科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２９　日現在

機関番号：　１７６０１
研究種目：　基盤研究(C)
研究期間：　2010～2013
課題番号：　２２５００２０６

研究課題名（和文）高次元データに対する事後確率分布構造の解析

研究課題名（英文）On the structure of the posterior probability distribution in high-dimensional data

研究代表者
伊達　章（Date, Akira）

宮崎大学・工学部・准教授

研究者番号：６０３２２７０７

交付決定額（研究期間全体）：（直接経費）　　3,200,000円、（間接経費）　　960,000円

研究成果の概要（和文）：画像，音声，自然言語，塩基配列などの大規模データに対し，確率モデルを構築し，認識，予測など確率推論を行なうことが，計算機性能の急速な向上に伴い可能になっている．ベイズ推論の本質の一つは，データを観測した後の事後確率分布の利用にあるが，その分布の構造については不明な点が多く，分布が奇妙な構造をもつことは十分考えられる．本研究では，事後確率分布から多数のサンプルを生成することで，事後確率分布の構造を反映した意味のある推定量を求める手法を開発した．単純な隠れマルコフモデル，格子型マルコフ確率場を用いて計算機実験をおこない，本手法の有効性を確認した．

研究成果の概要（英文）：Probabilistic generative models work in many applications of image analysis and speech recognition. In general, there is an observation vector y and a state vector x, and a joint dependency structure among them. The object of interest is, given y, the most meaningful configuration x and the posterior distribution Pr(x|y). In practice, the structure of the posterior distribution Pr(x|y) is hard to know, and it might have a peculiar structure, especially when x is high dimensional vector. In this project, we developed a method which finds a meaningful estimator by generating a large number of samples from posterior distribution. We performed computer experiments of simple hidden Markov models in which the various functions of the posterior probability distribution is obtainable. Based on the experiments, the effectiveness of the method was discussed.

研究分野：　総合領域

科研費の分科・細目：　情報学・感性情報学・ソフトコンピューティング

キーワード：　確率的情報処理　確率モデル　ベイズ推論　確率推論　事後確率　隠れマルコフモデル　マルコフ確率場

1. 研究開始当初の背景

物事の確率的な関係に着目し，観測できるデータから観測できない変数の値を推論する方法を確率推論という．計算機性能の急速な向上に伴い，画像，音声，自然言語，塩基配列などの大規模データの認識，予測などに対し，確率推論・ベイズ推論が利用されるようになってきていた．ここで，確率推論を行う場合には，あらかじめ対象の規則性を反映した確率分布を構築しておく必要がある．観測データを $\boldsymbol{y}$，解釈用の内部変数（隠れ変数，潜在変数ともよばれる）を $\boldsymbol{x}$ とすると，確率分布 $\{p(\boldsymbol{x})\}$ と $\{p(\boldsymbol{y}|\boldsymbol{x})\}$ の設計が重要になる．データ $\boldsymbol{y}$ を生成する能力を持つものは確率的生成モデルとよばれており，データ $\boldsymbol{y}$ から $\boldsymbol{x}$ を推定する際に威力を発揮する．例えば，入力 $\boldsymbol{y}$ に対し事後確率 $p(\boldsymbol{x}|\boldsymbol{y})$ を最大化する

$$\boldsymbol{x}_{\mathrm{MAP}} \quad = \quad \arg\max_{\boldsymbol{x}} \; p(\boldsymbol{x}|\boldsymbol{y})$$

を求める最適化問題として定式化される場合が多い．別の推定量がよい場合もあるが，ともかく，ベイズ推論の本質は事後確率分布 $\{p(\boldsymbol{x}|\boldsymbol{y})\}$ の利用にある．では，事後分布 $\{p(\boldsymbol{x}|\boldsymbol{y})\}$ はどのような構造を持つであろうか．事後確率分布の構造は，対象をどうモデル化するかに依存するため，一般的な議論はできない．しかし，単純なモデルの場合でさえ，その構造はよく分かっていない．この問は難問であるが，研究開始当初，少しでもこの問に答えることが期待されていた．

2. 研究の目的

問題の難しさを考えてみよう．$\boldsymbol{x}$ が n 次元で，その各要素が 2 値 0,1 をとるとする．各 $\boldsymbol{x}$ についての確率 $p(\boldsymbol{x})$ を決めれば，確率分布 $\{p(\boldsymbol{x})\}$ が定まる．理想的には，事前分布 $\{p(\boldsymbol{x})\}$ は，多数の場所にピークをもつように（図 1 左），一方，観測データ $\boldsymbol{y}$ が与えられた後の事後分布 $\{p(\boldsymbol{x}|\boldsymbol{y})\}$ は，ある特定の場所 $\boldsymbol{x}^*$ にピークをもつ分布であるように確率モデルを設計したい（図 1 右）．

