科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２４　日現在

機関番号：　１５２０１
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００３５０

研究課題名（和文）関数推定に基づく機械学習と生物統計の横断的研究

研究課題名（英文）Cross-disciplinary research between machine learning and biostatistics based on curve estimation

研究代表者
内藤　貫太（Naito, Kanta）

島根大学・総合理工学研究科（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）　　4,000,000円、（間接経費）　　1,200,000円

研究成果の概要（和文）：関数推定をツールとした機械学習と生物統計の横断的研究として、深化研究（理論研究）、展開研究（方法論開発）、および応用研究を定めたエフォートに基づき進めた。深化研究では論文5本が出版され、当初の計画以上の進捗を得た。展開研究では、論文3本が出版され、当初の計画通りの成果が得られた。応用研究では、ヒト胎児データへの応用を念頭に、ある種の擬等角写像の歪曲度の極限分布を導出した。また非線形回帰手法であるLMS法を非線形多変量回帰の枠組みに拡張したLMSR法を考案し、ヒト胎児データの解析に果敢に応用した。

研究成果の概要（英文）：As a cross-displinary research between machine learning and biostatistics, I have been addressing Deepened Research (Theoretical Research), Expanded Research(Developing Methodology), and Applied Research, with referring to the given effort. Five papers have been published in Deepened Research, which is the progress more than expected. I have published three papers in Expanded Research, so it certainly got the progress.In Applied Research, asymptotic distribution of dilatation of a certain quasi-conformal mapping was developed, which can be applied toanalysis of human fetus data. The LMS method, which is one of efficent nonlinear regression methods, has been extended to
nonlinear multivariate regression setting. The resultant method is called the LMSR method, and it has been applied to analysis of human fetus data.

研究分野：　統計科学

科研費の分科・細目：　情報学・統計科学

キーワード：　平滑化　関数推定　パターン認識　機械学習　高次元小標本　生物統計　漸近理論　特徴選択

１．研究開始当初の背景

「局所適合セミパラメトリック推測の新展開」（基盤研究（C）20500257）のサポート受け、平滑化の研究を進めた。その研究成果を踏まえ、関数推定をツールとして、以下の 3 つの研究を着想するに至った：

- 深化研究：関数推定手法そのものの理論的考察を深める研究。
- 展開研究：新しい方法の開発を目指す研究。
- 応用研究：胎児形態計測データの解析に関数推定を応用する研究。

全てが関数推定を通して絡み合う。この横断的な研究を推進することを試みた。

２．研究の目的

深化研究では、

- B-スプライン罰則回帰における漸近理論の構築
- 離散応答核型平滑化のセミパラメトリック改良版の挙動の理論構築

展開研究では、

- ブースティングや逐次最小化アルゴリズムに基づく密度推定と回帰手法の開発・提案
- パターン認識におけるフィッシャー判別手法の高次元改良版の挙動

応用研究では、

- ノンパラメトリック平滑化による歪曲度の推定を用いた胎児発生過程の調和度解析
- 次元縮小を含むような様々な関数推定手法による調和度解析

といった内容について成果を得ることを目的とした。

３．研究の方法

深化研究では、B-スプライン罰則回帰の漸近理論を構築する。用いる B-スプライン関数の次数が低い場合、例えば 0 次、1 次の場合の漸近理論の先行研究が存在する。しかしながら、スプライン平滑化が利用される場面では、多くは 3 次のスプラインが用いられる。そのため、一般次数の B-スプライン関数を用いる場合の漸近理論が必要であり、その確立を目指した。また、いわゆる Additive Model にこの結果を拡げ、多次元の説明変数の場合に拡張することを試みる。

展開研究においては、様々な「分布間距離」を特殊な場合として含む U-ダイバージェンスに基づく逐次最小化アルゴリズムによる関数推定法の開発と精度評価を行う。特に、密度推定の議論をまず完成させる。高次元パターン認識の問題については、いわゆるナイーブベイズと呼ばれるアプローチを改良する方法を開発・提案する。その際、正準相関分析から導かれる統計量に基づいた手法が我々のアイデアであり、その挙動を理論的・数値的に調べる。

応用研究においては、胎児発生過程の調和度解析で重要な歪曲度をノンパラメトリックに推定するアイデアを具現化し、その数値的振る舞いを検証した上で、データ解析への利用を検証する。

