DEIM2020 F8-3(day2 p54)

変分ベイズにおける最適解探索効率の検証

岡　威久馬†　　若林　　啓††

† 筑波大学情報学群知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †s1611490@u.tsukuba.ac.jp, ††kwakaba@slis.tsukuba.ac.jp

あらまし　データの背後に隠れたパターンを発見し，解釈可能な知識を抽出することは重要な課題である．このような課題に対して，確率モデルによる統計的データ分析に基づく手法が有効と考えられており，クラスタリングをおこなう際に用いられる混合ガウスモデル（GMM）や系列データを扱う際に用いられる隠れマルコフモデル（HMM）のようなモデルが提案されている．これらのモデルを最適化する手法として変分ベイズ法が広く用いられているが，変分ベイズ法は決定論的なアルゴリズムであり，一度局所最適解に陥ると抜け出すことができない．本研究では，この問題を解決するための手法として摂動を加えるアプローチに着目し，期待値のモンテカルロ近似を用いた手法や，確率的勾配降下法を用いた手法について，最適解の探索効率を検証する．いくつかの条件の下で行なった実験の結果を示し，それぞれの手法の特性について得られた示唆を報告する．

キーワード　統計的学習，機械学習，クラスタリングアルゴリズム, 変分ベイズ, GMM, HMM

1 はじめに

近年, 確率モデルの学習アルゴリズムを用いたデータ分析が注目されている. ビッグデータの背後に隠れたパターンを発見し, 解釈可能な知識を抽出することは重要な課題である. この課題は, 対象とするデータに応じて, 類似したデータのグループを発見するクラスタリングや, 系列データからパターンを見つける系列データマイニングなどと呼ばれている. このような課題に対して, 確率モデルによる統計的データ分析に基づく手法が有効と考えられており, クラスタリングをおこなう際に用いられる混合ガウスモデル (GMM) や系列データを扱う際に用いられる隠れマルコフモデル (HMM) のようなモデルが提案されている.

GMM や HMM といったモデルは, モデルの関数の形を決めるパラメータが存在する. 確率モデルを用いたデータ分析では, モデルのパラメータを最適化することによって, 未知のデータに対して正確な予測をおこなうことが可能となる. パラメータを最適化する手法は数多く存在し, その中でも反復的数値最適化手法である変分ベイズは, パラメータ最適化手法の中でも広く知られている手法である. そして, 変分ベイズは GMM や HMM のようなモデルに対して用いることができる [1] [2].

GMM や HMM のようなモデルがデータ分析において有効であるということは先行研究によって明らかになっている [3] [4] が, これらの手法は局所最適解に陥りやすいという問題を抱えている. 局所最適解とは, 最適化したい関数のすべての変域のうち, 一部の変域を取り出して考えたとき, その範囲の中で最適解となる場所である. この方法によって求めた最適解は, 一部の変域のみを調べているので, そこで最適化されたパラメータが全体の最適解とは限らない. 局所最適解において定まったパラメータが良い精度をもたらす場合もあるが, 精度の悪い結果になる場合もある. GMM や HMM のようなモデルでは, 多数のパラメータを用いるので, 悪い局所最適解に陥りやすいという問題がある.

局所最適解を抜け出し, それよりも良い最適解を探索する手法はいくつか提案されている. そのうちの一つに, 乱数による摂動を加えるというアプローチがあり, 確率的勾配降下法 (SGD) や期待値のモンテカルロ近似などの手法が先行研究によって提案されている. しかし, これらの手法は局所最適解を抜け出し, より良い最適解を見つける手法として有効であるということは明らかになっているが, どれだけ効率的に最適解を探索できているかということは明らかになっていない. 探索効率を検証することは, データの学習を高速化することや, データに合わせた最適な学習アルゴリズムの選択することに対して有益である. また, 機械学習アルゴリズムを用いてデータ分析をする際, 多層ニューラルネットワークなどの手法において, SGD を合わせた手法がさまざまなデータに対して用いられるが, 実際にはそれが最適な選択肢であるのかはまだあまり明らかにされていない. データサイズやデータの種類によっては, SGD よりも効率よく最適化がおこなえる手法が存在する可能性がある. よって, 探索効率を明らかにし, それぞれの手法を比較することでデータに合わせたアルゴリズムを提案することができる. また, それぞれの手法の良いところを取り入れ, 新たなアルゴリズムを提案することも可能である.

本研究では, 変分ベイズ手法を用いて, 摂動を加えない学習アルゴリズム, SGD, 期待値のモンテカルロ近似の 3 手法を GMM と HMM の 2 つのモデルに対して適用し, 最適解の探索効率を検証し明らかにする.

2 関連研究

近年, 深層学習や確率モデルを用いた学習アルゴリズムに関

する研究は盛んに研究行われており, 多くの手法が提案されている. 多層ニューラルネットワークを利用した学習アルゴリズムや, 潜在変数を持った確率モデルに対して用いられる EM アルゴリズムのような反復的数値最適化手法などがその例である. しかし, これらの手法は, データの学習中に局所最適解に陥り学習が進まなくなるという問題を抱えている. この問題を解決するための手法はいくつか提案されており, 現在でも盛んに研究されている対象である.