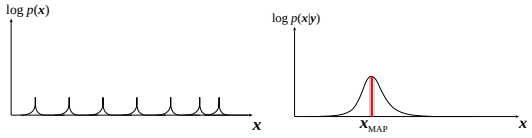

図 1

$\boldsymbol{x}$ が高次元ベクトルの場合，2^n-1 個の値を個別に指定することは現実的には不可能なため，通常，少数のパラメータを用い，高次元変数の確率分布をモデル化する．このとき，少数のパラメータで高次元変数の確率分布をモデル化するため，確率分布の構造に意図しない構造が埋め込まれてしまい，事後分布が不自然な構造を持つ可能性がある．

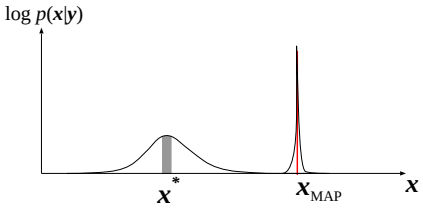

図 2

例えば，図 2 に示すような，事後分布が 2 つの山をもっている場合があり得る．ここで $\boldsymbol{x}_{\mathrm{MAP}}$ は，確かに事後確率を最大にする状態ではあるが，その近傍の状態に対する事後確率は比較的低い．一方，まったく別の場所に，なだらかな山が存在している．ここで，ある特定の状態 $\boldsymbol{x}$ に対し次の指標を考えよう．

$$b(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{y}) + \sum_{\boldsymbol{x}'} p(\boldsymbol{x}'|\boldsymbol{y})$$

ここで $\boldsymbol{x}'$ は，$\boldsymbol{x}$ の近傍の状態の集合とする．あらかじめ多数の $\boldsymbol{x}$ を事後分布 $p(\boldsymbol{x}|\boldsymbol{y})$ からサンプリングしておき，$b(\boldsymbol{x})$ を最大にする $\boldsymbol{x}$

$$\boldsymbol{x}^* \quad = \quad \arg\max_{\boldsymbol{x}} \; b(\boldsymbol{x})$$

を求めよう．これは $\boldsymbol{x}_{\mathrm{MAP}}$ とは別の推定量である．事後確率分布が図 2 に示す構造を持っていれば，$\boldsymbol{x}^*$ は，$\boldsymbol{x}_{\mathrm{MAP}}$ より，意味ある推定値である可能性が高い．

3. 研究の方法

(1) 事後確率分布の不自然な構造を検出する手法

本研究では，事後確率分布から多数のサンプルを生成することで，事後分布の構造を調べる新しい手法を提案した．$\boldsymbol{x}$ は高次元ベクトルであり，事後分布 $\{p(\boldsymbol{x}|\boldsymbol{y})\}$ の構造を知ることは容易ではない．ここでは，図 2 に示すような構造があった場合に，その存在を検出する手法を述べる．

(S1) 事後分布 $\{p(\boldsymbol{x}|\boldsymbol{y})\}$ にしたがう $\boldsymbol{x}^{\alpha}_{\mathrm{smp}}$ を生成する（$\alpha = 1, \cdots, m$）

(S2) $\boldsymbol{x}^{\alpha}_{\mathrm{smp}}$ と近傍関係にある状態 $\boldsymbol{x}^{\alpha'}_{\mathrm{smp}}$ を複数個生成する

(S3) 以下に示す $b(\boldsymbol{x}^{\alpha}_{\mathrm{smp}})$ の値を求める．

$$b(\boldsymbol{x}^{\alpha}_{\mathrm{smp}}) = p(\boldsymbol{x}^{\alpha}_{\mathrm{smp}}|\boldsymbol{y}) + \sum_{\boldsymbol{x}^{\alpha'}_{\mathrm{smp}}} p(\boldsymbol{x}^{\alpha'}_{\mathrm{smp}}|\boldsymbol{y})$$

(S4) $b(\boldsymbol{x}^{\alpha}_{\mathrm{smp}})$ を大きい順に並び替え，

$$b_{\max} \quad = \quad \max_{\alpha} b(\boldsymbol{x}^{\alpha}_{\mathrm{smp}})$$
$$\boldsymbol{x}^{*}_{\mathrm{smp}} \quad = \quad \operatorname*{argmax}_{\alpha} b(\boldsymbol{x}^{\alpha}_{\mathrm{smp}})$$