非線形多変量解析手法を応用し、胎児発生過程の多次元的な理解を与える。特に、発生過程の多次元スタンダードの構築のための新たな非線形多変量解析手法を構築する。

４．研究成果

23 年度の成果として、

深化研究では、B-スプライン罰則回帰の漸近理論を構築した。このような漸近理論は整備が不十分であったが、本研究により、漸近的結果に基づく B-スプライン平滑化での推測手法の適用が可能となった。この研究は更に発展し、説明変数が多次元での加法モデルにおける B-スプライン罰則回帰の漸近的結果の導出につながった。

展開研究においては、高次元小標本の枠組みでのパターン認識の問題について、正準相

関分析から導かれる統計量に基づいた手法を確立し、その挙動を調べた。この結果は論文にまとめ投稿し、採択に至った。この研究では2クラスでのパターン認識問題を議論したが、一般の多クラスの場合へ拡張が進められた。多クラスのパターン認識では、誤判別確率の漸近評価に関しては理論的研究がほとんどなされていない中、誤判別確率の漸近上界を与えた。この結果は2クラスでの結果を数理的に含んでいる。多クラスのパターン認識のための変数選択、いわゆる特徴選択の議論は少ない中、本提案手法に基づく特徴選択の方法も合わせて考案した。多クラスへの拡張が出来たことにより、文字認識、画像認識、遺伝子発現量に基づく疾病種類同定などへの応用が格段に広がった。

応用研究においては、変換の歪曲度をノンパラメトリックに推定するアイデアを具現化し、その理論計算を進めた。この理論計算は、変換のヤコビアンのノンパラメトリック推定に該当し、その汎関数の推定まで見据えると多くの応用を持つものである。また、様々なパラメトリック擬等角写像について、その歪曲度を理論的に求め、その推定量を開発した。

24 年度の成果として、

深化研究では、B-スプライン罰則回帰の漸近理論の研究を更に進めた。説明変数が多次元での加法モデル、離散応答を含めた一般化加法モデルにおけるB-スプライン罰則回帰の漸近的結果を導出し、1本の論文が出版され、1本が投稿済となった。また、セミパラメトリック回帰の枠組みでのB-スプライン罰則回帰の漸近的結果も導出し、1本の論文として出版した。

展開研究においては、高次元小標本の枠組みでのパターン認識の問題について、正準相関分析から導かれる統計量に基づいた手法を確立し、その挙動を調べた。この結果は1本の論文として出版された。この論文での理論を多クラスのパターン認識の問題へ拡張した。多クラスのパターン認識において、誤判別確率の漸近上界を与え、さらに特徴選択の手法を開発し、その性能をシミュレーションおよび実データへの適用を通して評価した。U ダイバージェンスの逐次最小化による密度推定に関する論文1本が出版された。

応用研究においては、スプライン平滑化のエレガントな応用であるLMS法の多次元化に取り組み、実用段階に達している。この多次元化により、胎児発生の多次元的スタンダードが得られることになる。

25 年度の成果として、

深化研究では、罰則付きスプライン平滑化の研究を更に進めた。特に、一般化線形回帰の枠組みにおける、罰則付きスプライン平滑化の漸近理論に関する論文が 1 本出版された。

展開研究では、多群の枠組みにおいて、正準相関に基づくパターン分析手法を提案し、高次元小標本の設定の元でその漸近挙動を理論的に調べた。特に、多群におけるナイーブ正準相関を考案し、ナイーブ正準相関ベクトルの一致性についての理論的結果は他に類を見ないものである。また、判別方向ベクトルに関してもその漸近挙動に関する結果を得ている。更に、多群高次元における特徴選択アルゴリズムを考案し、その有効性を数値実験と実データへの適用を通して確認した。これらの結果をまとめ、1 本の論文として発表、出版された。

応用研究では、生物の成長を記述するのに有効な非線形回帰手法である LMS 法を多変量に拡張する研究を進めた。正値データをアウトプットとする回帰分析では、アウトプットのべき変換を用いるのが有効であり、LMS 法においてもある種のべき変換を適用し、そこに含まれるパラメータを罰則付き尤度で推定する。多変量アウトプットへの拡張においては、多変量のべき変換を上手く導入する必

要があった。LMS 法を利用しつつ、相関構造を導入することで多変量べき変換を定義し、LMS 法を拡張したのが、考案した LMSR 法である。数値実験を通して機能することを確認し、胎児形態計測データへの適用を通して、LMS 法では抽出できなかった情報が抽出できることがわかり、その結果を論文にまとめている。

