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　元 年　 ６ 月　 ６ 日現在

機関番号：１２５０１
研究種目：基盤研究(C)（一般）
研究期間：2015～2018
課題番号：１５Ｋ００００４７

研究課題名（和文）高次元セミパラメトリック推測と機械学習

研究課題名（英文）High-dimensional semiparametric inference and machine learning

研究代表者
内藤　貫太（Naito, Kanta）

千葉大学・大学院理学研究院・教授

研究者番号：８０３０４２５２

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：3つのテーマそれぞれで成果を得た。テーマ「パターン認識」では、ナイーブ正準相関係数の高次元漸近理論、歪曲度による統計解析手法の構築が成果となる。テーマ「密度関数の推定」では、頑健な局所密度推定法の構築が成された。さらに、テーマ「回帰関数の推定」では、経験リスク最小化アルゴリズムによる回帰関数の推定法の構築と得られた推定量の理論的評価、説明変数が未知の低次元多様体に埋め込まれている設定でのノンパラメトリック回帰推定量の構築と理論的評価、ＬＭＳＲ法と呼ばれる非線形多変量回帰手法の構築とその応用が成果となる。

研究成果の学術的意義や社会的意義
学術的意義として、まず従来の統計解析手法をより広範なデータに適用可能とするための数理的拡張がなされた点が挙げられる。高次元データや外れ値を含むようなデータへの適用が可能となった。もう1点は、これまでになかった統計解析手法を構築した点である。特に、歪曲度を用いた多次元データの調和度解析や、多次元スタンダード曲線の構築法は、ヒト胎児の発生過程の解析を念頭に考案された。本研究で新たに考案されたこれらの手法により、ヒト胎児の臓器の発生について様々な知見を得ることができた点は、社会的意義となる。

研究成果の概要（英文）：Significant results have been obtained in each of three themes. In the theme "Pattern Recognition", asymptotic results for the naive canonical correlation coefficient have been developed, and a new statistical analysis based on the dilatation has been proposed. In the theme "Density Estimation", a robust version of local density estimation method has been proposed and its theoretical properties have been investigated. Furthermore, in the theme "Regression", an algorithm for regression based on the risk minimization has been considered and the performance of the resultant estimator has been clarified. Nonparametric kernel regression has been shown to work even in the setting where the explanatory variables are embedded into an unknown low dimensional manifold. A new method of nonlinear multivariate regression called the LMSR method has been proposed, and applied to analyze the development process of human fetuses.

研究分野：統計科学、数理統計学

キーワード：セミパラメトリック　関数推定　高次元　機械学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

「関数推定に基づく機械学習と生物統計の横断的研究」（基盤研究（C）23500350；2011～2013年度）のサポートを受け、以下のような研究を推進した経緯がある：

（１）関数推定手法の理論的研究、単純なモデルにおけるスプライン平滑化の漸近理論、核型平滑化とパラメトリック回帰のハイブリッド手法の開発

（２）ブースティングや逐次最小化アルゴリズムに基づく関数推定や、高次元小標本の設定における、パターン認識手法（判別分析）の開発

これらの研究成果を踏まえると同時に、高次元データ解析手法へのニーズが強いことを鑑みて、セミパラメトリック平滑化と、関数推定に基づく機械学習を融合していく研究を推進したいと考えた。その研究枠組みの元、具体的には以下の 3 つの研究テーマを着想するに至った：

①パターン認識：高次元の枠組みでは、従来の線形判別関数などをそのまま用いることができない。その難点を克服する手法を開発するとともに、医学分野へ応用可能な新たな手法を構築する。

②密度関数の推定：高次元データへの応用を視野に入れた、ロバストな密度関数推定の方法論と理論の構築を進め、その高次元データ解析への適用に取り組む。

③回帰関数の推定：回帰関数推定のための「経験リスク最小化アルゴリズム」を構築し、補正関数に基づくセミパラメトリック手法を、そのアルゴリズムに組み込むことを進める。説明変数が高次元の場合に拡張する。