を求める

(S5) $b_{\max}$ と $p(\boldsymbol{x}_{\text{map}}|\boldsymbol{y})$ の大小を評価し，$b_{\max} > p(\boldsymbol{x}_{\text{map}}|\boldsymbol{y})$ かつ $\boldsymbol{x}^*_{\text{smp}}$ の近傍に $\boldsymbol{x}_{\text{map}}$ が含まれていない場合，図 2 に示す事後確率分布の構造が存在すると判断する．これは，$\boldsymbol{x}^*_{\text{smp}}$ と $\boldsymbol{x}_{\text{map}}$ のハミング距離を計算すれば判断できる．

事後確率分布構造を調べるために必要なサンプル $\boldsymbol{x}_{\text{smp}}$ は，動的計画法を用いた計算により，得ることができる．$\boldsymbol{x}_{\text{smp}}$ と近傍関係にある状態 $\boldsymbol{x}'_{\text{smp}}$ として，本研究では，隠れマルコフモデルの実験では，(a) $\boldsymbol{x}$ の要素 1 つを反転，(b) $\boldsymbol{x}$ の要素を無作為に 2 つ反転，(c) $\boldsymbol{x}$ の遷移している箇所をずらす，(d) $\boldsymbol{x}$ に矩形波を 1 つ挿入，という 4 種類の方法を試した．

(2) 数理モデル：隠れマルコフモデル

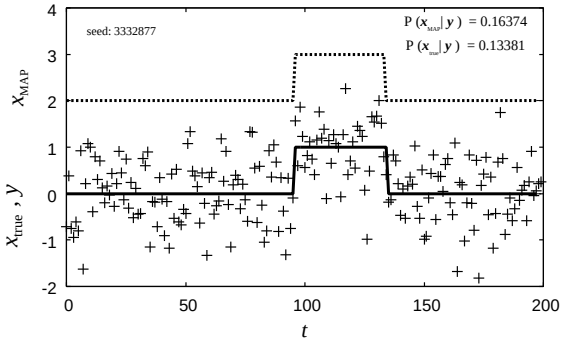

図 3

0,1 の 2 つの状態をとるマルコフ的情報源 $x_0, x_1, x_2, \cdots$ を考えよう．初期状態として x_0 が 0 をとる確率を 0.5，状態遷移行列として $\begin{bmatrix} 0.99 & 0.03 \\ 0.01 & 0.97 \end{bmatrix}$ を考える．初期状態を 0,1 のどちらかに決め，状態遷移を $N-1$ 回繰り返すと一つの信号 $\boldsymbol{x}_{\text{org}} = (x_0, x_1, \cdots, x_{N-1})$ が生成できる（図 3 の直線，$N = 200$）．この信号 $\boldsymbol{x}$ にノイズ $\boldsymbol{n}$ が加わった $\boldsymbol{y} = (y_0, y_1, \cdots)$ が観測される状況を考えよう．ここでノイズ $\boldsymbol{n}$ は各 x_i に独立に加わるとし（$y_i = x_i + n_i$），n_i は平均 0，分散 σ^2 の正規分布にしたがうとする．問題は $\boldsymbol{y}$（図 3，+）を観測し，もとの $\boldsymbol{x}$ を推定することである．ベイズ推論により元の信号を推定した結果 $\boldsymbol{x}_{\text{MAP}}$ を点線で（表示が重ならないよう $\boldsymbol{x}_{\text{MAP}} + 2$ を）表示している．

