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 ２８ 年　 ６ 月　 ９ 日現在

機関番号：１３９０１
研究種目：基盤研究(C)（一般）
研究期間：2012～2015
課題番号：２４５００３４０

研究課題名（和文）密度比の推定と計算の理論的展開とその応用

研究課題名（英文）Theory and Applications of Density Ratio

研究代表者
金森　敬文（Kanamori, Takafumi）

名古屋大学・情報科学研究科・教授

研究者番号：６０３３４５４６

交付決定額（研究期間全体）：（直接経費）　　4,100,000円

研究成果の概要（和文）：密度比とは，確率密度関数の比として定義される関数である．２つのデータドメインの間のマッチングを行う場合などに有用である．重要な応用例としては，共変量シフトの下での回帰分析や判別分析，統計的仮説検定，次元削減などが挙げられる．本研究課題では，高次元大規模データ解析への応用を念頭に置き，密度比の統計的な推定精度を向上させるための方法について研究を進めた．さらに，確率分布間の距離を表すダイバージェンスとの関連について研究を進めた．多くの推定アルゴリズムはダイバージェンスを用いて記述することができる．その統計的な有用性について，とくに回帰分析とロバスト統計の視点から，密度比との関連を考察した．

研究成果の概要（英文）：Density ratio is defined as the ratio of two probability densities. That is very useful when two distinct data domains are compared. Important applications of the density ratio includes regression analysis and classification problems under covariate shift, statistical test, and dimension reduction. In this study, we developed methods to improve the statistical accuracy and computational efficiency of density ratio estimation. Moreover, we studied the relation between density ration and statistical divergences which is regarded as a discrepancy measure between two probability distributions. Statistical usefulness of the divergence is considered from the view point of the density ratio.

研究分野：機械学習

キーワード：数理統計学　機械学習

１．　　研究開始当初の背景

２つの確率密度関数の比を密度比とよぶ．数理統計学や機械学習におけるさまざまな問題が，データから密度比を推定することに帰着される．

密度比推定として定式化される問題として，例えば，学習データとテストデータの分布が異なる状況のもとでの統計的学習(共変量シフトの下での学習)，外れ値検出，特徴量選択などが挙げられる．それぞれの定式化は以下のようになる．

共変量シフトでは，おもに回帰分析の問題を考える．独立変数の分布が，学習データとテストデータで異なるとき，通常の最小２乗法だと推定量に大きなバイアスが生じ，適切に予測を行うことができない．そこで分布の違いを，学習データの分布とテストデータの分布の比によって表し，学習データを適切に重み付けすることで，バイアスを補正することができる．

外れ値検出では，外れ値が混入していないことが保証されているデータと，外れ値が混入している可能性があるデータについて，それぞれの分布から定義される密度比を用いて，外れ値の特徴を抽出する．これにより，将来得られるデータが外れ値かどうかをオンライン処理により判定することを目指す．

特徴量選択では，予測のために重要な変数を抽出することを目指す．このために，情報理論で用いられる相互情報量や，これに類似の情報量尺度を用いて，独立変数のセットと従属変数の相関を計測する．これらの情報量尺度の計算において，データが相関している場合と独立な場合のそれぞれの分布から定義される密度比が現れる．データから密度比を通して情報量尺度を推定し，特徴量選択を行う手法が考案され，研究が進められている．

一方で密度比推定は，伝統的な統計学において，カーネル密度推定の密度比への拡張や，ケース・コントロールスタディなどへの応用などの話題が，散発的に研究されてきた．これらの知見は重要であるが，データ解析の基盤技術としての密度比推定という視点には至っていない．統計学や情報理論の分野では，それぞれ固有の問題と関連して，個別に密度比推定の研究が進められてきたという背景がある．

最近になり，とくに機械学習の分野において，密度比の重要性が認識され，密度比を用いた統計的推論の応用分野が爆発的に拡大している．上記に示した研究の背景や進捗状況を踏まえ，本研究では密度比推定の理論的基礎をさらに充実させ，大規模データへの適用を念頭に置いて，効率的な計算アルゴリズムを開発することを目標とする．さらに密度比推定の新たな応用を創成することを目指す．