例えば, 確率的勾配降下法 (SGD) は局所最適解を抜け出すための手法として広く用いられている. SGD は勾配法にランダム性を取り入れた手法のことで, 勾配法はある関数において最小値を求めるための最適化手法である. ある関数 $f(x)$ を最小化したいとき, 初期値を x, $f(x)$ の微分を $\frac{\partial f}{\partial x}$, ステップサイズを ρ とするとき, 勾配法の更新式は以下のようになる.

$$x' = x - \rho \frac{\partial f}{\partial x} \tag{1}$$

と表すことができる. ρ はハイパーパラメータであり, 任意の値を用いることができる. 式 (2.1) を反復的に用いることで最小値へとたどり着くことが可能となる. これを応用させた手法が SGD である. SGD はランダムに抽出した学習データのみを用いて求めた勾配に基づいてパラメータを更新する. このことから, ノイズを加える効果があり, より良い最適解へたどり着くことが可能となる. Smith [5] らは, 多層ニューラルネットワークにおいて, SGD を用いることで局所解よりも大域解に近づくことができることを示している. 本研究では, SGD の考え方を変分ベイズに適用し, ランダムにサンプリングしたデータのみで求めた勾配を用いる Stochastic Variational Inference(SVI) を最適解探索の検証の対象とする.

また, Valentin ら [6] は, EM アルゴリズムにおいて, softEM と hardEM という 2 つの手法を比較している. softEM は, EM アルゴリズムと同様の手法である. 一方, hardEM は期待値を計算した時, 一番高い確率を持つ要素に 1 を割り当て, それ以外の要素には 0 を割り当てるという方法をとり, 負担率を再割り当てした後に M ステップを実行する. [6] の実験によると, hardEM の方が精度が良いことが示されている. この事実は, 本研究において重要なことである. 本研究では期待値のモンテカルロ法を用いた近似を利用するが, hardEM と同様に負担率を 2 値変数に近似する. 再割り当てを実行する時にランダムな摂動を加えるという点のみが hardEM と異なる点である. このため, 変分ベイズに SEM を導入した手法も hardEM と同様に精度が向上することが見込める.

David ら [7] は, 混合ガウスモデルにおいて, EM アルゴリズム, EM アルゴリズムに期待値の近似を取り入れた手法 (SEM), SAEM, MCEM の 4 手法の学習を比較している. SEM は, E ステップを行なった後, 1 か 0 の再割り当てをおこなう S ステップとして実行される. hardEM の場合では, E ステップで負担率を計算した後, それぞれのデータの負担率の 1 番高い確率を 1, それ以外を 0 に再割り当てする. しかし, SEM では負担率の 1 番大きい要素を必ず 1 に置き換えるのではなく, 負担率にしたがってランダムに要素を選択する. そして選択された要素が 1 に置き換わり, それ以外の要素は 0 に置き換えられる. このため, hardEM のように 1 に割り当てられる要素は E ステップを実行した時点ではわからない. このランダム性があることにより, ノイズを加えることができ, 局所最適解を抜け出すことができる. SAEM は, SEM を改良したアルゴリズムで, 負担率の要素を 0 と 1 に再割り当てした後, 確率的勾配降下法のようにステップサイズを導入し, 以下のように更新する.

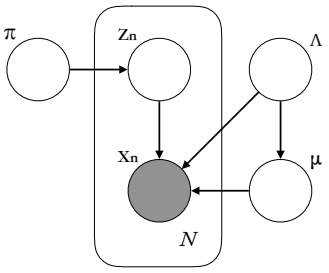

図 1　GMM のグラフィカルモデル

$$\theta^i = (1 - \rho_i)\theta^i_{EM} + \rho_i \theta^i_{SEM} \tag{2}$$

ρ_i はステップサイズであり, θ^i_{EM} は EM アルゴリズムの M ステップによって更新された値, θ^i_{SEM} は, M ステップ後に負担率の再割り当てが行われた値である. 最後に MCEM は, E ステップをモンテカルロ法によって近似する手法である. E ステップの計算が計算量的に困難であるときに用いられる. この手法は, 複数のサンプルをモンテカルロ法によって抽出することによって E ステップを実行する. 複数のサンプル抽出をおこなわず, 1 つのみ抽出する場合, MCEM は SEM へ帰着される.

この研究では, SEM が他の手法よりも良い推定をおこなっていることが報告された. [8] でも述べられているように, 通常の EM アルゴリズムは必ず局所最適解に陥るという問題があるが, 確率的に近似することで局所解から抜け出すことができる. SEM が他の手法よりも良い結果を得られた理由は, この EM アルゴリズムの問題をランダム性を取り入れることによって局所解を抜け出すことができたからである.

本研究では, 上で述べた SVI, SEM を変分ベイズに適用し, 最適解の探索効率を検証する. 変分ベイズとは, 未知パラメータに分布を仮定することで EM アルゴリズムを拡張する手法である. 複雑な確率モデルのパラメータを厳密な計算によって推定することが困難なとき, 変分ベイズを用いて分布を近似することにより, 現実的な時間でパラメータ推定をおこなうことができる.

3　混合ガウスモデルにおける最適解探索効率の検証

3.1　混合ガウスモデル (GMM)

混合ガウスモデル (GMM) とは, 複数の正規分布が重なり合ったモデルのことで, とても幅広い分野で応用されているモデルである. GMM のグラフィカルモデルを図 3.1 に示した. GMM のパラメータは, 正規分布の平均パラメータ $\boldsymbol{\mu}$, 精度パラメータ $\boldsymbol{\Lambda}$, そして観測することのできない潜在変数 $\mathbf{Z}$ が用いられる. 観

測データベクトルを $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}\}$, 対応する潜在変数を $\mathbf{Z} = \{\mathbf{z_1}, \mathbf{z_2}, \cdots, \mathbf{z_n}\}$ として, これらのパラメータが与えられた時, 観測データベクトルの条件付き分布は以下のように表される. ただし, $\mathbf{\Lambda}$ は分散の逆数である.

$$p(\mathbf{x}|\mathbf{Z}, \boldsymbol{\mu}, \mathbf{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K} N(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{\Lambda}_k^{-1})^{z_{nk}} \tag{3}$$

ここで,

$$N(\mathbf{x}_n|\boldsymbol{\mu}_k, \mathbf{\Lambda}_k^{-1}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{(|\mathbf{\Lambda}^{-1}|)^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\} \tag{4}$$

である. さらに, 混合比 $\boldsymbol{\pi}$ が与えられた時の $\boldsymbol{Z}$ の条件付き分布は次のように表すことができる.