２．研究の目的

テーマ「パターン認識」では、

・高次元小標本の枠組みにおいて有用な、ナイーブベイズ法によるパターン認識と関連する統計量の理論的評価

・ヒト胎児発生過程に見られる「調和」のパターンを“歪曲度”と呼ばれる新たな指標で評価する方法論の確立

テーマ「密度関数の推定」では、

・ロバストな密度推定法の開発とその応用

テーマ「回帰関数の推定」では、

・密度推定で議論した経験リスク最小化アルゴリズムを回帰分析の枠組みに拡張し、非漸近的誤差限界の導出

・説明変数が多次元で，それが実際には低次元に埋め込まれている設定での回帰関数の推定方法の確立とその精度評価

・ヒト胎児発生過程の多変量スタンダード構築のための非線形多変量回帰手法の開発とその精度評価

を目的として掲げた。

３．研究の方法

テーマ「パターン認識」では、高次元小標本の設定で標本共分散行列が特異になることから、その対角成分のみを用いて得られるナイーブ正準相関による判別関数と、現れる統計量の高次元漸近理論を構築する。また、等角写像論で知られている歪曲度を用いた統計解析手法を

構築し、ヒト胎児形態計測データへの応用を通して、その有効性を検証する。

テーマ「密度関数の推定」においては、パラメータを局所的に推定して得られる密度推定量の理論を確立する。特に、密度推定量の構成法を一般的なダイバージェンスを用いて与え、特別なダイバージェンスを用いることにより局所的にロバストな密度推定量が構築されることを示す。

テーマ「回帰関数の推定」では、スプライン関数と経験リスク最小化アルゴリズムを用いた回帰関数の推定方法を確立する。この方法では、推定量の構築だけでなく、用いるスプライン関数の選択も実装できる。説明変数が多次元の重回帰分析の枠組みで、特にその説明変数が低次元の多様体に埋め込まれている場合での回帰関数のノンパラメトリック推定が重要であり、この問題にバイアス縮小の観点から取り組む。非線形多変量回帰を用いたヒト胎児発生過程の多次元スタンダード構築のために、従来から知られていた LMS 法を多変量回帰の枠組みに拡張する。

４．研究成果

補助期間全体を通しての成果をまとめると、テーマ「パターン認識」では、高次元小標本の設定で標本共分散行列の対角成分のみを用いて得られるナイーブ正準相関係数の高次元漸近分布を導出した。ナイーブ判別分析の誤判別確率の漸近上界の構築に応用された（論文②）。等角写像論で知られている歪曲度を用いて対応のある多次元点群の調和度を定義し、ヒト胎児発生過程の調和度の統計解析手法を確立した（論文⑥⑦）。当初の研究目的はほぼ達成された。

テーマ「密度関数の推定」では、パラメータを局所的に推定して得られる密度推定量の方法と理論を確立した。特に、ベキ関数に基づくダイバージェンスを用いて局所的にロバストな密度推定量が構築されることを示すと同時に、従来のパラメトリックな推定量をリスクの意味で漸近的に改善することを証明している（論文③）。着想した研究は着実に進捗があった。

テーマ「回帰関数の推定」では、密度推定で議論されていた経験リスク最小化アルゴリズムを用いた回帰関数の推定方法を構築した。特に、スプラインをワードとした辞書として用いるアルゴリズムとすることで、得られる推定量の非漸近的誤差を導出すると共に、変数選択手法も同時に構築した（論文①）。重回帰分析の枠組みで、その説明変数が未知の低次元多様体に埋め込まれている場合でのノンパラメトリック回帰手法を構築し、特に、データ・シャープニングと呼ばれるバイアス縮小の方法が、この場合でも機能することを示した（論文④）。ヒト胎児発生過程の“多次元スタンダード”を構築のためには、従属変数が多次元である多変量非線形回帰の枠組みが必要であった。従属変数が１次元である非線形重回帰分析で従来から知られていた LMS 法を多変量非線形回帰の枠組みに拡張し、多次元従属変数間の相関構造を考慮した LMSR 法として提案された。これを用いて、ヒト胎児の臓器の発生過程を様々な角度から多次元的に解析することが可能となった（論文⑤）。本テーマでは当初の計画以上の成果が得られた。

５．主な発表論文等

〔雑誌論文〕（計 7 件、全て査読有）

①Takuma Yoshida and Kanta Naito,
Regression with stagewise minimization on risk function.
Computational Statistics and Data Analysis,
in press, https://doi.org/10.1016/j.csda.2018.12.011.

