様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年 5 月 15 日現在

機関番号：３２６９０
研究種目：基盤研究（Ｂ）
研究期間：平成 21 年度～平成 23 年度
課題番号：21330049
研究課題名（和文）高次元データの推測理論の開発と応用

研究課題名（英文）Statistical Inference for High-Dimensional Data and Its Applications

研究代表者　杉山　高一（SUGIYAMA TAKAKAZU）
創価大学・比較文化研究所・客員教授
研究者番号：70090371

研究成果の概要（和文）：多変量解析においては、変量の数が大きい場合の高次元データの分析法を発展させることが重要になっている。本研究では、変量の数が標本数より小さい場合には伝統的手法について高次元漸近理論を開発し、変量の数が標本数より大きい場合には高次元特有の方法の導入と高次元漸近理論の開発を目指し成果を得た。また、経済学への応用や関連する統計的基礎理論についても研究した。具体的には、以下の課題に取り組み、成果をあげた。
（1）高次元伝統的多変量手法の開発
（2）高次元現代多変量手法の開発
（3）高次元モデリング手法の開発
（4）シミュレーションによる研究と応用
（5）高次元計量経済統計手法の開発と応用"

研究成果の概要（英文）：研究成果の概要（英文）：In multivariate analysis, it is important to develop the statistical method to analyze the high-dimensional data when the number of variables is large. In this study, we have also constructed a high-dimensional asymptotic theory for the traditional method when the number of variables is smaller than the number of observations. The aim of our study is to develop the introduction of high-dimensional method and the method of high-dimensional asymptotic theory when the number of variables is greater than the number of observations. We also applied our method and the statistical development of high-dimensional asymptotic theory in economics.
More specifically, the challenges of the following, we have achievements.
（1）Development of traditional multivariate methods for high-dimensional data
（2）Development of modern multivariate methods for high-dimensional data
（3）Development of high-dimensional modeling techniques
（4）Research and the applications based on statistical simulation
（5）Development and applications of high-dimensional statistical econometric

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
21 年度	2,900,000	870,000	3,770,000
22 年度	2,400,000	720,000	3,120,000
23 年度	2,400,000	720,000	3,120,000
総　計	7,700,000	2,310,000	10,010,000

研究分野：経済学
科研費の分科・細目：計量経済学
キーワード：(1) 高次元多変量データ　(2) 高次元推測理論　(3) 高次元漸近理論
(4) 高次元モデリング手法　(5) シミュレーション　(6) 計量経済統計

１．研究開始当初の背景

近年、電子化された測定技術の進歩により科学の諸分野で、現象過程の連続的な計測・測定と、大量なデータの蓄積が可能になっている。この結果、大規模なデータ（ビッグデータ）の統計解析や、標本数より変量の数の方が大きい高次元小標本データの解析などが重要になっている。実際、米国国立科学財団主催のワークショップ「統計科学：21 世紀に対する挑戦と機会」の報告書(2003) においても、当面の重要課題は大規模データ分析であると述べている。また、そのようなデータ分析にアプローチする上で重要な課題を６つ挙げているが、その中の一つが高次元小標本データ問題である。分担者藤越は、これらについて「21 世紀の統計学への挑戦的課題と展望」(業績欄を参照) と題して、より詳しい解説を行なっている。

研究代表者杉山・分担者藤越は、最近の業績に見られるように、高次元の場合の主成分分析や判別分析に取り組んで来ている．また、Ledoit and Wolf (Ann. Statist., 30, 1081-1102, 2002), Schott (J.Multivariate Anal., 97, 827-843, 2006) などにおいても関連した研究が行なわれている。これらの先駆的な研究を通して、高次元データ解析の研究では、(a) 変量の数 p は、標本数 n より小さいが、大標本と比べ比較的大きい場合、
(b) 変量の数 p が標本数 n より大きい場合、
の両方の場合を研究する必要があると考えている。(a) の場合の研究はそれ独自に重要であると共に、(b) の場合において変数選択により(a) の場合に帰着されたりするからである。さらに、高次元漸近理論においては、主として
(c)$\lim p/n \to \gamma \in (0, 1)$ または$(0,\infty)$
という枠組で発展させることも重要である。(a) の場合、１つのアプローチとして伝統的な統計量に対して高次元漸近理論を展開することになるが、僅かな経験に基づくものであるが、このような結果が大標本漸近理論の改良になっており、このことが他の統計量についても期待されることにも注目している。