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{5}$$

混合比 $\boldsymbol{\pi}$ はクラスタ数 K と同じ要素数のベクトルであり, データに対して, それぞれの混合分布がどの程度負担しているかの割合となっており, クラスタ数を K とすると以下の条件を満たす.

$$\sum_{k=1}^{K} \pi_k = 1 \tag{6}$$

さらに潜在変数 $\boldsymbol{Z}$ は, K 次元の二値確率変数である.
次に，各パラメータに対して事前分布を導入する．内容は [9] を参照している.

平均パラメータ $\boldsymbol{\mu}$, 精度パラメータ $\mathbf{\Lambda}$, 混合比 $\boldsymbol{\pi}$ それぞれに事前分布を仮定する. $\boldsymbol{\pi}$ の事前分布には $\boldsymbol{\alpha}_0 = \{\alpha_1, \alpha_2, \cdots, \alpha_K\}$ のディリクレ分布を仮定する.

$$p(\boldsymbol{\pi}) = Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) \tag{7}$$

同様に, パラメータ $\boldsymbol{\mu}, \mathbf{\Lambda}$ をもつ正規分布については, 分布の自由度パラメータ ν_0，D × D の尺度行列 $\boldsymbol{W}_0$ のガウス-ウィシャート分布を導入する.

$$\begin{aligned} p(\boldsymbol{\mu}, \mathbf{\Lambda}) &= p(\boldsymbol{\mu}|\mathbf{\Lambda})p(\mathbf{\Lambda}) \\ &= \prod_{k=1}^{K} N(\boldsymbol{\mu}_k|\boldsymbol{m}_0, (\beta_0\mathbf{\Lambda}_k)^{-1}) W(\mathbf{\Lambda}_k|\boldsymbol{W}_0, \nu_0) \end{aligned} \tag{8}$$

3.1.1 事後分布の分解仮定

GMM で変分ベイズを適用するために全ての確率変数の同時分布を書き下す.

$$p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda}) = p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \mathbf{\Lambda})p(\boldsymbol{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\mathbf{\Lambda})p(\mathbf{\Lambda}) \tag{9}$$

この同時分布は, 図 3.1 のグラフィカルモデルをもとに分解されることがわかる. 次に, $\boldsymbol{X}$ が観測されたときの潜在変数とパラメータに分解した変分近似を考える. 分解は以下のように仮定される.

$$q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda}) = q(\boldsymbol{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda}) \tag{10}$$

3.1.2 近似分布 $q(\boldsymbol{Z})$ の計算

式 (3.9) のような分解仮定をすることで計算可能な解を得ることができる. 各因子に対応する逐次更新式は, カルバックライブラーダイバージェンスの最小化によって得られる一般的な形を利用することで得ることができる. 具体的には, 最適解を得たい因子以外の因子で期待値をとることで最適解を得ることができる. まず, 因子 $q(\boldsymbol{Z})$ についての更新式を導出する. 最適な因子分布の対数は以下のようになる.

$$\ln q^*(\boldsymbol{Z}) \propto \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\mathbf{\Lambda}}[\ln p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda})] \tag{11}$$

$$= \mathbb{E}_{\boldsymbol{\pi}}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})] + \mathbb{E}_{\boldsymbol{\mu},\mathbf{\Lambda}}[\ln p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \mathbf{\Lambda})] \tag{12}$$

式 (3.10) から式 (3.11) の式変形は, $\boldsymbol{Z}$ に依存しない項を定数に吸収させている. 式 (3.11) に式 (3.1), 式 (3.3) を代入して計算すると, 以下のようになる.

$$\ln q^*(\boldsymbol{Z}) \propto \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \omega_{nk} \tag{13}$$

ただし, $\ln \omega_{nk}$ は次のような式である. また, 式中の D はデータの次元である.

$$\ln \omega_{nk} = \quad \mathbb{E}[\ln \pi_k] + \tfrac{1}{2}\mathbb{E}[\ln |\mathbf{\Lambda}_k|] - \tfrac{D}{2}\ln(2\pi) \tag{14}$$

$$-\tfrac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\mathbf{\Lambda}_k}[(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \mathbf{\Lambda}_k (\mathbf{x}_n - \boldsymbol{\mu}_k)] \tag{15}$$

式 (3.12) の両辺の指数をとって正規化すると,

$$q^*(\boldsymbol{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} {r_{nk}}^{z_{nk}} \tag{16}$$

となる. ただし,

$$r_{nk} = \frac{\omega_{nk}}{\sum_{j=1}^{K} \omega_{nj}} \tag{17}$$

式 (3.15) より, r_{nk} は負担率を計算していることがわかる. これは, EM アルゴリズムでいうところの, E ステップをおこなっている.

