様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ５ 月 ２７ 日現在

機関番号：１２６０８
研究種目：若手研究(B)
研究期間：2012～2014
課題番号：２４７００１３９

研究課題名（和文）漸近理論に基づく潜在変数推定精度の近似計算法

研究課題名（英文）Approximation methods of accuracy of latent-variable estimation based on asymptotic error

研究代表者
山崎　啓介（Yamazaki, Keisuke）

東京工業大学・総合理工学研究科（研究院）・助教

研究者番号：６０３７６９３６

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：情報科学の分野ではデータ構造を調べることが重要な課題のひとつである。データ分析で用いられる階層モデルは観測変数と潜在変数の２つの変数を有する。本研究は潜在変数の推定に着目した。潜在変数の真値は与えられないため、推定精度を評価する方法は十分に研究されていない。我々は与えられたデータから潜在変数推定精度を計算し最適なモデル構造を発見する新たな方法を提案した。変数が過不足なく定義されている場合に提案手法が精度を近似できることを示した。変数に冗長性が生じている場合は付加的な情報が必要であることもわかった。潜在変数の一部が観測可能である場合について推定精度の導出を行い、提案手法拡張のための基礎を築いた。

研究成果の概要（英文）：In many information science fields, one of the important tasks is to investigate the structure of data. Hierarchical statistical models, which is often used for data analysis, consists of two types of variables: observable and latent variables. The present study focused on estimation of the latent variable and its accuracy. Since the true latent variables are not available or observable, a method to evaluate the accuracy has not thoroughly been studied. We proposed a new method to calculate the accuracy from the given observable data and to design the optimal structure of the model. The result shows that the method enables us approximate the accuracy when the latent variable is well-specified. When the variable has redundancy, the method requires additional information of the variable. Then, we theoretically revealed the asymptotic behavior of the accuracy for an extension of the proposed method when some parts of latent variables are observable.

研究分野：統計的機械学習

キーワード：数理統計学　最尤推定　ベイズ推定　潜在変数推定　教師なし学習　半教師あり学習

１．研究開始当初の背景

情報科学の多くの分野ではデータに潜む構造や隠れた法則を見つけることが重要な課題のひとつである。このための処理に用いられるのが混合分布やベイジアンネットワークなどの階層型統計モデルである。モデルは観測されたデータを表現する観測変数と、隠れた因子を表現する潜在変数の２種類の変数で構成される。これまで観測変数の予測に関する理論が構築され、その精度が明らかになった。一方、潜在変数は真の値が与えられないため、その推定法や精度について十分な検討がなされていない。

２．研究の目的

観測変数の予測では推定精度に関する研究が多くなされており、精度を表す誤差関数の理論値（漸近形）が導出されている。この漸近形を用いることで予測誤差の値をデータから近似する手法が与えられた。また、誤差を最小化するモデル構造を同定する手法も提案されている。観測変数はデータとして与えられるため、交差検証法などの計算により理論値が不明な場合でも近似値を求めることが可能である。
一方、潜在変数の推定精度を近似する場合、データにはこの変数が現れないため、上記の交差検証法のような近似計算は原理的に不可能である。
本研究では、潜在変数の推定精度を与えられたデータから計算する手法を確立し、最適な潜在変数の次元数やモデル構造の同定の基礎を築くことを目的とする。

３．研究の方法

交差検証法のようにデータから直接的に潜在変数の推定精度を計算するのは不可能である。そこで本研究では予測誤差における情報量基準に倣い、理論値（漸近形）に基づく近似計算のアプローチをとる。
研究代表者の過去の研究により、潜在変数の推定誤差をカルバックライブラーダイバージェンスで測った場合の漸近形が導出されている。本研究ではこの理論値と漸近的に同じ値を有し、かつデータから計算可能な関数を見つけることで近似値を求める。
理論値はモデルの潜在変数とデータを生成する情報源のそれとの関係により、以下の２通りの場合で導出されている。

(1)モデルの潜在変数が過不足なく情報源のそれを表現している場合：
事前の理論解析により、n をデータ数としカルバックライブラーダイバージェンスで真の潜在変数確率と推定確率の誤差 D(n)を定義すると、その漸近形は
$D(n) = f(w^*)/n + o(1/n)$
となることが分かっている。ここで f(w)はモデルが有するパラメータ w の関数、w*は情報源を表す真のパラメータである。関数 f(w)は推定法に依存しており、最尤推定では、
$f(w) = \mathrm{Tr}(I(w)J(w)^{-1}) - E$,
ベイズ推定では
$f(w) = \ln \det(I(w)J(w)^{-1})$
となる。ここで I(w)は観測変数と潜在変数の同時分布に関するフィッシャー情報行列、J(w)は観測変数のみの周辺確率に関するフィッシャー情報行列、E は単位行列である。誤差 D(n)の値を計算するためには真のパラメータ w*が必要であり、データからは直接求めることができない。そこで最尤推定量などのパラメータ推定を行うことでデータから f(w*)の値を近似する。