本研究の研究代表者はこれまで，共同研究者らと密度比推定のための計算アルゴリズムなどを提案してきた．密度比は確率密度の比であるので，分子と分母の確率密度をそれぞれデータから推定し，それらの比をとるという素朴な推定法が考えられる．しかし，これでは全く信頼性のない推定量になってしまうことが，実用上指摘されている．推定量の統計的な信頼性を改善するために，本研究の研究代表者らは，最小２乗法を用いて，データから密度比を直接推定する方法を提案した．通常の回帰分析とは異なり，密度比推定では出力データに対応するデータが存在しないにも関わらず，最小２乗法が有効に働く点が極めてユニークである．この研究は大いに注目を集め，機械学習におけるトップ国際会議にアクセプトされ，さらにトップジャーナルに掲載された．また，発表当初からよく引用されていることが確認できる．提案手法はおもにパラメトリックモデルに基づく手法であり，今後，より複雑で大規模なデータ解析への応用を念頭に置いて研究を進展させる必要があると考え，本研究課題を提案した．

２．研究の目的

研究の背景とこれまでの研究成果を踏まえて，本研究課題の目的を以下のように定め，研究を推進する．

(1) いままで提案された手法では対処できないような，複雑なデータや問題設定に対応するために，パラメトリック推定法だけでなく，密度比のセミパラメリック推定法やノンパラメリック推定法を提案する．セミパラメリック推定は，統計モデルが無限次元の未知パラメータを含むが，興味のあるパラメータは有限個という設定である．密度比推定においては，密度比の値が大きく異なる部分空間を推定する，などの問題がセミパラメリック推定として定式化される．またノンパラメリック推定は，無限次元の未知パラメータを推定する推定手法である．もし高精度のノンパラメトリック推定が可能になれば，柔軟で高精度な推論が可能になる．

(2) 高次元小標本データから精度よく密度比を推定するためのスパース推定法を開発する．高次元データの密度比を推定するために，高次元の基底関数をもつ統計モデルを仮定する．一方で，小標本データに対処するために，数多くの基底関数のなかから，重要な基底関数を適応的に選択する必要がある．このような推定パラダイムを総称してスパース学習などとよぶ．このような考え方は，情報理論では compressed sensing という名称で近年爆発的に重要性が増している．スパース学習の考え方を密度比推定に導入し，現代の大規模高次元データを高い信頼性で処理するための統計手法を開発することが必須の課題である．

(3) 統計的ノイズに強い計算アルゴリズムを開発する．現在，数理最適化の分野の進展が著しい．最適化の分野で開発されたさまざまな計算・最適化手法を機械学習のアルゴリズムに応用することで，推定における計算性能

が向上することが期待される．しかしその際，単に最適化の手法を統計に応用すればよいというものではない．データにはランダムネスが含まれることを考慮した上で，統計的ノイズ対してロバストであるような計算アルゴリズムを設計する必要がある．

３．研究の方法

本研究の初年度では，主に密度比推定の統計的性質の解明に重点をおく．とくに研究目標の (1)，(2) について，いままでに得られている成果を進展させる形で研究を進める．

(1)の研究計画：

さまざまな推定法の基礎となるパラメトリックモデルのもとでの密度比推定について，理論的基盤を築く．そこで得られた知見に基づいて，より複雑で現実的な問題設定，すなわち，密度比のセミパラメトリック推定やノンパラメトリック推定などの研究テーマに打ち込む．

密度比推定の統計的性質をより深く正確に理解するため，まず最初に密度比のパラメトリック推定に対して，クラメール・ラオ不等式に対応するような基本的な推定限界を導出することを目指す．さらに，漸近理論や高次漸近理論の建設について検討する．密度比のパラメトリック推定における 2 次漸近有効な推定量の構成まで到達できれば，一定の成果を得たと言えるであろう．

さらに，密度比のセミパラメトリック推定へ研究を進める．この問題は確率分布の推定におけるセミパラメトリック推定の一種として解釈できる可能性がある．しかし密度比推定のセミパラメトリック推定を新しい統計的問題として定式化することは，応用の可能性を広げるものと考える．密度比の次元削減問題やダイバージェンス推定問題を密度比のセミパラメトリック推定の視点から捉え，信頼性の高い推定量を提案する．さらに密度比のセミパラメトリック推定の理論基盤を築く．例えば，通常のセミパラメトリック推定における有効スコアに対応する推定量を密度比に対して考察する．これらの研究を通して，密度比のセミパラメトリック推定に対する理解を深める．さらに応用からのフィードバックを取り入れることで，新たな理論構築への方向性を探る．