②Mitsuru Tamatani and Kanta Naito,
High dimensional asymptotics for the naive Hotelling T^2 statistic in pattern recognition.
Commnications in Statistics-Theory and Methods,
in press, DOI: 10.1080/03610926.2018.1517217.

③Spiridon Penev and Kanta Naito,
Locally robust methods and near-parametric asymptotics.
Journal of Multivariate Analysis, **167** (2018), 395-417.

④Kanta Naito, Shouta Shimizu, Jun Udagawa and Hiroki Otani,
The LMSR method for providing a multidimensional understanding of growth standard in human fetuses.
Statistical Methods in Medical Research, **27** (2018),2809-2830.

⑤Masaki Kudou and Kanta Naito,
Data sharpening on unknown manifold.
Commnications in Statistics-Theory and Methods, **46** (2017), 11721-11744.

⑥Kanta Naito, Akifumi Notsu, Jun Udagawa and Hiroki Otani,
Statistical analysis with dilatation for development process of human fetuses.
Statistical Methods in Medical Research, **26** (2017),176-200.

⑦Hiroki Otani, Jun Udagawa and Kanta Naito,
Statistical analyses in trials for the comprehensive understanding of organogenesis and histogenesis in humans and mice.
Journal of Biochemistry, **159** (2016), 553-561.

〔学会発表〕（計 13 件）

①内藤　貫太
“ダイバージェンスに基づく局所密度推定の漸近理論”
研究集会　第 20 回ノンパラメトリック統計解析とベイズ統計
慶応大学　2019 年 3 月 26 日

②内藤　貫太
“歪曲度のノンパラメトリック推定”
科研費シンポジウム『多変量データ解析法における理論と応用』
広島大学　2018 年 12 月 15 日

③内藤　貫太
“Nonparametric estimation of dilatation”
統計関連学会連合大会
中央大学　2018 年 9 月 13 日

④Kanta Naito
“Regression on stagewise minimization on risk function”
5th IMS-APRM, 2018, 29th July, Singapore.

⑤内藤　貫太
“歪曲度のノンパラメトリック推定について”
研究集会　第 19 回ノンパラメトリック統計解析とベイズ統計
慶応大学　2018 年 3 月 28 日

⑥内藤　貫太
“Locally robust density estimation and near parametric asymptotics”
科研費研究集会
筑波大学　2017 年 12 月 2 日

⑦吉田拓真、内藤　貫太
“Regression with stagewise minimization on risk function”
日本数学会
山形大学　2017年9月12日

⑧内藤　貫太、スピロ・ペネフ
“Locally robust density estimation and near parametric asymptotics”
統計関連学会連合大会
南山大学　2017年9月4日

⑨Kanta Naito
“Density Estimation with Minimization of U-divergence”
PNU MATH Forum, 1st December 2016, Pusan Naitional University, Korea.（招待講演）

⑩Kanta Naito
“High dimensional asymptotics for the naive Hotteling T^2 statistics in pattern recognition”
China-Japan Joint Workshop on Mathematics and Statistics, 9th October 2016, Northeast Normal University, China.（招待講演）

⑪内藤　貫太
“Kernel naïve Bayes for high dimensional pattern recognition”
統計関連学会連合大会
金沢大学　2016年9月6日

⑫Kanta Naito
“Kernel Naive Bayes for High Dimensional Pattern Recognition”
4th IMS-APRM, 2016, 29th July, Hong Kong.（招待講演）

⑬Kanta Naito
“Kernel Naive Bayes for Highdimensional Pattern Recognition”
International Workshop on Mathematical Sciences in Dalian, 31st October 2015, Dalian University of Technology, China.

６．研究組織

(2)研究協力者
研究協力者氏名：吉田　拓真
ローマ字氏名：Yoshida Takuma

研究協力者氏名：玉谷　充
ローマ字氏名：Tamatani Mitsuru

研究協力者氏名：野津　昭文
ローマ字氏名：Notsu Akifumi