２．研究の目的

多変量解析においては、変量の数が大きい場合の高次元データの分析法を発展させることが重要になっている。このため、本研究では、変量の数が標本数より小さい場合には伝統的手法について高次元漸近理論を開発し、変量の数が標本数より大きい場合には高次元特有の方法の導入と高次元漸近理論の開発を目指す。また、応用や関連する統計的基礎理論についても研究する。

具体的には、上記(a)、(b)、(c) の研究に焦点を当てながら、高次元データ解析のための推測理論と応用を発展させることを目的にしている。とくに、(a) の場合には、これまでの伝統的な手法が利用できるが、従来の大標本漸近理論は利用できないので、(c) のような高次元漸近的枠組での理論を発展させる。(b) の場合には、高次元特有の手法を提案すると共に、それらの統計的性質を高次元の立場から明らかにする。さらに、各手法の信頼性を数値的に検証し、実データへの応用を試みることを目的としている。

３．研究の方法

研究の方法では、以下のそれぞれの課題に取り組んでいる。

（１）高次元伝統的多変量手法の開発

多変量解析の伝統的手法である、主成分分析、判別分析、多変量検定、正準相関分析、経時データ解析、などを取り上げる。主成分分析、判別分析、多変量検定に関しては、ある程度の成果が得られているが、手法の比較などは十分でない。また、すべての場合、多変量正規性のもとでの成果であり、これを多変量非正規の場合に展開する。また、ブートストラップ法、パーミュテーションテスト等を取り入れた分析法あるいは結果の信頼性についても研究する。

（２）高次元現代多変量手法の開発

因子分析や共分散構造モデルなどは潜在変数モデルとして、統一的に扱うことができる。このようなモデル対し、伝統的手法の高次元漸近理論を展開する。また、高次元特有の方法の開発に取り組む。分担者狩野は、潜在変数モデルに関して多くの成果を（研究業績を参照）発表しており、この課題に対して準備ができている。この研究では、多変数データの構造方程式モデリングの QOL 評価への応用も試みる。さらに、サポートベクターマシーン、独立成分分析、クラスター分析、大量データ（ビッグデータ）を用いたデータマイニング手法の予後予測の問題、など現代多変量手法の高次元への拡張も目指している。

（３）高次元モデリング手法の開発

高次元の場合、重要な変数を選び出すための変数選択法は一層重要になる。このため、大標本理論に基づく AIC 基準などを高次元漸近理論に基づく基準に拡張・発展させ、新たな高次元基準にもとづく変数選択法を提案する。また、複雑な自然現象や社会現象を解明するための非線形モデリング手法の開発と、その過程におけるモデル評価法についても研究する。この研究には研究業績欄に掲載した藤越、小西によるいくつかの論文が基になる。

（４）シミュレーションによる研究と応用

本研究で開発された高次元データの分析法において、縮約された情報量(統計量) の特性を知ることは極めて重要である。この統計量の挙動を調べるシミュレーションアルゴリズムの研究を行う。また、見出されたいくつかの統計量の中で、どの統計量が優れているかを絞り込むこと、それが時間をかけて研究するに値する統計量であるか否かを知るのに、研究の初期段階での統計的シミュレーション実験は有効である。これを当該研究分野での研究に生かし、さらに統計的シミュレーションによって分析法の頑健性を含めた分析結果の信頼性を研究する。また、ＤＮＡデータなどの実データ分析に取り組む。この研究には研究業績欄に掲載した杉山等によるいくつかの論文が基になる。

(5) 統計的・数学的基礎理論の開発

本研究課題の展開において、多変量高次元標本分布論の研究が必要になるが、これを、精密論、大標本漸近理論、高次元漸近理論、ブートストラップ法、パーミュテーションテストなどのアプローチから研究する。また、最近では計算機による数値計算の発展とともに、新たなアルゴリズムが応用されるようになって来ているが、その数学基礎についても研究する。この研究には研究業績欄に掲載した杉山、藤越、小西、若木等によるいくつかの論文が基になる。