有限次元のパラメトリックモデルを仮定せず，観測データに依存して適応的に基底関数やモデルを定めて推定する方法を，ノンパラメトリック推定とよぶ．主に機械学習の分野でさかんに研究が進められている，再生核ヒルベルト空間に付随するカーネル関数を用いた推定法は，効率的な計算アルゴリズムが存在するなど，実用上非常に有用であることが知られている．密度比推定に対しても，再生核ヒルベルト空間を用いる推定量を自然に導出することができると考えられる．さらにその推定精度や正則化項の適切な決め方について，理論的な考察を行う．ノンパラメトリック推定法は，どのようなデータに対しても大抵は適用可能で，一応の解析結果を与える．しかし，得られた結果に対する統計的な信頼性を評価する上で，理論的な考察は欠かすことはできない．近年発展しているカバリング・ナンバーやカーネル関数のスペクトルを用いる理論的な方法によって，再生核ヒルベルト空間をモデルとするノンパラメトリック推定法に対して精度保証を与えることが可能である．密度比の推定に対しても同様のアプローチが有効であると考え，研究を進める．

(2)の研究計画：

高次元小標本データは近年爆発的に研究が進展している．例えば線形回帰分析では，L1 正則化項を用いて変数選択を行う LASSO などの推定法などが代表的である．密度比推定においても高次元小標本データにおける変数選択は重要である．これは(1)におけるセミパラメトリック推定として定式化することも可能と考えられるが，通常の統計的漸近理論の枠組の外にある統計的問題と捉えることが肝心である．なぜなら，高次元データから小数の特徴量を選択する問題では，スパース性のパターンに対する統計的一致性の問題など，特有の問題が設定されるためである．密度比推定に対して，近年のスパース学習などの知見を取り入れながら，高次元小標本データに対する密度比推定の変数選択問題や，基底関数選択に対するスパース・パターンの統計的一致性の問題に取り組む．

また判別分析の問題では，入力 x に対する出力 y の条件付き確率の推定可能性とスパース性との関連が議論されている．推定結果がスパースになりすぎる推定量では，推定不可能となる確率値が存在する．例えばサポートベクトルマシンでは，推定結果が非常にスパースになるため確率値が 0.5 に等しくなるかどうかのみ，推定が可能になる．出力の予測を行うだけなら条件付き確率が 0.5 以上か以下かが分かれば十分なので，サポートベクトルマシンは有力な判別アルゴリズムとしてその地位を築いている．このようなスパース性と推定可能性とのトレードオフを，密度比推定に対して理論的に定量化する．これにより「知りたいこと」と「用いるべき推定法」との間の関係が直接的に与えられることになる．これは，応用に大きなインパクトを与える研究成果となることが予想される．

本研究の２年目以降は，研究目標の (1)，(2) から(3)計算アルゴリズムの開発へと重点を移行しつつ研究を進める．大規模データに対応可能な，効率的な計算アルゴリズムの開発を進める．さらに統計解析言語Ｒを用いたソフトウェア開発を進める．これを通して密度比推定の応用の可能性を広げ，密度比推定に関して理論サイドに求められる，新しい研究テーマの発掘を目指す．(3)の研究計画：推定量を提案することができても，効率的に計算できなければ，%結果が得られないようで

は，広く応用されるような知的技術にはなりにくい．最新の数値計算技術を取り入れることを前提として，密度比の推定アルゴリズムを設計する．最適化の分野でも近年，確率最適化やロバスト最適化などの定式化により，データにノイズが含まれていることを前提とした問題を取り扱うことが増えてきている．例えば，確率最適化やロバスト最適化などは，データのランダムネスを考慮した最適化手法として理論的にも応用面でも発展している．これらの研究で得られている知見を積極的に取り入れ，密度比の効率的な計算アルゴリズムを提案する．さらに統計解析言語Rを用いてアルゴリズムの実装を行い，ライブラリとして公開する．効率的な計算を実現するために，現代のマルチコアCPUに対応した並列処理を積極的に利用する計算アルゴリズムを提案することを目指す．