3.1.3 近似分布 $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda})$ の分解

$q(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda})$ を求める前に, 前の節で求めた負担率 r_{nk} をもとに計算できる 3 つの統計量を定義する. これは, 後の計算式を簡単にするためである.

$$N_k = \sum_{n=1}^{N} r_{nk} \tag{18}$$

$$\bar{x}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk} x_n \tag{19}$$

$$\boldsymbol{S}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^T \tag{20}$$

次に, 近似分布 $q(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda})$ について考える. これは $q(\boldsymbol{Z})$ と同様で, 考えている因子以外の因子で期待値をとることで最適化することができる. 最適化な分布 $q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda})$ の対数は以下のようになる.

$$\ln q^*(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}) \propto \ln p(\boldsymbol{\pi}) + \sum_{k=1}^{K} \ln p(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k) + \mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})] + \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}[z_{nk}] \ln N(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}) \quad (21)$$

この式の右辺について, $\boldsymbol{\pi}$ と $\boldsymbol{\mu}, \boldsymbol{\Lambda}$ が別々の項に分解されていることがわかる. また, $\boldsymbol{\mu}, \boldsymbol{\Lambda}$ が k 個の和で表されていることより, 指数をとると k 個の積になることがわかる. 以上の議論により, $q(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda})$ は以下のように分解される.

$$q(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}) = q(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k) \quad (22)$$

3.1.4 近似分布 $q(\boldsymbol{\pi})$ の計算

式 (3.20) の $q(\boldsymbol{\pi})$ について, 最適な分布を求める. 式 (3.19) から $\boldsymbol{\pi}$ に関係する項を取り出すと $\ln q^*(\boldsymbol{\pi})$ は,

$$\ln q^*(\boldsymbol{\pi}) \propto \ln p(\boldsymbol{\pi}) + \mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})] \quad (23)$$

右辺第 1 項は式 (3.5), 第 2 項は式 (3.3) を代入して期待値を計算すると,

$$\ln q^*(\boldsymbol{\pi}) \propto (\alpha_0 - 1)\sum_{k=1}^{K} \ln \pi_k + \sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk} \ln \pi_k \quad (24)$$

両辺の指数をとると, ディリクレ分布になる.

$$q^*(\boldsymbol{\pi}) = Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}) \quad (25)$$

ここで, $\boldsymbol{\alpha}$ の k 番目の要素 α_k は $\alpha_k = \alpha_0 + N_k$ である.

3.1.5 近似分布 $q(\boldsymbol{\mu},\boldsymbol{\Lambda})$ の計算

$q(\boldsymbol{\pi})$ の計算と同様, 式 (3.20) から $\boldsymbol{\mu}, \boldsymbol{\Lambda}$ に関係する項を取り出して計算する. そうすると, $\ln q^*(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)$ は,

$$\ln q^*(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k) \propto \sum_{k=1}^{K} \ln p(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k) + \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}[z_{nk}] \ln N(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}) \quad (26)$$

ここで, 式 (3.25) 第 1 項について, 式 (3.6) と同じ分解を考えて, $p(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k) = p(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)p(\boldsymbol{\Lambda}_k)$ より,

$$\ln q^*(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k) \propto \sum_{k=1}^{K}\{\ln p(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) + \ln p(\boldsymbol{\Lambda}_k)\} + \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}[z_{nk}] \ln N(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}) \quad (27)$$

式 (3.26) に式 (3.6) を代入して両辺の指数をとり整理すると, この近似分布はガウス-ウィシャート分布となる.

$$q^*(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k) = N(\boldsymbol{\mu}_k|\boldsymbol{m}_k,(\beta_k\boldsymbol{\Lambda}_k)^{-1})W(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k,\nu_k) \quad (28)$$

各パラメータの更新式は以下のように定義される.

$$\beta_k = \beta_0 + N_k \quad (29)$$

$$\boldsymbol{m}_k = \frac{1}{\beta_k}(\beta_0\boldsymbol{m}_0 + N_k\bar{x}_k) \quad (30)$$

$$\boldsymbol{W}_k^{-1} = \boldsymbol{W}_0^{-1} + N_k\boldsymbol{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{x}_k - \boldsymbol{m}_0)(\bar{x}_k - \boldsymbol{m}_0)^T \quad (31)$$

$$\nu_k = \nu_0 + N_k \quad (32)$$

これらの更新式は, EM アルゴリズムでいうところの, M ステップの計算をしている. 最後に, 式 (3.26) 中の $\mathbb{E}[z_{nk}]$ の具体的な計算であるが, これは式 (3.13) の $\ln \omega_{nk}$ を正規化することで得られる.

以上より, GMM における変分ベイズのアルゴリズムは以下の Algorithm1 のようになる.

Algorithm 1 GMM における変分ベイズアルゴリズム

Require: Data $\{x_n\}_{n=1}^N$, iteration number T
  Initialize distributions $q(\pi), q(Z), q(\mu,\Lambda)$
  for $i = 1$ to T **do**
    Update $q(Z)$ by Eq.(16)
    Update $q(\pi), q(\mu,\Lambda)$ by Eq.(25), (28)
  end for

3.2 期待値のモンテカルロ近似の導入 (GMM)

期待値のモンテカルロ近似の手法を変分ベイズに導入する. 上でも述べたように, 変分ベイズでも EM アルゴリズムと同じ手続きをおこなって近似分布を最適化するので, E ステップにおいて負担率を確率的 EM アルゴリズムと同様の考え方でハードに再割り当てをすることは可能である. 実際に, David ら [10] は, トピックモデルにおいてモンテカルロ近似を用いている. 変分ベイズに負担率の近似を導入した手法のアルゴリズムは以下の Algorithm2 のようになる.

Algorithm 2 変分ベイズに期待値のモンテカルロ法を導入した手法

Require: Data $\{x_n\}_{n=1}^N$, iteration number T
  Initialize distributions $q(\pi), q(Z), q(\mu,\Lambda)$
  for $i = 1$ to T **do**
    for all n **do**
      Choose random element r_{nk}
      $r_{nj} \leftarrow \delta_{j=k}$ for all $j = 1$ to K
    end for
    Update $q(Z)$
    Update $q(\pi), q(\mu,\Lambda)$
  end for

3.3 確率的勾配降下法 (SGD) の導入 (GMM)

変分ベイズにおいて SGD を導入する. Hoffman ら [11] は, LDA というモデルにおいて, 変分ベイズに確率的勾配降下法を導入した手法 (SVI) を提案した. また, Foulds ら [12] は, LDA において, SVI を発展させた手法を提案している. このように, SVI は活発に研究され有用な手法であることがわかる. また, Chris [13] は, GMM において SVI を用いた計算手法を示している. 以下の式の展開は, [13] をもとにしている.