(2)モデルの潜在変数が情報源のそれと比べ冗長な場合：
(1)と同様の誤差関数に対し、ベイズ推定における漸近形が導出されており,
$D(n) = g \ln n / n + o(\ln n/ n)$
で表される。ここで係数 g は真のパラメータと潜在変数の冗長性の両方に依存する。
(1)の場合と異なり、真のパラメータの推定だけでは係数 g は求めることができない。そこでベイズ自由エネルギーF_B(n)と変分自由エネルギーF_V(n)を用いて漸近形が D(n)と同じ形式になるような関数を作る。ちなみに２つの自由エネルギーはデータから計算できる関数である。観測変数が離散の場合には
$g \ln n = F_V(n) - F_B(n) + o(\ln n)$
が示せる。つまり自由エネルギーの差を n で除した量は漸近的に D(n)と等しい。この関係を用いて誤差を近似する。

４．研究成果

精度の近似を評価するため、予め人工的に情報源を定めデータを発生させた。本研究の成果は３．に対応する２つの場合の精度近似法の評価と手法拡張のための新たな理論解析とをあわせ合計３つの項目からなる。

(1)精度近似手法の評価（モデルの潜在変数が過不足なく情報源のそれを表現している場合）：
離散データを表現する混合二項分布を情報源とした場合、観測変数における最尤推定量 w_{ML}を用いて近似を行った。つまり
$D_{ML}(n) = f(m_{ML})/n$
を計算し D(n)を近似した。漸近的に
$D(n) = D_{ML}(n) + o(1/n)$
となることが示せるため、理論的な保証のある近似法である。最尤推定、ベイズ推定ともに良好な結果を得た。
連続データを表現する混合正規分布を情報源とした場合、分散・共分散行列をパラメータに含めると、観測変数における最尤推定量

w_{ML}は発散することが知られている。EM アルゴリズムを用いて w_{ML}を探索すると w* と大きく異なる推定量となることを確認した。つまりこのモデルでは w_{ML}を用いた近似計算は信頼性が低い。そこで変分ベイズ法を用いたパラメータ推定量 w_{VB}を代入し、D_{VB}(n) = f(w_{VB})/n
を近似値とした。w_{VB}の漸近挙動は解明されていないため D(n)と D_{VB}(n)の差を理論的に調べることは困難であるが、実験的に良い近似を与えることが確認できた。

(2) 精度近似手法の評価（モデルの潜在変数が情報源のそれと比べ冗長な場合）：
誤差D(n)の漸近形の係数gは離散データの場合に導出されているため、混合二項分布を情報源として評価を行った。
２つの自由エネルギーの差により D(n)を近似した結果、(1)に比べ近似精度が低下することがわかった。係数 g はベイズ推定における事前分布には依存しない。しかしながら事前分布のハイパーパラメータにより実験結果が大きく変化した。漸近形の高次項の影響が強いこと、自由エネルギーの計算精度に問題があることの２つが原因と考えられる。近似計算の精度を上げるのが今後の課題である。

(3)提案手法拡張のための理論解析：
上記２つの近似計算手法の実験的評価によって、潜在変数の冗長性がなければ高精度な近似が可能であることがわかった。潜在変数の次元や範囲などの情報が手に入れば、それを基に冗長性を排除できる。そこで潜在変数の一部が観測可能な場合について考察を行った。これは半教師あり学習と呼ばれる。
本研究では、半教師あり学習のクラスター分析において、最尤推定、ベイズ推定の誤差の漸近形を導出した。教師なし学習（潜在変数の情報がない場合）と同様に最尤推定と比較し、ベイズ推定が漸近的に高精度となることを証明した。
さらに潜在変数の次元が観測可能なものとそうでないもので変化する場合についても解析を行った。潜在変数に変化が起こるものの、観測変数とあわせた推定結果が向上するため、与えられる情報を無視するのに比べ、情報を用いたモデルのほうが高精度になることを証明した。

本研究で提案した近似手法は(2)の場合に自由エネルギーの計算を必要とする。しかしながらこの計算量は変分自由エネルギーに比べ非常に多く、エネルギー値の精度を上げるのが困難である。そこで変分ベイズ法による潜在変数推定の精度を実験的に求めた。ベイズ推定に近ければ、少ない計算量で高精度な推定の実現が期待できる。
潜在変数に冗長性がある場合とない場合の両方で数値実験を行った。冗長性がない場合は最尤推定に近く、ベイズ推定に比べ精度が劣ることがわかった。また冗長性が存在する場合はベイズ推定よりは低いが、事前分布の影響などは近い挙動を示すことがわかった。今回は実験的な比較であるが、漸近論などを用い理論的な比較を行うことが今後の課題である。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計５件）

① Keisuke Yamazaki, “On Bayesian Clustering with Structured Gaussian Mixture”, Journal of Advanced Computational Intelligence and Intelligent Informatics, 査読有, 8(6), pp.1007-1012, 2014