４．研究成果

初年度，当初の研究計画では，密度比のパラメトリック推定だけでなく，セミパラメトリック推定についても研究を進める予定であった．また，新たに提案されたパラダイムである密度差についても，積極的に研究を進めることを計画していた．実際には，密度比と密度差の推定をスコアに基づいて行う方法に関する研究が急速に進展し，そちらを重点的に進めることとなった．具体的には，密度比と密度差の推定量の性質の違いについて，ロバスト統計の観点から研究を進めた．さらに密度比推定の安定化のための推定法について考察を進めた．この結果，密度比と密度差の両方に関して，安定して推定を行うための方法を提案することができた．さらに密度差に対するバイアス補正推定量を提案した．以上の結果は，実際のデータ解析を行う上で非常に重要である．スコアに基づく方法は，さまざまなデータ解析に対して応用可能な汎用的手法であり，統計的決定論の立場から密度比や密度差の推定を考える上で基本的なツールとなり得る．したがって，スコアを用いた統計的推論という視点からの研究は，既存の統計的手法を拡張するための理論的基盤として，今後の進展が大いに期待される．さらに当該年度は，密度比の半教師付き学習への応用においても研究成果を得ることができた．これにより，さらに密度比の応用が広がったことになる．以上を鑑みると，当初の研究計画に完全に沿ってはいないが，内容としては，理論サイドから極めてインパクトのある研究成果を得ている．今後，スコアを用いたパラメトリック密度比推定や密度差推定の枠組を理論的に拡張し，推定の安定性や精度の向上を目指すことになる．さらに，セミパラメトリック推定への展開についても考察を深め，実践的統計手法の開発に本格的に着手することを計画している．当該年度は，そのための基盤を腱固に構築することができたと，高く評価するものである．
２年目以降は特に，高次元大規模データ解析への応用を念頭に置きつつ，密度比推定において自然に導入される制約式に着目し，統計的な推定精度を向上させるための方法について研究を進めた．その成果は，T. D. Nguyen, et al., "Constrained Least-Squares Density-Difference Estimation" として IEICE Transactions on Information and Systems から査読有り論文として出版されている．さらに，数理統計学における重要な概念であるダイバージェンスの研究を進めた．ダイバージェンスは関数空間上の距離を拡張した概念であり，多くの推定アルゴリズムはダイバージェンスを用いて記述することができる．Bernoulli に掲載された査読有り論文，T. Kanamori and H. Fujisawa,"Affine Invariant Divergences associated with Proper Composite Scoring Rules and their Applications"では，とくにデータ変換に対する不変性の概念から出発し，いままで提案された推定量をまったく新しい方向に拡張するようなダイバージェンスのクラスを導出した．その統計的な有用性について，とくに回帰分析とロバスト統計の視点から考察した．提案したダイバージェンスは密度比推定においても重要な役割を果たすことが予想される．
更に本研究課題の最終年度において，いくつかの進展がみられた．これまでの成果を踏まえ，本年度は特に離散確率分布に着目し，データから分布を推定するための方法を提案した．この方法では，経験分布による局所化という新しい分布の変換法を提案し，計算効率を大幅に向上させることに成功した．さらに斉次ダイバージェンスとよばれる相対不変性を満たすダイバージェンスを用いることで，正規化定数の計算が不要になり，大規模モデルに適用することが可能になる．提案した統計手法が優れた精度と計算効率を達成することを，理論的，数値的に確認した．経験分布による局所化では，通常の統計モデルと経験分布から定義される密度比に関連する分布を新たな統計モデルとみなし，統計的推論に応用する．この研究成果は，論文 Takenouchi T, Kanamori T., "Empirical Localization of Homogeneous Divergences on Discrete Sample Spaces." としてまとめられ，国際会議 NIPS 2015 の場でポスター＆スポットライトとして発表された．関連して，ロバスト推定を行うための統計的手法について考察し，研究成果を数本の論文にまとめた．

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 18 件）
① T. Kanamori, H. Fujisawa,

"Robust Estimation under Heavy Contamination using Unnormalized Models". Biometrika, vol. 102, no. 3, pp. 559-572, Sep. 2015.

② A. Takeda, S. Fujiwara, T. Kanamori, "Extended Robust Support Vector Machine Based on Financial Risk Minimization". Neural Computation, vol. 26, num. 11, pp. 2541-2569, Nov. 2014.

③T. Kanamori and H. Fujisawa, "Affine Invariant Divergences associated with Proper Composite Scoring Rules and their Applications". Bernoulli, vol. 20, No. 4, pp. 2278-2304, Nov. 2014

④T. Kanamori and A. Takeda, "A Numerical Study of Learning Algorithms on Stiefel Manifold". Computational Management Science, vol. 11, Issue 4, pp 319-340, Oct. 2014.

⑤A. Takeda, T. Kanamori, "Using Financial Risk for Analyzing Generalization Performance of Machine Learning Models". Neural Networks, vol. 57, pp. 29-38, Sep, 2014.

⑥T. D. Nguyen, M. C. du Plessis, T. Kanamori, M. Sugiyama, "Constrained Least-Squares Density-Difference Estimation". IEICE Transactions on Information and Systems, vol. E97-D, no. 7, pp. 1822-1829, July, 2014.

⑦T. Kanamori, "Scale-Invariant Divergences for Density Functions". Entropy, vol 16(5), pp. 2611-2628, May 2014.