SVI では, まずデータをランダムに抽出する. 抽出するデータ

サイズは 1 つ, もしくは複数である. 複数のデータを抽出する場合, 抽出したデータサイズをバッチサイズという. そして, E ステップと M ステップは通常通りおこない, 最後にステップサイズ ρ を用いて M ステップで更新したパラメータを更新する. t 回目のイテレーションにおいて, ρ は次のように更新する.

$$\rho^{(t)} = (t+\tau)^{-\kappa} \tag{33}$$

τ と κ はハイパーパラメータである.
更新する近似分布のパラメータをまとめて $\boldsymbol{\lambda}$, t 回目のイテレーションにおいて M ステップで更新されたパラメータを $\hat{\boldsymbol{\lambda}}$ とすると更新式は以下のようになる.

$$\boldsymbol{\lambda}^{(t)} = (1-\rho_t)\boldsymbol{\lambda}^{(t-1)} + \rho_t \hat{\boldsymbol{\lambda}} \tag{34}$$

3.4 実　　験

3.2 節から 3.4 節で述べたそれぞれの手法を用いて, 最適解探索に関する実験をおこなった.

3.5 実 験 方 法

それぞれのモデルのデータへの当てはまりの良さを示す指標として対数尤度を用いる. 対数尤度は対象としているモデルがどれだけデータに適合できているかを示す指標であり, データがそれぞれ独立に生成されると仮定すると次のように表すことができる.

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \ln p(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{35}$$

$\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}\}$ であり, $\boldsymbol{\theta}$ はパラメータベクトルである. GMM におけるパラメータベクトルは, $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}\}$ である.

イテレーション回数は 1,000 回で固定とする. 収束判定を用いる場合, 各手法で収束する速さが異なるため比較する際に不都合である. イテレーション回数を固定すれば, 各手法を同じ条件で比較することができる. また，対数尤度の他に，パラメータ推定をおこないその推定したパラメータ行列のフロベニウス距離を求め，正解パラメータのフロベニウス距離と比較する.

3.5.1 実験データ

実験に用いるデータは, 3 つの等方的な正規分布が重なった 2 次元混合分布からランダムに生成したデータを使う. 生成した分布の平均パラメータは, [0, 0], [4, 4], [7, 1] とし, 分散パラメータは, 対角成分が全て 1.0 の対角行列を用いる. また, 分散パラメータは, 1.0 以外に 3.0, 5.0 と値を変えて実験をおこなった. また, データサイズは, 300 件, 3,000 件, 30,000 件の 3 種類である. この実験は, 分散がそれぞれ 1.0, 3.0, 5.0 の場合で異なるデータを用いて 5 回ずつ実験をおこなった.

3.5.2 実 験 結 果

実験をおこなった結果, 各分散における平均対数尤度の値は表 1 のようになった. また, 表について, VI は変分ベイズ, VI+SEM は変分ベイズに期待値のモンテカルロ法を合わせた手法, VI+SGD は変分ベイズに SGD を合わせた手法を表している. 縦軸はデータの件数を表している. 表 1 に記載されている対数尤度の値は, 5 種類の異なるデータに対して対数尤度を求め, それらの値を平均した値となっている.

表 1　各分散における各手法の平均対数尤度 (GMM)

分散	データ数	VI	VI+SEM	VI+SGD
1.0	300 件	-4.0256	-4.0246	-4.4453
	3,000 件	-3.9274	-3.9024	-4.6508
	30,000 件	-3.9064	-3.9304	-4.6662
3.0	300 件	-4.9067	-4.8740	-5.0416
	3,000 件	-4.8618	-4.8618	-4.8762
	30,000 件	-4.8608	-4.8608	-4.8650
5.0	300 件	-5.0861	-5.0840	-5.1969
	3,000 件	-5.0891	-5.0895	-5.2184
	30,000 件	-5.0927	-5.0920	-5.2241

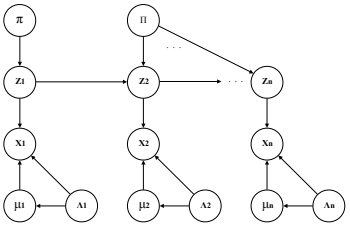

図 2　HMM のグラフィカルモデル

各値について見てみると, 分散が 1.0 かつデータが 3,000 件や分散 3.0 かつデータが 3,00 件の値では SEM+VI の手法が最も良い値となっている．一方で，VI + SGD の対数尤度はどの分散の値，データ件数を比較しても最も低い値となった.

次に，各手法において正解パラメータと推定パラメータの差を計算した値を下の表に示す．ここでは，データ数 30,000 件，分散 5.0 の場合についての結果となっている.

表 2　正解パラメータと推定パラメータのノルムの差 (GMM)

	μ	Λ
VI	0.3506	1.222
VI+SEM	0.0795	0.0191
VI+SGD	2.0551	16.3804

表の値から，SEM を用いた手法の推定結果が最も正解パラメータとの距離が近いという結果となった.