②山崎啓介, “多種粒子 TASEP を表現する混合分布モデルと統計的粒子クラスタリングについて”, 日本応用数理学会論文誌, 査読有, 24(4), pp.357-372, 2014

③ Keisuke Yamazaki, "Asymptotic Accuracy of Distribution-Based Estimation for Latent Variables", Journal of Machine Learning Research, 査読有, 13, pp.3541-3562, 2014

④ Takuto Naito, Keisuke Yamazaki, “Asymptotic Marginal Likelihood on Linear Dynamical Systems”, IEICE-ED, 査読有, E97-D, 4, pp.884-892, 2014

⑤ Keisuke Yamazaki, Daisuke Kaji, "Comparing Two Bayes Methods Based on the Free Energy Functions in Bernoulli Mixtures", Neural Networks, 査読有, 44, pp.36-43, 2013

〔学会発表〕（計１６件）

①中村文士，山崎啓介，“交通流映像からの速度決定則のモデル化とグループ分け”，電子情報通信学会 ITS 研究会，北海道大学（北海道札幌市），ITS2014-53，155-160，2015 年 2 月 23 日

② Keisuke Yamazaki, "The Optimal Hyperparameter for Bayesian Clustering and its Application to the Evaluation of Clustering Results", Proc. of SCIS-ISIS2014, 北九州国際会議場（福岡県北九州市）, pp. 961-965, 2014 年 12 月 4 日

③Fumito Nakamura, Keisuke Yamazaki, "Two Statistical Methods for Grouping Vehicles in Traffic Flow Based on Probabilistic Cellular Automata", Proc. of

SCIS-ISIS2014，北九州国際会議場（福岡県北九州市），pp. 956-960，2014 年 12 月 4 日

④山崎啓介，“ 変分ベイズ法における潜在変数推定の精度について ”，IBIS2014 ワークショップ，名古屋大学（愛知県名古屋市），2014 年 11 月 18 日

⑤中村文士，山崎啓介，“ 交通流モデル ZRP における変分ベイズ法について ”，IBIS2014 ワークショップ，名古屋大学（愛知県名古屋市），2014 年 11 月 18 日

⑥山崎啓介，"クラスバランスが変化する状況下でのベイズクラスタリング精度について"，IBIS2013 ワークショップ，東京工業大学（東京都目黒区），2013 年 11 月 12 日

⑦大原成裕，山崎啓介，渡辺澄夫，"学習結果の対応づけを用いた自己組織化写像のねじれ評価"，IBIS2013 ワークショップ，東京工業大学（東京都目黒区），2013 年 11 月 12 日

⑧山崎啓介，“ 多種粒子 TASEP における統計的粒子識別について ”，第 19 回交通流のシミュレーションシンポジウム，名古屋大学（愛知県名古屋市），2013 年 12 月 16 日

⑨Keisuke Yamazaki，"Accuracy of Latent Variable Estimation with the Maximum Likelihood Estimator for Partially Observed Hidden Data"，ISITA，Honolulu (USA)，2012 年 10 月 30 日

⑩ Keisuke Yamazaki，"MCMC Sampling on Latent-Variable Space of Mixture of Probabilistic PCA"，SCIS-ISIS，神戸国際会議場（兵庫県神戸市），2012 年 11 月 23 日

⑪山崎啓介，"完全データと不完全データの混合におけるベイズ潜在変数推定の精度"，電子情報通信学会情報論的学習理論と機械学習研究会，京都キャンパスプラザ（京都府京都市），IBISML2012-11，pp.73-77，2012 年 6 月 20 日

⑫山崎啓介，"制約がある分布でのベイズクラスタリングについて"，日本神経回路学会全国大会(JNNS2012)，名古屋工業大学（愛知県名古屋市），2012 年 9 月 21 日

⑬山崎啓介，渡辺一帆，梶大介，"自由エネルギーによる潜在変数推定精度の計算法"，電子情報通信学会情報論的学習理論と機械学習研究会，筑波大学（東京都文京区）IBISML2012-44，pp. 75-81，2012 年 11 月 7 日

⑭小林浩一、山崎啓介，"交通流の時空図における ZRP のパラメータ推定と能動学習"，電子情報通信学会情報論的学習理論と機械学習研究会，筑波大学（東京都文京区），IBISML2012-47，pp. 97-101，2012 年 11 月 7 日

⑮梶大介、山崎啓介，"複数クラスラベル付きバイナリーデータ群からの相関抽出方法と投薬データ解析への応用"，第 15 回情報論的学習理論ワークショップ(IBIS2012)，筑波大学（東京都文京区），ディスカッショントラック，2012 年 11 月 8 日

⑯金井政宏、山崎啓介，"2 種粒子 TASEP における有効データ長分布からのパラメータ推定について"，第 18 回交通流のシミュレーションシンポジウム，名古屋大学（愛知県名古屋市），2012 年 12 月 13 日

６．研究組織

(1)研究代表者

山崎　啓介（YAMAZAKI，Keisuke）
東京工業大学　大学院　総合理工学研究科・助教
研究者番号：60376936