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　６ 年　 ４ 月　 ２ 日現在

機関番号：３２７０２
研究種目：基盤研究(C)（一般）
研究期間：2020～2023
課題番号：２０Ｋ１１７１２

研究課題名（和文）標本分布の歪みに対処した新たな高次元統計解析の開発

研究課題名（英文）Development of new high-dimensional statistical analysis to deal with skewness of sample distribution

研究代表者
兵頭　昌（Hyodo, Masashi）

神奈川大学・経済学部・教授

研究者番号：００７１１７６４

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：高次元統計解析における仮説検定において、正規近似に基づく近似検定が数多く提案されている。これらの検定は、次元pが1000～10000のように膨大であれば、十分な精度を有することが明らかにされている。一方で、次元pが10～500のように中程度のデータにおいては、検定統計量の分布に歪みが生じるため、正規近似が機能しないという問題点がある。このような問題に対して、いくつかの解析的な方法を応用することで、分布の歪みへ対処した新しい近似検定法を提案した。

研究成果の学術的意義や社会的意義
高次元データにおける近似的な仮説検定の多くは、中心極限定理を利用した漸近的な精度保証を行っている。しかし、漸近理論と有限次元のデータに乖離があるため実用性と説得性に欠ける。そこで、本研究では、エッジワース展開や検定統計量の適切な変換を与えることでより正確な漸近分布を導出する。このようなアプローチは古典的な大標本統計学ではよく用いられるが、高次元データにおいては十分に研究されているとは言えないため、古典的な多変量解析における漸近理論を大幅に発展させる可能性があると期待できる。

研究成果の概要（英文）：Many approximate tests based on normal approximation have been proposed for hypothesis testing in high-dimensional statistical analysis. It has been revealed that these tests have sufficient accuracy when the dimension p is very large, such as 1000 to 10000. On the other hand, there is a problem that normal approximation does not work for data with medium dimension p, such as 10 to 500, because the distribution of test statistics is distorted. To address these problems, we proposed a new approximation test method that deals with distribution distortion by applying several analytical methods.

研究分野：統計科学

キーワード：高次元データ　正規化変換　　誤差限界　多重比較　多変量分散分析　　歪度　漸近正規性　一致性

様　式　C－１９、F－１９－１（共通）

１．研究開始当初の背景

様々な媒体、経路を通じて大規模データが、驚くほど低コストで入手できるようになった現在、多変量解析手法に対する学術界やビジネス界からのニーズは非常に高まっている。しかしながら、伝統的な多変量解析手法の多くは、直接には、大規模データへは応用できない困難な点が横たわっている。その典型的な問題点は、「高次元データ小標本問題」である（図 1:実データ）。

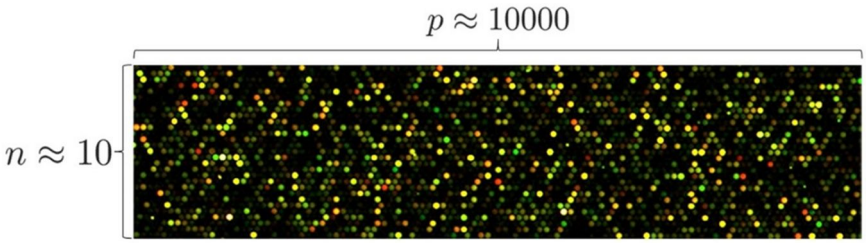

図 1：発現量を蛍光光度計で測定した DNA マイクロアレイデータ。このデータは、次元数（遺伝子数） p が約 1,000～50,000 程度あるのに対し、サンプルサイズ（被験者数） n はわずかに 10 ～100 程度であり、
高次元小標本問題がおきる代表的なデータである。