4 隠れマルコフモデル (HMM) における最適解探索効率の検証

4.1 隠れマルコフモデル (HMM)

天気データやセンサーデータなど, 系列データを扱う際に用いられるモデルが隠れマルコフモデル (HMM) である. HMM のグラフィカルモデルは図 4.1 のようになる. 観測データ $\boldsymbol{x}_i$ はそれぞれに対応する潜在変数 z_i, $\boldsymbol{\mu}_i$, $\boldsymbol{\Lambda}_i$ から生成される. GMM では潜在変数 $\boldsymbol{Z}$ を導入していたが, HMM でも同様に潜在変数を導入する. しかし, HMM では, 各潜在変数 z_n はそれぞれ前の状態の潜在変数に依存するという仮定をおく. そうすると, 遷移確率 $\boldsymbol{\Pi}$ が与えられたときの潜在変数 $\boldsymbol{Z}$ の条件付き確率は以下のようになる.

$$p(\boldsymbol{Z}|\boldsymbol{\Pi}) = p(z_1|\boldsymbol{\pi}) \prod_{n=2}^{N} p(z_n|z_{n-1}, \Pi_n) \tag{36}$$

遷移確率は, $\boldsymbol{\Pi}_2, ..., \boldsymbol{\Pi}_K$ を行にもつ K × K の行列である. $\boldsymbol{\pi}$ は初期確率である. 観測モデルは扱うデータの種類によって異なるが, 本研究では観測データは連続量とする. したがって, 観測モデルは正規分布とする.

$$p(\mathbf{x}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N} \prod_{k=1}^{K} N(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{nk}} \tag{37}$$

ここで, 正規分布は式 (3.2) の通りである.

4.2 HMM における変分ベイズ

4.2.1 事前分布の導入

HMM も GMM と同様に, パラメータに分布を仮定することで変分ベイズを用いることができる. 遷移確率 Π には, ディリクレ分布を仮定する. 式は,

$$p(\boldsymbol{\Pi}) = \prod_{k=1}^{K} Dir(\boldsymbol{\Pi}_k|\boldsymbol{\alpha}_k^0) \tag{38}$$

$$\boldsymbol{\alpha}_k^0 = \{\boldsymbol{\alpha}_{k,1}^0, ..., \boldsymbol{\alpha}_{k,K}^0\} \tag{39}$$

となる. また, $\boldsymbol{\mu}$ と $\boldsymbol{\Lambda}$ の同時分布 $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ は, GMM と同様の事前分布なので,

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \\ &= \prod_{k=1}^{K} N(\boldsymbol{\mu}_k|\boldsymbol{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1}) W(\boldsymbol{\Lambda}_k|\boldsymbol{W}_0, \nu_0) \end{aligned} \tag{40}$$

4.2.2 事後分布の分解仮定

まず, 全ての確率変数の同時分布の分解は以下のようになる.

$$p(\boldsymbol{X}, \boldsymbol{\Pi}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{Z}|\boldsymbol{\Pi}) p(\boldsymbol{\Pi}) p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) p(\boldsymbol{\Lambda}) \tag{41}$$

次に, データ $\boldsymbol{X}$ が観測されたときの変分近似を考える. 分解は以下のように仮定する.

$$q(\boldsymbol{Z}, \boldsymbol{\Pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{Z}) q(\boldsymbol{\Pi}) q(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \tag{42}$$

$$= q(\boldsymbol{Z}) \prod_{k=1}^{K} q(\boldsymbol{\Pi}_k) \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \tag{43}$$

4.2.3 近似分布 $q(\boldsymbol{Z})$ の計算

GMM において潜在変数の近似分布を最適化したときと同様に, HMM で扱う潜在変数 $\boldsymbol{Z}$ の近似分布は, $\boldsymbol{Z}$ 以外の変数で期待値を計算することにより最適化することができる.

$$\begin{aligned} \ln q^*(\boldsymbol{Z}) &\propto \mathbb{E}_{\boldsymbol{\Pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}}[\ln p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] \\ &= \mathbb{E}_{\boldsymbol{\Pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}}[\ln p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{Z}|\boldsymbol{\Pi}) p(\boldsymbol{\Pi}) p(\boldsymbol{\mu}|\boldsymbol{\Lambda}) p(\boldsymbol{\Lambda})] \\ &= \mathbb{E}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}}[\ln p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \mathbb{E}_{\boldsymbol{\Pi}}[\ln p(\boldsymbol{Z}|\boldsymbol{\Pi})] \end{aligned} \tag{44}$$

式 (4.11) の計算を進め, 両辺の指数をとると以下のようになる. なお, 以下の式の表記は部分的に [14] を用いている.

$$q^*(\boldsymbol{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} (b_{n,k}) \prod_{n=1}^{N} \prod_{k=1}^{K} \prod_{j=1}^{K} (a_{k,j})^{z_{n,k} z_{(n+1),j}} \tag{45}$$

ただし, $b_{n,k}, a_{k,j}$ は以下である.

$$a_{k,j} = \exp(\ln \mathbb{E}[\ln \pi_{k,j}]) = \Psi(\alpha_{k,j}) - \Psi(\sum_{j=1}^{K} \alpha_{k,j}) \tag{46}$$

$$b_{n,k} = \exp(\ln \mathbb{E}[p(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)]) \tag{47}$$

となる. 式 (4.13) の最右辺 Ψ はディガンマ関数を表している. 詳細は [9] などを参照されたい.