⑧T. Kanamori and M. Sugiyama, "Statistical Analysis of Distance Estimators with Density Differences and Density Ratios". Entropy, vol. 16 (2), pp. 921-942, Feb. 2014.

⑨T. Kanamori, A. Ohara, "A Bregman extension of quasi-Newton updates II: analysis of robustness properties". Journal of Computational and Applied Mathematics, vol. 253, pp. 104-122, Dec. 2013.

⑩M. Sugiyama, T. Kanamori, T. Suzuki, M. C. du Plessis, S. Liu, I. Takeuchi, "Density Difference Estimation". Neural Computation, vol. 25(10), pp. 2734-2775, Oct. 2013.

⑪T. Kanamori, A. Takeda, T. Suzuki, "Conjugate Relation between Loss Functions and Uncertainty Sets in Classification Problems". Journal of Machine Learning Research, vol. 14, pp. 1461-1504, June, 2013.

⑫M. Sugiyama, S. Liu, M. C. du Plessis, Y. Yamanaka, M. Yamada, T. Suzuki, T. Kanamori, "Direct Divergence Approximation between Probability Distributions and Its Applications in Machine Learning". Journal of Computing Science and Engineering, vol. 7, no. 2, pp.99-111, June, 2013.

⑬M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, M. Sugiyama, "Relative Density-Ratio Estimation for Robust Distribution Comparison". Neural Computation, vol. 25, No. 5, pp. 1324-1370, May 2013.

〔学会発表〕(計 28 件)

①Takenouchi T, Kanamori T. "Empirical Localization of Homogeneous Divergences on Discrete Sample Spaces". The Neural Information Processing Systems (NIPS 2015), poster & spotlight, 2015.

②Kanamori T. "Legendre Transformation in Machine Learning". Workshop: Information Geometry for Machine Learning, December 2014.

③Fujisawa, H., Kanamori T. "Affine invariant divergences with applications to robust statistics". The 7th International Conference of the ERCIM WG on Computational and Methodological Statistics (ERCIM 2014), the University of Pisa, Italy, 6-8 December 2014.

④Kanamori T., Fujisawa, H. "Affine Invariant Divergences and their Applications". The 3rd Institute of Mathematical Statistics, Asia Pacific Rim Meeting, June 29-July 3, 2014.

⑤Sugiyama M., Kanamori T., Suzuki T., Plessis M., Liu S., Takeuchi I.
"Density-Difference Estimation".
The Neural Information Processing Systems (NIPS 2012), Lake Tahoe, Nevada, United States, 3-8 Dec., 2012.

⑥Kanamori T., Takeda A.
"Non-Convex Optimization on Stiefel Manifold and Applications to Machine Learning".
The 19th International Conference on Neural Information Processing (ICONIP 2012), Doha, Qatar, 12-15 Nov., 2012.

⑦Takeda A. Kanamori T., Mitsugi H.
"Robust optimization-based classification method".
The 21st International Symposium on Mathematical Programming (ISMP 2012), Berlin, Germany, 19-24 Aug., 2012.

⑧Kanamori T., Suzuki, T., Sugiyama, M.
"f-divergence estimation and two-sample test under semi-parametric density ratio models".
The 2nd Institute of Mathematical Statistics, Asia Pacific Rim Meeting (ims-APRM 2012), Tsukuba, Japan, 2-4 July, 2012.

〔図書〕(計 4 件)
①薩摩順吉・大石進一・杉原正顕, 編, 金森敬文 他 執筆, 応用数理ハンドブック, 項目「アンサンブル学習」, p582-583, 朝倉書店, 2013,

②伏見 正則 (監修, 翻訳), 逆瀬川 浩孝 (監修, 翻訳), 金森 敬文 他 翻訳, モンテカルロ法ハンドブック, 項目「確率的最適化」9 頁, 朝倉書店, 2014,

③杉山 将・井手 剛・神嶌 敏弘・栗田 多喜夫・前田 英作監訳, 金森 敬文 他 翻訳, 統計的学習の基礎―データマイニング・推論・予測―,「モデルの評価と選択」47 頁, 共立出版, 2014

④金森 敬文 著, 統計的学習理論, 講談社, 2015
〔産業財産権〕
○出願状況(計 0 件)

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況(計 0 件)

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
金森 敬文 (KANAMORI Takafumi)
名古屋大学 情報科学研究科・教授
研究者番号：60334546

(2)研究分担者
(　　)

研究者番号：

(3)連携研究者
(　　)

研究者番号：