補助期間全体を通しての成果をまとめると、深化研究では、B-スプライン罰則回帰の漸近理論を加法モデル、一般化線形加法モデルについて構築し、セミパラメトリックな枠組みでの漸近理論についても理論的結果を導出した。それらの成果は論文①⑥⑦である。関数推定手法の理論研究のその他の成果は論文④⑧となる。深化研究（理論研究）は当初の計画以上に進捗があったと考えている。

展開研究では、U－ダイバージェンスの逐次最小化に基づく密度推定に関する研究の成果が論文③として、またナイーブ正準相関に基づく 2 群パターン認識手法の提案とその漸近理論に関する成果が論文⑤、更にナイーブ正準相関に基づく手法の多群への拡張およびその漸近理論に関する成果は論文②として公表された。当初の研究目的はほぼ達成されたと考えている。

応用研究では、胎児形態計測データの解析への応用を念頭に、擬等角写像の歪曲度の計算を蓄積した。ある種の歪曲度の推定量に関する漸近分布を導出した（学会発表④）。非線形回帰分析手法の LMS 法を、非線形多変量回帰の枠組みに拡張した LMSR 法を開発し、それを用いて胎児形態計測データの解析を行い、成果をまとめた論文を現在作成中である。論文としての成果は今後になるが、着想した研究は着実に進捗があった。

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８件）

① Takuma Yoshida and Kanta Naito
Asymptotics for penalized splines in generalized additive models.
Journal of Nonparametric Statistics, **26**, (2014), 269-289.
DOI: 10.1080/10485252.2014.899360

② Mitsuru Tamatani, Kanta Naito and Inge Koch
Multi-class discriminant function based on canonical correlation in high dimension low sample size.
Bulletin of Informatics and Cebernetics, **45**, (2013), 67-101.
http://bic.math.kyushu-u.ac.jp

③ Kanta Naito and Shinto Eguchi
Density estimation with minimization of U-divergence.
Machine Learning, **90**, (2013), 29-57.
DOI: 10.1007/s10994-012-5298-3

④ Junmei Jing, Inge Koch and Kanta Naito
Polynomial histograms for multivariate density and mode estimation. *Scandinavian Journal of Statistics*, **39**, (2012), 75-96.
DOI: 10.1111/j.1467-9469.2011.00764.x

⑤ Mitsuru Tamatani, Inge Koch and Kanta Naito
Pattern recognition based on canonical correlations in a high dimension low sample size context.
Journal of Multivariate Analysis, **111**, (2012), 350-367.
DOI: 10.1016/j.jmva.2012.04.011

⑥ Takuma Yoshida and Kanta Naito
Asymptotics for penalized additive B-spline regression.
Journal of the Japan Statistical Society, **42**, (2012), 81-107.
http://www.jss.gr.jp/ja/journal/index.html

⑦ Takuma Yoshida and Kanta Naito
Semiparametric penalized spline regression.
Bulletin of Informatics and Cybernetics, **44**, (2012), 65-86.
http://bic.math.kyushu-u.ac.jp/

⑧ Masaru Kanba and Kanta Naito
Selection of smoothing parameter for one-step sparse estimates with Lq penalty.
Journal of Data Science, **9**, (2011), 565-584.
http://www.jds-online.com/

①～⑧全て査読有

〔学会発表〕（計５件）

①内藤貫太、玉谷充、Inge Koch
高次元小標本における naïve canonical correlation に基づく多群判別－特徴選択
統計関連学会連合大会
2013 年 9 月 11 日, 大阪大学

②Kanta Naito and Shinto Eguchi
Density estimation with minimization of U-divergence
IMS-APRM2012
2012 年 7 月 4 日、筑波

③Kanta Naito, Mitsuru Tamatani and Inge Koch
Pattern recognition based on canonical correlation in high dimension low sample size context
科研費シンポジウム「多変量解析の新展開」
2012 年 1 月 20 日、那覇市

④内藤貫太、野津昭文、宇田川潤、大谷浩
Statistical Analysis with Dilatation for development Process of Human Fetuses
科研費シンポジウム「生命科学と統計学」
2011 年 11 月 5 日、大阪大学

⑤内藤貫太、玉谷充、Inge Koch
高次元小標本における正準相関に基づくパターン認識―実際的側面
統計関連学会連合大会
2011 年 9 月 6 日、九州大学

６．研究組織
(1)研究代表者
内藤　貫太　（Naito, Kanta）
島根大学大学院総合理工学研究科・教授
研究者番号：80304252

(2)研究分担者
吉田　拓真（Yoshida, Takuma）
鹿児島大学大学院理工学研究科助教
研究者番号：80707141