4.2.4 近似分布 $q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ の計算

$q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ も $q(\boldsymbol{Z})$ と同様に潜在変数 $\boldsymbol{Z}$ 以外の変数で期待値を計算することにより求めることができる. 計算結果は, GMM で導出したときと同様, ガウス-ウィシャート分布となる.

$$q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = N(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta_k\boldsymbol{\Lambda}_k)^{-1}) W(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, \nu_k) \tag{48}$$

4.2.5 近似分布 $q(\boldsymbol{\Pi})$ の計算

変数 Π 以外で期待値を計算すると, ディリクレ分布となる.

$$q^*(\boldsymbol{\Pi}) = Dir(\boldsymbol{\Pi}|\boldsymbol{\alpha}) \tag{49}$$

4.3 E ステップの計算

GMM の E ステップでは負担率を計算していた. HMM の E ステップでも同様に負担率を計算する. 負担率 $r_{n,k}$ は以下のような式になる.

$$r_{n,k} = \frac{\alpha_{n,k}\beta_{n,k}}{\sum_{k=1}^{K} \alpha_{n,k}\beta_{n,k}} \tag{50}$$

式 (4.15) の $\alpha_{n,k}$ と $\beta_{n,k}$ は, Baum-Welch アルゴリズムを用いることで表すことができる. 導出方法は, [15] や [16] を利用している. [15] では, 正規分布ではなく, ポアソン分布を用いた導出方法であるが, 負担率の計算方法は同じである. $\alpha_{n,k}$ と $\beta_{n,k}$ は以下のように表される.

$$\alpha_{n,k} = b_{n,j} \sum_{k=1}^{K} a_{n,k}\alpha_{n-1,k} \tag{51}$$

$$\beta_{n,k} = \sum_{k=1}^{K} \beta_{n+1,k} a_{n,k} b_{n+1,k} \tag{52}$$

式 (4.15), (4.16), (4.17) より, 負担率は再帰計算によって求められる.

4.3.1 M ステップの計算

E ステップと同様, GMM と同じように計算する. まず, あとの計算を簡単にするため, 3 つの統計量を定義する. これは, 実際には GMM と同じ式で表すことができる. よって, 式 (3.16) から式 (3.18) を HMM の M ステップでも用いる.

さらに, HMM における観測データには正規分布を仮定しているので, HMM で更新する正規分布のパラメータも GMM で示した式と同様の式になる. よって, ガウス-ウィシャート分布の更新式は GMM と全く同じ式になり, 式 (3.28) から式 (3.31) によって表される.

また, 遷移確率 $\boldsymbol{\Pi}$ のディリクレ分布 $q(\boldsymbol{\Pi})$ の更新式は以下のようになる.

$$\alpha_{k,j} = \alpha_{k,j}^{(0)} + \sum_{n=1}^{N-1} h(z_{n,k}, z_{(n+1),j}) \tag{53}$$

ただし, $h(z_{n,k}, z_{(n+1),j})$ は以下のように定義される.

$$h(z_{n,k}, z_{(n+1),j}) = \frac{\alpha_{(n-1),k} a_{n,k} b_{k,j} \beta_{n,j}}{\sum_{l=1}^{K} \sum_{m=1}^{K} \alpha_{(n-1),l} a_{l,m} b_{n,m} \beta_{n,m}} \tag{54}$$

4.4 期待値のモンテカルロ近似の導入 (HMM)

GMM と同様, HMM でも期待値を近似する. 式 (4.17) の計算によって求めた負担率を, 1 と 0 の 2 値に再割り当てする.

4.5 確率的勾配降下法 (SGD) の導入 (HMM)

HMM に対して SVI を導入する. この手法は, Foti ら [17] によって提案されている. まず, バッチサイズ S を定め, 全データから S だけデータをサンプリングする. E ステップではサンプリングされたデータを使って負担率を計算し, M ステップでは E ステップで求めた負担率から統計量を計算し, 近似分布のパラメータを更新する. また, M ステップにおける各パラメータの更新では, GMM と同様ステップサイズ ρ を用いる.

4.6 実　　験

4.2 節から 4.3 節について述べた手法について, GMM と同様の実験をおこなった.

4.6.1 実 験 方 法

GMM 同様, データへの当てはまりの良さの指標として対数尤度を計算する. 式は (3.37) のようになる. また, イテレーション回数は 1,000 回で固定とする. さらに，正解パラメータと推定パラメータのフロベニウス距離を比較する.

本実験で用いるパラメータは, 平均パラメータ, 分散パラメータ, 遷移確率である. 正解の平均パラメータは以下のような行列である.

$$\boldsymbol{\mu} = \begin{pmatrix} 0.0 & 0.0 \\ 0.0 & 11.0 \\ 9.0 & 10.0 \\ 11.0 & -1.0 \end{pmatrix}$$

さらに, 正解の分散パラメータは以下のような対角行列である. また, 分散は 1.0 の要素が 3.0, 5.0 の場合の合計 3 種類で実験をおこなった.

$$\boldsymbol{\Lambda} = \begin{pmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0.1 \end{pmatrix}$$

最後に遷移確率 $\boldsymbol{\Pi}$ は次のような値を持つ行列である.

$$\boldsymbol{\Pi} = \begin{pmatrix} 0.7 & 0.2 & 0.0 & 0.1 \\ 0.3 & 0.5 & 0.2 & 0.0 \\ 0.0 & 0.3 & 0.5 & 0.2 \\ 0.2 & 0.0 & 0.2 & 0.6 \end{pmatrix}$$

データサイズは 300 件, 3,000 件, 30,000 件の 3 種類である.

4.6.2 実 験 結 果

実験の結果, 平均対数尤度は表 3 のようになった.

表を見てみると, 大部分の分散の値, データ件数において SEM+VI が最も良い値を示し，VI+SGD が最も低い値を示した.

表 3　各分散における各手法の平均対数尤度 (HMM)

分散	データ数	VI	VI+SEM	VI+SGD
1.0	300 件	-3.9238	-3.7750	-4.2346
	3,000 件	-3.7712	-3.7708	-3.7781
	30,000 件	-3.7665	-3.7664	-3.7774
3.0	300 件	-4.9067	-4.8740	-5.0416
	3,000 件	-4.8618	-4.8618	-4.8762
	30,000 件	-4.8608	-4.8608	-4.8650
5.0	300 件	-5.3593	-5.3363	-5.5239
	3,000 件	-5.3503	-5.3503	-5.3635
	30,000 件	-5.3494	-5.3496	-5.3522

次に，正解パラメータと推定パラメータの距離の差を求めた値を次の表に示す．また，データ数，分散の値は GMM と同じである.