このような 「p（次元）>>n（サンプルサイズ）」 の仮定の下で正当化される研究は、2000 年以前は統計学の全般の分野に渡ってほとんど存在しなかった。通常の統計理論の枠組みでは、次元 p は一定のもとで n→∞ である漸近論が基本であったが、この仮定は高次元データでは全く役立たない。2000 年代になり、高次元データの研究が徐々に進み、既存の統計解析の限界が理論的に示され、新たな統計学の必要性が認識され始めた。2010 年代に入って、理論と応用の両面から統計学が飛躍的に向上し、新たな統計学として高次元統計解析が誕生した。高次元における検定で利用される検定統計量の多くは、p→∞であるとき漸近正規性（極限分布が正規分布）が成り立つ。この性質を応用し、有限次元においては、正規分布を近似分布として用いることとなる。高次元における検定理論では、このような正規近似が主流であり、次元が 1,000～10,000 程度であれば実用上十分な精度を有することが既に明らかにされている。一方で、次元が 10～500 程度（中程度）の場合は、高次元統計解析における検定統計量の実際の分布は、正規分布に比べて歪みをもつため正規近似の近似精度が極端に悪化するという問題がある（図２を参照）。

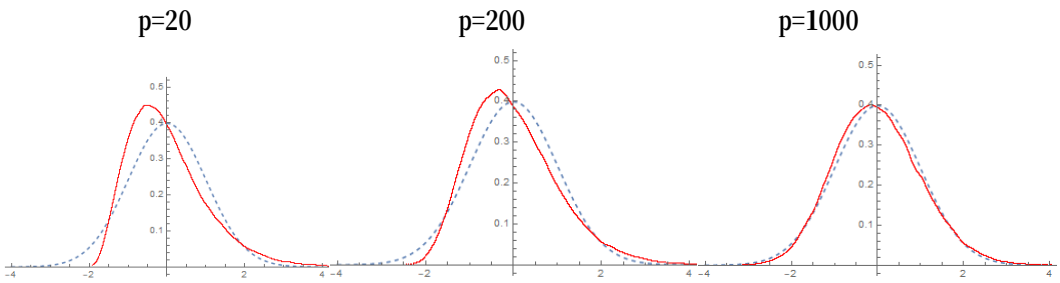

図 2:高次元統計解析におけるデンプスター検定統計量の次元 p を変化させたときの分布の概形をシミュレーションで算出した図である。正規分布（点線）と検定統計量の実際の分布（実線）を表している。

２．研究の目的

次元が 10～500 程度（中程度）の場合に対して、従来の正規近似に代わる分布の歪みに対処した高精度な近似分布を提示し、それを応用した幾つかの仮説検定問題における近似検定を提案

することを目的とする。

３．研究の方法

目的のための１つのアプローチとして、モーメントマッチングやエッジワース展開を応用し、中程度の次元において分布の歪みに対処した新たな近似分布を導出し、その近似精度を理論的あるいは数値的に確認し、各検定問題へ応用する。

