様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月　 ３ 日現在

機関番号：１４２０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１１７０６

研究課題名（和文）標本数問題に関する情報幾何学的アプローチ

研究課題名（英文）Sample size problem in view of information geometry

研究代表者
椎名　洋（Sheena, Yo）

滋賀大学・データサイエンス学系・教授

研究者番号：８０２４２７０９

交付決定額（研究期間全体）：（直接経費）　　900,000円

研究成果の概要（和文）：パラメトリックモデルに真の分布が含まれない場合を想定した。モデルの中で一番真の分布に近い分布（Information Projection）と、最尤推定量をパラメーターに代入して得られる予測分布（Estimative Density)の近さを、カルバックライブラーダイバージェンスを用いて測り、その期待値をリスクにしたとき、そのリスクの漸近的な挙動がどうなるかについて研究した。１）リスクを標本数の二次オーダーまで漸近展開し、２）ダイバージェンスとベイズ誤差率との関係を求めた。その結果を利用して、３）与えられたモデルに必要な標本数について、一定の基準を設けることに成功した。

研究成果の学術的意義や社会的意義
単純な統計モデルから巨大な深層学習モデルまで、様々な確率モデルが構築されている。その際、モデルのパラメーターを学習させるためには、どれくらいの大きさの標本が必要か（いわゆる、標本数問題）については、はっきりした基準がなかった。本研究では、パラメーターを座標としたモデルの集まりの中で最良の点（Information Projecton）に、学習済みモデル（標本から得られる最尤推定量でパラメーターを置き換えた予測分布、Estimative density）が十分近くなるために、どの程度の大きさの標本が必要かという観点から、標本数問題に一定の答えを出している。この点に、本研究の最大の意義がある。

研究成果の概要（英文）：We assume that the parametric model does not include the true distribution. The proximity of the distribution closest to the true distribution in the model (Information Projection) to the predictive distribution obtained by substituting the maximum likelihood estimator for the parameters (Estimative Density) was measured using the Kullback-Leibler divergence, and its expected value, i.e. risk was used to study the asymptotic behavior. We studied the asymptotic behavior of the risk; 1) the risk was asymptotically expanded to quadratic order of the sample size, and 2) the relationship between divergence and Bayes error rate was obtained. Using the results, 3) we succeeded in establishing certain criteria for the sample size required for a given model.

研究分野：数理統計学

キーワード：標本数問題　リスクの漸近展開　情報幾何　予測分布　ダイバージェンス

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

パラメトリックな確率モデルに関して、推定された分布が真の分布にどの程度近いかは、統計学の分野において、非常に重要なテーマであり、多くの研究がなされてきた。しかしながら、多くの場合、真の分布がモデルの中に存在することを条件にしており、近さを測る方法としても、パラメーター同士の距離を二乗誤差で測るという方法を採用した場合の研究が非常に多かった。

真の分布がモデルの中に存在するという状況は、実際には極めて珍しく、またパラメーター間の二乗誤差は、幾何学的な「不変性」をみたしておらず、確率変数を変換すると違った結果になってしまうという欠点がある。

また、密度関数に関してもある特定の形を想定したものが多く、多層パーセプトロンのような密度関数が陽に得られない複雑な確率モデルには適用できない結果が多かった。

２．研究の目的

真の分布がモデル外にあることを前提として、分布の近さをダイバージェンスで計測することで、上記のような欠点を克服することができる。

二つの確率分布をダイバージェンスで測り、さらに真の確率分布に基づく期待値をとったリスクで、モデルが十分に真の分布に近いかを測定する。これによって、与えられたモデルに対して、現在の標本の大きさで十分かどうかという、いわゆる「標本数問題」に対して一定の答えを提供することが研究目的となる。逆の見方をすると、現在の標本の大きさのもとで学習したモデルの大きさは適切か、つまりモデルが複雑すぎて（パラメーターが多すぎて）過学習を起こしていないかを判断する基準を提供することにもなる。

３．研究の方法

理論的な研究なので、先行文献・研究を調べて研究の目的が誤っていないかを確認した後は、各種の数学的手法を使いながら、理論的な結果を導出するための計算を行うこと、そして、数学的な証明に、トライ&エラーで挑戦した。理論的な成果が得られた後は、理論を実際のデータに適用するためのプログラミングを行い、コンピューター上でいくつかの現実のデータに関して、推定量の計算、リスクの計算を行った。これによって、理論的な結果をどう使えばよいかを例示することができ、通常の計算資源のもとでどの程度計算に時間がかかるかも、ある程度判明した。

