 $C \subseteq \mathrm{N}$ に対し、 $\boldsymbol{x}_C \equiv (x_n)_{n \in C}$, $\boldsymbol{z}_C \equiv (z_n)_{n \in C}$ とする。すると p_{ijk} は以下のように分解できる。

$$p_{n,i,j,k} = p\big(\boldsymbol{x}_{\mathrm{O}(p)}, z_l = i\big)\, p\big(\boldsymbol{x}_{\mathrm{F}(s)}, z_m = j \mid z_l = j\big)$$
$$p\big(\boldsymbol{x}_{\mathrm{B}(n)}, z_n = k \mid z_l = i, z_m = j\big)$$

これを効率的に計算するため、以下の４種の確率を導入する。

$$I_n[k] \equiv p\big(x_{\mathrm{I}(n)} \mid z_n = k\big)$$
$$O_n[k] \equiv p\big(x_{\mathrm{O}(n)}, z_n = k\big)$$
$$F_n[i,k] \equiv p\big(x_{\mathrm{F}(n)}, z_n = k \mid z_l = i\big)$$
$$B_n[i,j,k] \equiv p\big(x_{\mathrm{B}(n)}, z_n = k, \mid z_l = i, z_m = j\big)$$

これらの確率は以下の動的計画法により効率的に計算できる。

$$I_n[k] = \phi_{d,k,x_n} B_c[k, 0]$$
$$O_n[k] = \sum_{i=1}^{K}\sum_{j=1}^{K} O_n[i,j,k]$$
$$O_n[i,j,k] = O_p[k]\, F_s[i,j]\, B_t[i,j]\, \phi_{d,k,x_n} \theta_{d,i,j,k}$$
$$F_n[i,k] = I_n[k] \sum_{j=1}^{K} F_s[i,j]\, \theta_{d,i,j,k}$$
$$B_n[i,j] = \sum_{k=1}^{K} B_n[i,j,k]$$
$$B_n[i,j,k] = I_n[k]\, B_t[i,k]\, \theta_{d,i,j,k},$$

ここで c は n の長子であり、t は n の弟である。これより目的の確率は以下のように計算できる。

$$p_{n,i,j,k} = O_p[i]\, F_s[i,j]\, B_n[i,j,k]\,.$$

提案手法の最悪計算量は $O(NK^3)$ である。しかし、仮定 $\boldsymbol{A}$ によってその計算量は減少する。例えば LDA を表現する $\boldsymbol{M}$ に対する計算量は $O(NK)$ であり、HMM に対する計算量は $O(NK^2)$ となり、これはそれぞれのモデル専用の学習アルゴリズムと同じである。

5. 実験

5.1 人工データ

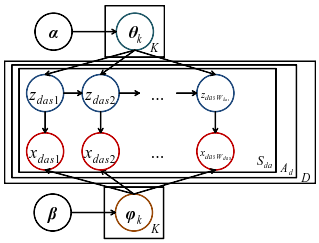
提案法を正解モデルが分かる人工データに適用した。図 1 の順序木を階層情報として持つ正解モデルを１２種類用意し、各モデルから L 文書、L 文、L 単語の計 L^3 語からなるデータセットを生成した。なおクラスタ数は $K = 5$, 語彙数は $V_1 = V_2 = 0$, $V_3 = 500$ とした。正解モデルと生成されたモデルを表 3 に示す。ここで探索幅は $w = 3$ とし、データサイズは $L = 10, 30, 50$ と変化させた。表中の赤字は間違って推定された仮説を表す。表より、簡単なモデルに関しては少ないデータ数から正解モデルを復元できていることが分かる。また、複雑なモデルに関しても、データ数を増やすほど正確にモデルが復元できており、$L = 50$ のときに全データに対して正しくモデルの生成が行えていることがわかる。

5.2 実データ

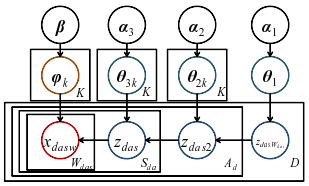
提案法を実データに適用した。ここでは Reuters-21578 [1] の一部を利用した。このデータセットは 1987 年の Reuters に