表 4　正解パラメータと推定パラメータのノルムの差 (HMM)

	μ	Λ
VI	0.0050	0.0558
VI+SEM	0.0004	0.0249
VI+SGD	0.3099	0.1009

表より，SEM を用いた手法によって推定されたパラメータが最も正解パラメータに近いことがわかる.

5 結　　論

本研究では, 変分ベイズにおいて, 摂動を加える各手法の最適解探索効率を GMM と HMM の 2 つのモデルにおいて検証した. GMM においては, 部分的に SEM が対数尤度, パラメータ推定の両方で最も良い性能を示した. これは, 通常の変分ベイズでは, 対数尤度がある値で収束してしまい局所最適解に陥ってしまうが, SEM では期待値をランダムに近似することで摂動が加わり, 局所最適解をうまく抜け出すことができたからである. 一方で SGD は，SEM 同様通常の変分ベイズよりも良い性能を示すかと思われたが, 比較した手法の中で最も性能が低い結果となった. SGD を用いた手法は，各イテレーションで全データから 1 つもしくは複数データをランダムにサンプル抽出するが, 実験結果の平均対数尤度やパラメータ推定がうまくいっていないことから，SGD による摂動の加え方が大きすぎると言える. SEM は毎回全データを用いて学習するが, SGD は毎回異なるデータを抽出して学習するので SEM に比べてうまく学習が進まないのである. 同様の結果が HMM でも示された. HMM では，大部分の実験設定において SEM を用いた手法が最も良い値を示し, SGD を用いた手法が最も性能の低い結果となった. GMM の場合と同様の理由から SGD はうまく学習をおこなうことができなかったと言える.

2 つのモデルにおいて変分ベイズの各手法を検証した結果, 摂動が比較的小さい SEM は性能がよく, 摂動を大きく加える SGD は加えない手法よりも性能が低いことがわかった. SGD は多層ニューラルネットワークなど様々な手法で応用されているとて

も広く知られた手法であり, その手法の有効性は数多くの文献によって示されている. しかし, 変分ベイズにおいて局所最適解を抜けだすための手法として考察してみると, 本研究で示したように SGD よりも良い性能を示す手法が存在する. SGD は全データを使って学習することが計算量的に困難である際に用いると効果を発揮するが, 全データを使って学習をおこなうことができる際には, SEM のような摂動をあまり加えない手法を用いる方が良い.

今後の課題としては, 各手法の実行時間を計測し比較することである．SGD は大量のデータから一部のデータを取り出す操作をおこなうことにより学習を高速化する効果があるため，実行時間において比較すると良い結果となる可能性がある．また，本研究では，GMM と HMM という比較的扱いやすいモデルを対象に実験をおこなったが，さらに複雑なモデルを対象として実験し，最適解探索効率に関する一般的な結論を得る必要がある．さらに, SEM を拡張した手法である MCEM や SEM を改良した SAEM を変分ベイズに導入し, 摂動の加わり方を調べる必要もある.

謝　　辞

本研究の一部は，JSPS 科研費（課題番号 JP16H02904, JP19K20333）の助成によって行われた．

文　　献

[1] Zoubin Ghahramani and Matthew J. Beal. Variational inference for bayesian mixtures of factor analysers. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pp. 449–455. MIT Press, 2000.

[2] ZOUBIN GHAHRAMANI. An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 15, No. 01, pp. 9–42, 2001.

[3] Noam Shental, Aharon Bar-Hillel, Tomer Hertz, and Daphna Weinshall. Computing gaussian mixture models with em using equivalence constraints. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pp. 465–472, 2003.

[4] Padhraic Smyth. Clustering sequences with hidden markov models. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, NIPS'96, pp. 648–654, 1996.

[5] Sam Smith and Quoc V. Le. A bayesian perspective on generalization and stochastic gradient descent. 2018.

[6] Valentin I. Spitkovsky, Hiyan Alshawi, Daniel Jurafsky, and Christopher D. Manning. Viterbi training improves unsupervised dependency parsing. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pp. 9–17, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[7] Didier Chauveau Gilles Celeux and Jean Diebolt. On stochastic versions of the em algorithm. Technical report, INRIA, 1995.

[8] 修功上田, 良平中野. 確定的アニーリング em アルゴリズム. 電子情報通信学会論文誌. D-2, 情報・システム 2-情報処理, Vol. 80, No. 1, pp. 267–276, jan 1997.

[9] C.M. Bishop. パターン認識と機械学習 (下). 丸善, 2012.

[10] David M. Mimno, Matthew D. Hoffman, and David M. Blei. Sparse stochastic inference for latent dirichlet allocation. In *Proc. ICML*, 2012.

[11] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

[12] James Foulds, Levi Boyles, Christopher DuBois, Padhraic Smyth, and Max Welling. Stochastic collapsed variational bayesian inference for latent dirichlet allocation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pp. 446–454, 2013.

[13] In-Depth Variational Inference Tutorial. `https://chrisdxie.files.wordpress.com/2016/06/in-depth-variational-inference-tutorial.pdf`.

[14] C. A. McGrory and D. M. Titterington. Variational bayesian analysis for hidden markov models. *Australian & New Zealand Journal of Statistics*, Vol. 51, No. 2, pp. 227–244, 2009.

[15] 須山敦志. ベイズ推論による機械学習入門. 講談社, 2017.

[16] Christian Gruhl and Bernhard Sick. Variational bayesian inference for hidden markov models with multivariate gaussian output distributions. 05 2016.

[17] Nick Foti, Jason Xu, Dillon Laird, and Emily Fox. Stochastic variational inference for hidden markov models. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pp. 3599–3607. Curran Associates, Inc., 2014.