４．研究成果

多変量解析においては、変量の数が大きい場合の高次元データの分析法を発展させることが重要になってきている。このため、本研究では、変量の数が標本数より小さい場合には伝統的手法について高次元漸近理論を開発し、変量の数が標本数より大きい場合には高次元特有の方法の導入と高次元漸近理論の開発を目指した。研究目的に沿った５つの研究課題、高次元伝統的多変量手法の開発、高次元現代多変量手法の開発、高次元モデリング手法の開発、シミュレーションによる研究と応用、高次元計量経済統計手法の開発と応用について、それぞれ役割分担を定めて取り組んだ。５つの課題は高次元データの分析法として有機的に結びついており、研究代表者、研究分担者が必要に応じて適宜集まって総合的に研究することにより、それぞれの研究課題における最新の問題を確認し、お互いに協力しながら当該研究目的を遂行した。また、経済学への応用や関連する統計的基礎理論について研究した。以下では各課題ごとに主要研究成果を述べる。

(1) 高次元伝統的多変量手法の開発に関する研究成果: 高次元漸近的枠組のもとで正準相関係数の漸近展開を導出し、これまでの大標本漸近展開近似の大幅な改良を与えた。また、2 母集団における共分散行列の第 j 番目に大きい固有値の同等性検定を行うための検定統計量を提案し、一般の母集団のもとで検定統計量の極限分布を導出した。遺伝子型データ等の高次元データの多変量解析、ある高次元多変量モデルにおける尤度比統計量の漸近分布の導出等に関しても成果を上げた。

(2) 高次元現代多変量手法の開発に関する研究成果: 母数間に制約のある潜在構造モデルにおける新たな推定値の算出方法を提案した。それは EM アルゴリズムの変形版で、正しい MLE への収束可能性を大幅に高めることに成功した。高次元漸近的枠組のもとで判別分析、質的データの判別分析、主成分正準相関係数の変数選択問題で、有用な変数選択基準等を求めた。また、遺伝子型データの特性とその高次元多変量解析を研究し、高次元主成分分析における固有値・固有ベクトルの推定法を提案した。

(3) 高次元モデリング手法の開発に関する研究成果:大規模モデルの推定と変数選択の新しい手法である lasso 推定を拡張して、非線形現象のモデル化に適用する研究に取り組み、新たな非線形モデリング手法を提唱した。高次元データ解析に関する検定統計量の漸近分布に対する共分散構造分析、あるガウスモデルの下での関数クラスター分析等で成果を得ている。

(4) シミュレーションによる研究と応用に関する研究成果:モデル化の過程において、ベイズ推論によって計算機上で実行するアルゴリズムを組み込むことによって、従来、適切に捉えることが困難であった局所的に変動する高次元の現象や変化点を自然に取り組むことができる柔軟なモデリング手法となった。欠測値のある場合にどのような対応をとるのが適切か調べて、欠測値の処理の方法論の優劣をシミュレーションによって研究した。ロジスティック回帰モデルにおける回帰係数の推測問題で新しい提案を行った。さらに、動画像に基づく移動オブジェクトの大きさに関する統計的レジストレーション等の研究を行い成果を得た。

(5)高次元計量経済手法の開発に関する研究成果:国全体のマクロ経済などを説明する場合に用いられる同時方程式モデルにおけるパラメータの推定量の性質について考察した。2 段階最小二乗法、制限付き最尤推定法などの推定量に対して、標本数と外生変数の数が共に大きくなると言う高次元漸近的枠組のもとで、漸近展開を導出し漸近的性質を明らかにした。さらに、国全体のマクロ経済などを説明する場合に用いられる動的パネル構造方程式の推定問題について統計的性質を明らかにし、動学的パネルデータ分析について多くの知見を得た。