ID	Correct Model	$L = 10$	$L = 30$	$L = 50$
1.	dMMM	N-dep, P-det, P-det	N-dep, P-det, P-det	N-dep, P-det, P-det
2.	sMMM	I-det, N-dep, P-det	I-det, N-dep, P-det	I-det, N-dep, P-det
3.	dLDA	I-det, P-det, P-dep	I-det, P-det, P-dep	I-det, P-det, P-dep
4.	sLDA	I-det, I-det, P-det	I-det, S-dep, P-det	I-det, I-det, P-dep
5.	dHMM	I-det, P-det, P-dep	S-dep, P-det, P-det	S-dep, P-det, P-det
6.	sHMM	I-det, S-dep, P-det	I-det, S-dep, P-det	I-det, S-dep, P-det
7.	wHMM	P-det, P-det, P-det	N-dep, P-det, B-dep	N-dep, P-det, B-dep
8.	dHMM + wMMM	I-det, P-det, P-dep	I-det, P-det, P-dep	S-dep, P-det, P-dep
9.	sHMM + wMMM	P-det, P-det, P-det	N-dep, B-dep, P-dep	I-det, S-dep, P-dep
10.	dLDA + sHMM	P-det, S-dep, P-det	S-dep, B-dep, P-det	I-det, P-det, B-dep
11.	dLDA + wHMM	P-det, P-det, P-det	S-dep, P-det, B-dep	I-det, B-dep, P-det
12.	dLDA + sHMM + wMMM	P-det, P-det, P-det	S-dep, B-dep, P-det	I-det, B-dep, P-dep

表 3: 正解モデルと生成されたモデル。



(1) wHMM



(2) dMMM + aMMM + sMMM

図 2: 最終モデル。1 つ目のデータセットに対しては単語レベルの HMM が、2 つ目のデータセットに対しては 3 層の MMM が生成された。

掲載された新聞記事の集合であり、本稿ではそのうち 3 月の記事を利用した。データは日数 29、記事数 10,535、文数 79,15、語彙数 31,057 の 4 階層であり、0-9 の数字を含む単語はすべて “NUM” に置き換えた。このデータより以下の 2 種のデータセットを作成した。1 つ目は頻出の 5,000 語を利用し、2 つ目は頻出の 100 語をストップワードとして取り除いた上で上位 5,000 語を利用した。この 2 つのデータセットに対し、提案法を適用した。ただし、データセットサイズに対して学習計算量が $O(NK^3)$ のモデルは効率的に学習できないため、探索から除外した。クラスタ数は $K = 10, 20, 30$ と変化させ、最も変分自由エネルギーが高い結果を最終モデルとした。なお探索幅は $b = 3$ とした。図 5.2 に最終モデルを示す。1 つ目のデータセットに対しては単語レベルの HMM が推定された。このデータセットは頻出単語をそのまま利用しているため、(数, 単位) や (be 動詞, 冠詞), (冠詞, 名詞) などのパターンが頻出しており、これをモデルとして表現している。一方、2 つ目のデータセットに対しては 3 層 MMM が推定された。このデータセットでは頻出単語をストップワードとして取り除いたため、文中の単語の順序が破壊されており、文を bag-of-words として扱うモデルが生成された。このように提案法はデータの性質にあったモデルを自動的に生成できてる。

更に比較対象として同データセットに LDA を適用した。ここでクラスタ数は $K = 10, 20, \ldots, 100$ とし、最も高い変分自由エネルギーを最終結果とした。表 5.2 に最終モデルと LDA の変分自由エネルギーを示す。表より提案法をは変分自由エネルギーという尺度の元では LDA より良いモデルを生成できていることがわかる。

Model	First dataset	Second dataset
day-LDA	-8.739×10^6	-4.891×10^6
article-LDA	-8.299×10^6	-4.609×10^6
sentence-LDA	-8.554×10^6	-4.842×10^6
Generated model	-7.658×10^6	-4.555×10^6

表 4: 最終モデルと LDA の変分自由エネルギー。

6. おわりに

階層情報を持つデータに対する潜在変数モデルの自動生成法を提案した。提案法は変分自由エネルギーをスコアとし、各階層の潜在変数の依存関係を局所探索により推定する。人工データを用いた実験より、提案法はデータが生成されたモデルを復元できることを示した。また、実データよりデータの特徴に合ったモデルを生成できることがわかった。

参考文献

[1] Reuters-21578 text categorization test collection. `http://www.daviddlewis.com/resources/testcollections/reuters21578/`.

[2] DM Blei, TL Griffiths, MI Jordan, and JB Tenenbaum. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In *NIPS*, 2003.

[3] DM Blei, AY Ng, and MI Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[4] Lan Du, Wray Buntine, and Huidong Jin. A segmented topic model based on the two-parameter Poisson-Dirichlet process. *Machine learning*, 81(1):5–19, July 2010.

[5] Roger Grosse and RR Salakhutdinov. Exploiting compositionality to explore a large space of model structures. In *UAI*, 2012.

[6] Do-kyum Kim, G Voelker, and LK Saul. A Variational Approximation for Topic Modeling of Hierarchical Corpora. In *ICML*, volume 28, 2013.

[7] David J C Mackay. Bayesian interpolation. *Neural computation*, 4(3):415–447, May 1992.

[8] Thomas P Minka. Estimating a Dirichlet distribution, 2000.

[9] YW Teh and MI Jordan. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.