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　２ 年　 ６ 月　 ８ 日現在

機関番号：１４２０１
研究種目：若手研究(B)
研究期間：2016～2019
課題番号：１６Ｋ１６０１８

研究課題名（和文）大標本高次元データに対するノンパラメトリック手法の開発

研究課題名（英文）Development of nonparametric methods with large sample sizes and high-dimension

研究代表者
姫野　哲人（HIMENO, Tetsuto）

滋賀大学・データサイエンス学部・准教授

研究者番号：４０４５２７３４

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：高次元データ（変数の数が多いデータ）の解析手法の多くは、データの分布が正規分布に従うという強い仮定の下で提案されたものが多かったが、近年、そのような強い条件が無くても使用な可能な手法の開発が急速に行われてきている。しかし、それらの研究の中には現実的でない仮定のものや、特定の項目を調べることに特化した手法も多く、一般化された手法の開発は行われていなかった。そこで、本研究では制約をできる限り緩和したうえで、数多くの仮説検定を包含した一般的な手法の開発を行った。

研究成果の学術的意義や社会的意義
近年、コンピューターやセンサーの性能向上により、多様なデータを同時に扱う機会が増えてきた。このようなデータが得られたとき、そのデータの背後にある性質（平均など）を調べることは重要である。しかし、古典的な方法は変数の数が増えたとき、または、データを発生させる構造（分布）が複雑なケースのときには適用できない。古典的な方法において制約されていた様々な条件を緩和することで、ビッグデータに対しても対応可能な手法を開発した。提案手法は、古典的な手法の一般化・拡張であり、学術的にも意義のあるものである。

研究成果の概要（英文）：Statistical methods for high-dimensional data used to have a strong condition that the distribution is normal. In recently years, several methods are developed under a condition with non-normal distribution. However, most of these methods specialize in a particular construction and the condition is not pragmatic. A comprehensive method has never been developed. Thus, I studied comprehensive statistical tests under a weak condition for high-dimensional data.

研究分野：数理統計学

キーワード：漸近理論　高次元データ　ノンパラメトリック

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

情報通信技術の発展により、科学研究、健康・医療、経済、社会、環境、ウェブなどのあらゆる分野でデータがリアルタイムに集積されるようになり、ビッグデータとよばれるデータが注目を集めるようになってきた。ビッグデータの分析は国内外、分野問わず注目されており、**ビッグデータから価値ある情報を抽出することは、社会的にも重要な課題である。**しかし、多くの場合、これらのデータはほとんど分析されないままであることが多い。**大規模データ（標本サイズも変数の数（種類）も膨大なデータ）に対する分析手法の開発**は急務である。

２．研究の目的

本研究では、ビッグデータに対する分析手法の 1 つとして、**大標本高次元データに対し、分布等の仮定をできる限り緩めた手法の開発を目的とする**。これにより、各種データに対し、分析前の事前調査がほとんど必要なくなり、誤った手法を用いることによるミスリーディングを減らすことができる。この目的は、これまでに研究されてきた大標本高次元に対する手法は、正規性が仮定されているものや、正規性が仮定されていなくともそれに類する強い仮定が設定されているものも多く、一方でそのような仮定が満たされているかどうかのチェック方法も確立されていなかったということに基づいている。

３．研究の方法

本研究ではビッグデータの解析手法に焦点を当て、分布を問わない分析手法の開発を行うことが目的である。そのために、先行研究である Himeno and Yamada (2014) 及び Yamada and Himeno (2015) で行った手法を応用する。これらの論文では、評価関数（平均ベクトルに関する検定では、平均ベクトルの大きさに関する関数）に平均ベクトルの推定量を代入し、そのことにより発生するバイアス（ずれ）を修正し、基準化することで新たな検定手法を提案している。これを、線形モデルにも応用し、**線形モデルの係数に関する検定手法の提案**を行う。