次に、上記で得られた成果の国内外におけ

る位置づけとインパクト、今後の展望などについて記述する。主要成果は、Journal of Multivariate Analysis 、 Journal of Statistical Planning and Inference 、 Communications in Statistics; Simulation and Computation、Journal of Forcasting、などの国外の統計学分野の国際的雑誌に掲載されている。また、国内の統計分野の代表的雑誌である Journal of Japan Statistical Society、応用統計学、などに掲載されている。このことにより、各成果は国内外において高い評価を得ている。とくに、高次元漸近理論に関しての成果は、その有効性と斬新性が高く、統計分野の理論的発展に大きく貢献するものと確信している。

最近、高次元問題に関連した研究において、多変量回帰モデルにおける情報量規準 AIC などが高次元漸近的枠組みのもとで一致性をもつという、斬新な結果を示している。今後、このような研究を他の多変量モデルで発展させることを考えている。また、多変量非線形問題にも高次元の立場から重点的に取り組むことが期待される。

なお、コンパクトな研究集会を２回行い、年度の最後（2012 年 1 月）に全体の研究集会を開催し、研究成果を広く公開した。研究成果を纏めた冊子、杉山高一編集（2012 年 2 月）「Proceedings of Statistical Inference for High-Dimensional Data and Its Applications」（絢文社）を作成し当該分野の研究者に配布した。また、藤越康祝・杉山高一著（2012 年 2 月）「多変量モデルの選択」（朝倉書店）を出版し、100 冊ほどを当該分野の研究者に配布した。これらは研究成果の学界及び社会への普及に役だち、高次元データの推測理論と応用のさらなる発展に繋がる成果である。

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　57 件）

1. Fujikoshi Y., Satoh T. and Sugiyama T.、Asymptotic distribution of the contribution ratio in high-dimensional principal component analysis 、American Journal mathematical and Management Sciences、査読有、2012 (accepted)
2. 竹田 裕一、暗号で用いる乱数検定における新しい手法の提案、Proceedings of Statistical Inference for High-Dimensional Data and Its Applications、査読無、2012、90-93
3. Tsukada, S.、Unbiased estimator for a covariance matrix under two-step monotone incomplete sample 、Communications in Statistics, Theory and Methods、査読有、2012 (accepted)
4. Kano Y. and Takai K.、Analysis of NMAR missing data without specifying missing-data mechanisms in a linear latent variate model、Journal of Multivariate Analysis、査読有、102 巻、2011、1241-1255
5. Seo T., Sakurai,T. and Fujikoshi, Y.、LR tests for two hypotheses in profile analysis of growth curve data、SUT Journal of Mathematics、査読有、47 巻、2011、105-118
6. 大倉征幸, 鎌倉稔成、小標本かつ応答変数発現確率が高い場合のロジスティック回帰モデルにおける回帰パラメータの検定法、応用統計学、査読有、23 巻 2011、41-51
7. Konishi, S. 、Nonlinear regression modeling via the lasso-type regularization、Journal of Statistical Planning and Inference、査読有、140 巻、2010、1125 – 1134
8. Kamakura, T. and Murakami, H. 、A saddlepoint approximation to the limiting distribution of a k-sample Baumgartner statistic、Journal of the Japan Statistical Society、査読有、39 巻、2010、133-141
9. Yamamoto, T. 、Forecasting in Large Cointegrated Systems 、Journal of Forecasting、査読有、26 巻、2009、631-650
10. Kunitomo, N. 、Asymptotic Expansions and Higher Order Properties of Semi-Parametric Estimators in a System of Simultaneous Equations、Journal of Multivariate Analysis、査読有、100 巻、2009、1727-1751
11. Kano, Y. 、Simple computation of maximum likelihood estimates in latent class model with equality and constant constraints、Communications in Statistics - Simulation and Computation、査読有、38 巻、2009、654-665
12. Murakami, H. and Sugiyama, T. 、Permutation Test for Equality of Individual an Eigenvalue from a Covariance Matrix in High-Dimension、Communications in Statistics - Simulation and Computation、査読有、38 巻、2009、1675-1689
13. Fujikoshi, Y. and Sakurai, T. 、High-dimensional asymptotic expansions for the distributions of