確率変数間の依存性グラフを図 4 に示す．このモデルは隠れマルコフモデルと呼ばれ様々な領域で利用されており，理論的に興味深い性質をもつ．

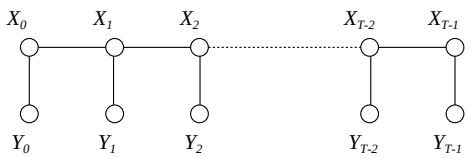

図 4

(3) 数理モデル：隠れマルコフ場

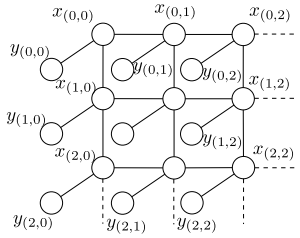

図 5

$K \times K$ 個のノードを持つ格子型マルコフ確率場を考える（図 5）．$\boldsymbol{x}$ は確率変数，$\boldsymbol{y}$ は観測データ，ノード間の枝が依存関係を示している．観測データ $\boldsymbol{y}$ については，$\boldsymbol{x}$ に正規分布のノイズが加わるものとする．隠れマルコフモデルの場合と同様に，事後確率分布から正確にサンプリングした $\boldsymbol{x}$ から近傍を生成し，b を計算する．近傍状態は，以下の 2 種類の方法で生成した．(a) $\boldsymbol{x}$ の要素 1 つを反転，(b) $\boldsymbol{x}$ の要素 2 つを反転．近傍の生成に関しては，どちらの方法も選択した要素を 0 なら 1，1 なら 0 と反転させた．要素 1 つを反転する場合，K^2 通りの近傍が得られる．また，要素を 2 つ反転する場合，可能な近傍状態は ${}_{K^2}\mathrm{C}_2$ 通りあり，K が大きくなると近傍数は膨大になる．そのため，近傍の数が 500 種類以上となる場合，${}_{K^2}\mathrm{C}_2$ 通りの中から 500 種類を無作為に選択して得ることにした．このように得た近傍から，$b_{\max}$ を計算し，事後確率分布構造を調べた．

4. 研究成果

(1) 主な成果

以下では，典型的な 2 例の解析結果を示す．$\boldsymbol{x}_{\text{org}}$ および $\boldsymbol{y}$ は，事前分布 $p(\boldsymbol{x})$ およびデータモデル $p(\boldsymbol{y}|\boldsymbol{x})$ を用い生成した（$\boldsymbol{x}$ の次元数 $N = 200$，ノイズの標準偏差 $\sigma = 0.7$，サンプルの生成数 $m = 10{,}000$）．

事後確率が自然な構造をもつ一例を図 6 に示す．横軸は遷移回数 N，縦軸は状態 $(0,1)$ の遷移を表す（縦軸の数字自体には意味がない）．$\boldsymbol{x}_{\text{org}}$ は元の系列，$\boldsymbol{y}_{\text{obs}}$ は $\boldsymbol{x}_{\text{org}}$ にノイズの加わった観測データ，$\boldsymbol{x}_{\text{map}}$ は推定により得られた最も尤もらしい系列，$\boldsymbol{x}_{\text{map-2nd}}$ は 2 番目に尤もらしい系列，$\boldsymbol{x}_{\text{smp}}$ は事後確率分布より得られたサンプル 10 例を表している．グラフ左側の値は $\boldsymbol{x}_{\text{org}}$，$\boldsymbol{x}_{\text{map}}$ の事後確率，及び $\boldsymbol{x}_{\text{smp}}$ の b の値である．また，d は $\boldsymbol{x}_{\text{map}}$ と $\boldsymbol{x}_{\text{map-2nd}}$，及び $\boldsymbol{x}_{\text{smp}}$ とのハミング距離を表している．また，表示が重ならないよう，$\boldsymbol{x}_{\text{map}}$，$\boldsymbol{x}_{\text{map-2nd}}$，$\boldsymbol{x}_{\text{smp}}$ をずらして出力している．

図 6 に示す例の場合，$p(\boldsymbol{x}_{\text{map}}|\boldsymbol{y}) = 0.07848$ であり，それぞれの $b_{\max}$ の値は，要素 1 つを反転させた近傍の場合は 0.22751，要素 2 つを反転させた近傍の場合は 0.14108，遷移箇所をずらした近傍の場合は 0.39378，矩形波を挿入した場合の近傍は 0.07943 と，すべて $b_{\max} > p(\boldsymbol{x}_{\text{map}}|\boldsymbol{y})$ となっていた．また，$\boldsymbol{x}^*_{\text{smp}}$ と $\boldsymbol{x}_{\text{map}}$ のハミング距離は，すべての場合で 0 となっているため，$\boldsymbol{x}^*_{\text{smp}}$ と $\boldsymbol{x}_{\text{map}}$ が一致しているこ

とが分かる．そのため，これらの結果から，事後確率分布は，図 1 右に示すような自然な構造をもつと推察できる．

図 7 に示す例の場合，$p(\boldsymbol{x}_{\mathrm{map}}|\boldsymbol{y}) = 0.05082$ である．$b_{\max}$ の値は，要素 1 つを反転させた近傍の場合は 0.14707，要素 2 つを反転させた近傍の場合は 0.08980，遷移箇所をずらした近傍の場合は 0.21915（図 7）と $b_{\max} > p(\boldsymbol{x}_{\mathrm{map}}|\boldsymbol{y})$ となっているが，$\boldsymbol{x}_{\mathrm{map}}$ とのハミング距離はそれぞれ 11，10，11，$\boldsymbol{x}_{\mathrm{map-2nd}}$ との距離も 12，11，12 であった．そのため，$\boldsymbol{x}^*_{\mathrm{smp}}$ の近傍に $\boldsymbol{x}_{\mathrm{map}}$，$\boldsymbol{x}_{\mathrm{map-2nd}}$ は含まれていないことが分かる．また，$\boldsymbol{x}^*_{\mathrm{smp}}$ の形に注目すると，$\boldsymbol{x}_{\mathrm{org}}$ に非常に近い形をしていることが分かる．これらの結果から，事後確率分布の構造に奇妙な構造が潜んでいると判断できる．