４．研究成果

(1) 高次元データに関する代表的な研究の一つに線形判別問題がある。高次元データに対する線形判別問題では、特定の条件下では一致性をもつ（誤判別確率が 0 に収束する）ことが知られている。しかし、一致性を持たないケースでは、判別境界をどのように設定するかが重要となる。

2 標本問題（観測データ $\boldsymbol{x}$ が群 Π_1 と Π_2 のどちらかから得られているという状況）において、線形判別関数 W に対し、判別ルール $W > c \Rightarrow x \in \Pi_1$（$W < c \Rightarrow x \in \Pi_2$）を考える。一般には、2 つの誤判別確率が等しくなる（$\mathrm{P}(\mathrm{W} < \mathrm{c} \mid x \in \Pi_1) = \mathrm{P}(\mathrm{W} > \mathrm{c} \mid x \in \Pi_2)$となる）ように判別境界 c を設定することが多い。しかし、判別問題において、2 種類の誤判別の意味が等しいことは少ないであろう。例えば、ある病気の罹っているかどうかの判定を行う場合、病気に罹っていない人を罹っていると判断することと、病気に罹っているにも関わらず罹っていないと判断することは同列には扱えない（後者のミスは命にかかわるため、その確率はできる限り小さくするべきである）。もちろん、一方についてないがしろにしてよいわけではない。

そこで、（正規性の条件は仮定するが）一方の誤判別確率 $\mathrm{P}(\mathrm{W} < \mathrm{c} \mid x \in \Pi_1)$ を漸近的に指定した値（0.05 や 0.01 など）に収束するという状況下で、もう一方の誤判別確率 $\mathrm{P}(\mathrm{W} > \mathrm{c} \mid x \in \Pi_2)$ がどのような値に収束するかどうかを導出した（Yamada et al., 2017）。この結果から、2 種の誤判別確率をともに設定値以下とするために必要なサンプルサイズを得ることが可能となる。

(2) 高次元データに対する正規性の検定はこれまであまり進んでこなかった。この要因としては、正規性の検定ではしばしば 3 次モーメントや 4 次モーメントの推定量に基づき検定が行われるが、高次元データに対するこれらのモーメントの推定が困難なことによると考えらえる。Koizumi et al. (2014)や Hanusz et al. (2017)では、正規性の下で 3 次モーメント（の 2 乗）や 4 次モーメントの推定量を導出し、正規性の検定を行っている。しかし、これらの統計量が非正規性の下でどのような振る舞いをするかは検討されていない。Himeno and Yamada (2014)では、非正規性の下での 4 次モーメントの一致推定量を導出し、これに基づく正規性の検定を導出した。しかし、この方法では 3 次モーメントを用いていないため、分布のゆがみに対応できず、いくつかの分布に対する検出力が不十分であった。

そこで、非正規性の下での 3 次モーメントの一致推定量を導出し、新たな正規性の検定を導出した（Yamada and Himeno, 2019）。数値実験の結果から、提案手法は様々な分布に対して十分な検出力が得られることを確認した。

(3) これまで高次元データに対する研究は多くの研究者によって行われており、多標本問題（Cai et al., 2014, Chen et al., 2019）、線形仮説問題（Zhou et al., 2017）、two-way MANOVA（Zhou et al.,

2020）などが挙げられる。しかし、これらの手法はすべてケースバイケースであり、統一的な議論を行えていない問題がある。そこで、本研究では GMANOVA モデルにおける線形仮説問題に対する検定統計量を導出した。提案手法は過去に提案されている様々な検定統計量（多標本問題、ANOVA モデル、profile モデル等）に帰着し、かつ、さらに一般的な問題への拡張を行った。本成果については、現在投稿中である。

<引用文献>

・Cai, T. T., Liu, W., and Xia, Y.(2014), Two-sample test of high dimensional means under dependence, *Journal of the Royal Statistical Society: Series B Statistical Methodology*, **76**, 349-372.