canonical correlations、J. Multivariate Anal.、査読有、100 巻、2009、231-242

〔学会発表〕（計　55 件）

1. Takeda, Y., Hashiguchi, H. and Sugiyama, T.、Numerical computation on distributions of statistics expressed by generalized hypergeometric function、ISI2011、2011 年 8 月、アイルランド
2. Fujikoshi Y.、High-Dimensional Approximations of the Coefficients on Linear Discriminant Functions and Canonical Correlation Variates 、ISI2011、2011 年 8 月、アイルランド
3. Kano, Y.、Bias of the Direct MLE for NMAR Missingness: Theoretical Approach、IMPS2011、2011 年 7 月、香港
4. Kamakura, T.、Inference for the Coefficient Parameters of the Logistic Regression from a Small Sample、、ISI2011、2011 年 8 月、アイルランド
5. Tsukada, S. and Yamada, T.、Asymptotic distribution for latent root of covariance matrix under two-step monotone incomplete data、、ISI2011、2011 年 8 月、アイルランド
6. Hiroaki Chigira and Taku Yamamoto、The Effect of Estimating Parameters on Long-Term Forecasts for Cointegrated Systems 、19th International Conference on Computational Statistics、2010 年 8 月、フランス
7. Jan Dolinsky、Konishi, S.、Echo State Networks with Non-Monotonous Activation Functions、2010 統計関連学会連合大会、2010 年 9 月、早稲田大学
8. 狩野裕、井上高継、無視不可能な欠測におけるＭＬＥのバイアス評価、2010 年度統計関連学会連合大会、2010 年 9 月、早稲田大学
9. 佐藤 整尚、国友 直人、景気判断と平滑化問題：GDP 公表値を巡って、2010 統計関連学会連合大会、2010 年 9 月、早稲田大学
10. 鎌倉 稔成、小椋 透、大草 孝介、スポーツ解析と統計教育、2010 統計関連学会連合大会、2010 年 9 月、早稲田大学
11. Sugiyama, T.、Approximations of percentile point of individual latent root on Wishart distribution、ISI2009、2009 年 8 月、南アフリカ
12. 藤越康祝、同時方程式モデルにおける高次元漸近理論、統計関連学会、2009 年 9 月、同志社大学

〔図書〕（計　6 件）

1. 藤越康祝、杉山高一、朝倉書店、多変量モデルの選択、2011、208
2. 國友直人、朝倉書店、構造方程式と計量経済学、2011、218
3. 山本拓　他、知泉書館、動学的パネルデータ分析、2011、338
4. 菅民郎、藤越康祝、現代数学社、質的データの判別分析：数量化 2 類、2011、264
5. Yasunori Fujikoshi, Vladimir V. Ulyanov, Ryoichi Shimizu 、Wiley 、Multivariate Analysis: High-dimensional and Large-Sample Approximations、2010、533
6. 小西貞則、岩波書店、多変量解析入門 - 線形から非線形へ - 、2010、320

〔産業財産権〕
○出願状況（計　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◇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
杉山 高一（SUGIYAMA TAKAKAZU）
創価大学・比較文化研究所・客員教授
研究者番号：70090371

(2) 研究分担者
藤越 康祝（FUJIKOSHI YASUNORI）
広島大学・名誉教授
研究者番号：40033849

研究分担者
山本 拓（YAMAMOTO TAKU）
日本大学･経済学部･教授
研究者番号：10150031

研究分担者
鎌倉 稔成（KAMAKURA TOSHINARI）
中央大学･理工学部･教授
研究者番号：40150031

研究分担者
狩野 裕（KANO YUTAKA）
大阪大学･基礎工学研究科･教授
研究者番号：20201436

研究分担者
村上 秀俊（MURAKAMI HIDETOSHI）
防衛大学校･専任講師
研究者番号：60453677

研究分担者
塚田 真一（TUKADA SHINNICHI）
明星大学･教育学部･教授
研究者番号： 10319022

研究分担者
竹田 裕一（TAKEDA YUICHI）
神奈川工科大学・教育センター・准教授
研究者番号：90349241

研究分担者
酒折 文武（SAKAORI FUMITAKE）
中央大学･理工学部･准教授
研究者番号：60453677

(3)連携研究者
国友 直人（KUNITOMO NAOTO）
東京大学経･済学研究科
研究者番号：10153313

小西貞則（KONISHI SADANORI）
中央大学･理工学部･教授
研究者番号：40150031