４．研究成果

（１）まず、真の分布がモデルの外側にあるので、どんなに工夫（学習）しても最終的にたどり着けるのは、モデルの中で真の分布に一番近い場所（Information Projection）であることを確認した。一方、モデルのパラメーターの推測方法としては、最も標準的な方法である最尤法を用いた場合を考えた。すなわち、最尤推定量をパラメーターに代入して得られる予測分布（Estimative Density）が、Information Projection にどれくらい近いかを考えた。また、二つの分布の近さは、不変性を満たすように Divergence を使って計測したが、今回は最も使用頻度の高い Kullback-Leibler Divergence で測った。さらに、標本分布（真の分布からの同一独立標本）に基づく期待値をとったものをリスクと呼ぶが、最終的にこのリスクを使って、Information Projection と予測分布の近さを計測した。

一般的な形でリスクを陽に表現することは、モデルがかなり単純な場合にしか可能でないので、どのような複雑なモデルでも導出可能な形でリスクを求めるために、標本数に関する漸近展開でリスクを表現することした。

（２）漸近展開の標本数の一次のオーダーの部分は、どんなモデルにも共通で、常に p/2n で与えられることが判明した。また、二次のオーダーに関する部分は非常に複雑で、モデルの Information Projection 周辺での幾何学的な性質に依存する。これに関しても、理論上は明示的な答えを得たが、実際に各種の幾何学的な量を計算して現実問題に応用するのは非常に難しいくらいの複雑さであった。

（３）次にモデルを指数型分布族に限定して、リスクの漸近展開をおこなった。指数型分布族は既存のよく知られた、非常に幅広い種類の確率分布をそのうちに含んでいる。そこに織り込む変数（特徴量）を豊富に用意すれば十分に Information Projection を真の分布に近づけることが可能であり、理論的にも様々に優れた点(例えば、Maximum Entropy 則など)があることが、設定の理由である。

（４）結果として、二次のオーダー項も一般の場合に比べてかなり簡便に表現できることになり、これで現実の問題に対する適用がかなり楽になったが、二次のオーダー項は「自然パラメーター」で表現されているので、これを「平均値パラメーター」に変換する必要がある。パラメーターは未知であり、標本から推測する必要があるが、平均値パラメーターは標本から簡単に推測できるので、推測された平均値パラメーターから自然パラメーターの推測値を求める必要があるからである。自然パラメーターを、平均値パラメーターによって陽に表現することは難しく、微分方程式の解として求める必要がある。この解法を与えるアルゴリズムを考案したことも論文の重

要な成果である。具体的には、方程式の解をあたえる基本的な方法である Newton-Raphson 法を適用する際に使う勾配行列を、乱数から計算できる標本分散共分散行列で置き換えるところが重要な工夫である。

（５）以上の結果を用いると、二次のオーダーまでで近似されたリスクを、標本から推測することができるが、最後の問題はそのように推測されたリスクの値がどれくらい小さい場合に、Information Projection と予測分布が近いと判断したらよいかの基準作りである。ダイバージェンスの値自体や、その期待値であるリスクの大きさ自体の解釈は難しい。そこで、より直観的な意味合いをもつベイズ誤差率を用いて、ダイバージェンスの大きさに意味を持たせることにした。すなわち、ダイバージェンスがある値以下であったときに、ベイズ誤差率がどれくらい 0.5 に近いかを示す命題を証明した。この命題を逆につかうことで、どれくらいリスクが小さければよいかを示す基準点を考案することができた。例えば、ベイズ誤差率が 0.45 であるためには、ダイバージェンスが 0.02 であることが十分条件であるので、先の方法で推測されたリスクが 0.02 以下であれば、十分に Information Projection と予測分布は近いと判断できる。この場合、標本の大きさは、与えられたモデルの複雑さのもとでは、十分であるという結論になる。もしリスクが 0.02 以上であれば、標本数をもっと増やす、あるいはモデルをもっとシンプルにする（パラメーターの数を減らす）ことによって対処することになる。

（６）具体的な例として、ワインの格付けと化学成分の同時分布を指数型分布モデルで構築し、その際にリスクが十分に小さくなっているかを検証した。また、有限離散分布の場合の例として、アワビの直径の分布をヒストグラムにする問題を考え、同様にリスクの推定を行い必要な標本の大きさについて考察した。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Yo Sheena	４．巻 64
２．論文標題 Convergence of estimative density: criterion for model complexity and sample size	５．発行年 2023年
３．雑誌名 Statistical Papers	６．最初と最後の頁 117-137
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s00362-022-01309-9	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関