隠れマルコフ場（2 次元格子型，木型）においても同様な解析をおこなった．その結果，奇妙な構造をもつ例を検出できたが，上記隠れマルコフモデルの場合と比較し，直観的には理解しにくい結果が得られた（詳細は省略）．

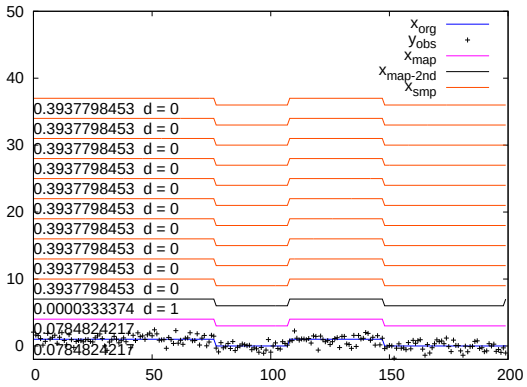

図 6

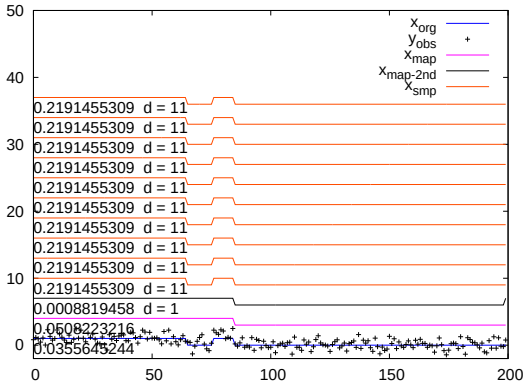

図 7

(2) 成果の国内外での位置付けと impact

高次元データに対し，本研究のように，正確な事後確率を計算している研究は，著者の知る範囲では見当たらない．また，事後確率分布から正確にランダムサンプルをおこない，事後確率分布を調べている研究も見当たらない（どちらも近似計算はよく見かける）．

(3) 今後の展望

高次元データの近傍の生成手法が今後の課題である．高次元空間では，近傍の状態があまりにも多すぎ，すべてを把握できない．この点を解決することが今後の課題である．

5. 主な発表論文等

【雑誌論文】(3 件)

[1] 伊達 章: 自己組織神経回路モデルによる情報表現の獲得, Telecom Frontier, no.81, pp.1–9, 2013. 査読無
http://www.scat.or.jp/frontier/frontier81/date.pdf

[2] 伊達 章, 倉田耕治: ボルツマンマシンを応用したトポグラフィックマッピングの形成モデルについて, 電子情報通信学会技術研究報告, vol.112, no.480, pp.203–208, 2013. 査読無
http://ci.nii.ac.jp/naid/110009713279

[3] 伊達 章: メトロノームの同期現象 +，偏微分方程式と現象：PDEs and Phenomena in Miyazaki 2010, pp.1–12, 2010. 査読無
http://www.cc.miyazaki-u.ac.jp/math/ppm/ppm2010/

【学会発表】(3 件)

[1] 伊達 章: 自己組織ダイナミクスのシミュレーション, 第 5 回 符号と力学系ワークショップ, （福岡県二日市）， 2014 年 3 月 21 日.

[2] 伊達 章: ボルツマンマシンを応用したトポグラフィックマッピングの形成モデルについて, 電子情報通信学会ニューロコンピューティング研究会，玉川大学工学部（東京都町田市）2013 年 3 月 15 日.

[3] Akira Date: A mathematical model of associative memory with a sense of feeling of knowing, Japan-Germany Joint Workshop on Computational Neuroscience, 沖縄科学技術大学院大学（沖縄県国頭郡） 2011 年 3 月 3 日.

6. 研究組織

(1) 研究代表者

伊達 章 （DATE, Akira）
宮崎大学・工学部・准教授
研究者番号 : 60322707

(2) 連係研究者

倉田 耕治 (KURATA, Koji)
琉球大学・工学部・教授
研究者番号 : 40170071