・Chen, S. X., Li, J., and Zhong, P. S.(2019), Two-sample and ANOVA tests for high dimensional means, *Annals of Statistics*, **47**, 1443-1474.

・Hanusz, Z., Enomoto, R., Seo, T., and Koizumi, K.(2018), A monte Carlo comparison of Jarque-Bera type tests and Henze-Zirkler test of multivariate normality, *Communications in Statistics – Simulation and Computation*, **47**, 1439-1452.

・Himeno, T. and Yamada, T.(2014), Estimates for some functions of covariance matrix in high dimension under non-normality and its applications, *Journal of Multivariate Analysis*, **130**, 27-44.

・Koizumi, K. and Hyodo, M., and Pavlenko, T.(2014), Modified Jarque-Bera Type Tests for Multivariate Normality in a High-dimensional Framework, *Journal of Statistical Theory and Practice*, **8**, 382-399.

・Yamada, T. and Himeno, T.(2014), Testing homogeneity of mean vectors under heteroscedasticity in high-dimension, *Journal of Multivariate Analysis*, **139**, 7-27.

・Yamada, T., Himeno, T., and Sakurai, T.(2017), Asymptotic cut-off point in linear discriminant rule to adust the misclassification probability for large dimension, *Hiroshima Mathematical Journal*, **47**, 319-348.

・Yamada, T., and Himeno, T.(2019), Estimation of multivariate 3rd moment for high-dimensional data and its application for testing multivariate normality, *Hiroshima Mathematical Journal*, **47**, 319-348.

・Zhou, B., Guo, J., and Zhang, J. T. (2017), High-dimensional general linear hypothesis testing under heteroscedasticity, *Journal of Statistical Planning and Inference*, **188**, 36-54.

・Zhou, B., Guo, J., and Zhang, J. T. (2020), An L^2-norm based test for high-dimensional two-way MANOVA, *Science Sinica Mathematica*, **50**, 729-750.

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3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 Yamada Takayuki、Himeno Tetsuto	４．巻 34
２．論文標題 Estimation of multivariate 3rd moment for high-dimensional data and its application for testing multivariate normality	５．発行年 2019年
３．雑誌名 Computational Statistics	６．最初と最後の頁 911～941
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s00180-018-00865-9	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yamada Takayuki、Himeno Tetsuto、Sakurai Tetsuro	４．巻 46
２．論文標題 Interval estimation in discriminant analysis for large dimension	５．発行年 2017年
３．雑誌名 Communications in Statistics - Theory and Methods	６．最初と最後の頁 9042～9052
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/03610926.2016.1202282	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yamada Takayuki、Himeno Tetsuto、Sakurai Tetsuro	４．巻 47
２．論文標題 Asymptotic cut-off point in linear discriminant rule to adjust the misclassification probability for large dimension	５．発行年 2017年
３．雑誌名 Hiroshima Mathematical Journal	６．最初と最後の頁 319～348
掲載論文のDOI（デジタルオブジェクト識別子） なし	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4件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 姫野哲人、山田隆行
２．発表標題 一般化した分布の下での高次元MANOVA問題
３．学会等名 2017年度統計関連学会連合大会
４．発表年 2017年

１．発表者名
姫野哲人

２．発表標題
高次元データに対する検定手法の開発

３．学会等名
統計サマーセミナー2017

４．発表年
2017年

１．発表者名
Takayuki Yamada, Tetsuto Himeno, Tetsuro Sakurai

２．発表標題
Asymptotic cut-off point in linear discriminant rule which adjusts misclassification probabilitywhen the dimension is relatively large compared to the sample sizes

３．学会等名
2016 International Conference for JSCS 30th Anniversary in Seattle（国際学会）

４．発表年
2016年

１．発表者名
山田隆行、姫野哲人

２．発表標題
高次元データに対する 3 次モーメントを使った正規性の診断

３．学会等名
第30回日本計算機統計学会シンポジウム

４．発表年
2016年

〔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