本研究では、特に下記の５つの研究課題について研究を実施した：

(A)　平均ベクトルの同等性検定

(B)　多変量分散分析

(C)　(A)で導出した結果を応用した多変量多重比較法

(D)　プロファイル分析

(E)　分散共分散行列の同等性検定

尚、各課題（A）～（E）において、以下の3点を可能な限り研究した。

(i)　提案近似検定が従来の正規近似を収束レートの意味で改善していることを理論的に明らかにする

(ii)　漸近検出力関数を導出する

(iii)　数値シミュレーションによる有限次元・有限標本における理論的結果の精度検証を行う

４．研究成果

- 研究課題（A）について、L2 型検定統計量の正規化変換を導出し、それが分布の歪みを補正する近似になっていること、従来の正規近似を収束レートの意味で改善していることを理論的に示した。さらに、漸近検出力関数を導出し、数値シミュレーションによる有限次元・有限標本における理論的結果の精度検証を行った。これらの結果は学術誌 Communications in Statistics – Theory and Methods へ掲載された。高次元における仮説検定に関する研究の多くは、中心極限定理を利用した漸近的な精度保証を行っている。しかし、漸近理論と有限次元のデータに乖離があるため実用性と説得性に欠ける。そこで、本研究では、より実用的な「非漸近的な誤差評価」を精度保証として利用している点が特徴である。ここで、「非漸近的な誤差評価」とは、「有限の p に対する真の分布と近似分布の誤差限界を与える。」ことである。
- 研究課題(B)について、高次元における 2 元配置分散分析法を提案した。この結果は、学術誌 Journal of Multivariate Analysis へ掲載された。先行研究として、1 元配置分散分析に関する研究は存在するが、より煩雑な問題である 2 元配置分散分析法へ拡張することに成功した。今後は、多元配置への拡張が考えられる。
- 研究課題(C)および(D)に関して、研究課題 (A)を応用した方法を提案し、研究成果を纏め、学術誌への投稿するための論文を作成した（Communications in Statistics – Theory and Methods へ投稿を検討中)。
- については、分布の歪みが生じる１つの状況として、母集団へある種のファクターモデルを仮定した場合を想定し研究を行った。母集団へある種のファクターモデルを仮定した場合、等分散性の検定に用いられる従来のＬ２ノルム型の検定統計量（より正確には従来のＬ２ノルム型の検定統計量へ適当な補正を施した検定統計量）が高次元において重み付きカイ２乗分布へ収束することを理論的に証明した。さらに、この漸近分布に現れる重みを、高次元因子モデルの推測理論を応用し推測することにより新たな近似検定を提案した。また、局所対立仮説の下での漸近的な検出力関数の導出も行った。これらの漸近的な結果を確認するために、有限次元・有限標本における第１種の過誤確率や検出力を調べるためのシミュレーションを実施した。シミュレーションの結果から、正規近似が機能しないような状況において提案法の第 1 種の過誤確率と名目有意水準が一致することを確認した。これらの研究成果を、統計関連学会連合大会および日本計算機統計学会第 37 回シンポジウムにて報告した。
- その他、判別分析（高次元データにおける判別分析の変数の冗長性に関する研究）や課題以外の仮説検定問題（ρV 係数に基づく独立性の検定、一般化分散の同等性検定）についても研究成果を上げることができた。

５．主な発表論文等

〔雑誌論文〕　計8件（うち査読付論文　7件／うち国際共著　1件／うちオープンアクセス　0件）

１．著者名 Hyodo Masashi、Nishiyama Takahiro、Pavlenko Tatjana	４．巻 195
２．論文標題 A Behrens-Fisher problem for general factor models in high dimensions	５．発行年 2023年
３．雑誌名 Journal of Multivariate Analysis	６．最初と最後の頁 105162～105162
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jmva.2023.105162	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 該当する

１．著者名 Hyodo Masashi、Watanabe Hiroki、Nakagawa Shigekazu、Nakagawa Tomoyuki	４．巻 52
２．論文標題 Normalizing transformation of Dempster type statistic in high-dimensional settings	５．発行年 2022年
３．雑誌名 Communications in Statistics - Theory and Methods	６．最初と最後の頁 8096～8113
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/03610926.2022.2056749	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 Nakagawa Tomoyuki、Watanabe Hiroki、Hyodo Masashi	４．巻 184
２．論文標題 Kick-one-out-based variable selection method for Euclidean distance-based classifier in high-dimensional settings	５．発行年 2021年
３．雑誌名 Journal of Multivariate Analysis	６．最初と最後の頁 104756～104756
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jmva.2021.104756	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 Watanabe Hiroki、Hyodo Masashi、Nakagawa Shigekazu	４．巻 179
２．論文標題 Two-way MANOVA with unequal cell sizes and unequal cell covariance matrices in high-dimensional settings	５．発行年 2020年
３．雑誌名 Journal of Multivariate Analysis	６．最初と最後の頁 104625～104625
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jmva.2020.104625	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 Hyodo Masashi、Nishiyama Takahiro、Pavlenko Tatjana	4．巻 178
2．論文標題 Testing for independence of high-dimensional variables: ρV-coefficient based approach	5．発行年 2020年
3．雑誌名 Journal of Multivariate Analysis	6．最初と最後の頁 104627～104627
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jmva.2020.104627	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 Hyodo Masashi、Nishiyama Takahiro、Pavlenko Tatjana	4．巻 157
2．論文標題 On error bounds for high-dimensional asymptotic distribution of L2-type test statistic for equality of means	5．発行年 2020年
3．雑誌名 Statistics & Probability Letters	6．最初と最後の頁 108637～108637
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.spl.2019.108637	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 Watanabe Hiroki、Hyodo Masashi、Sugiyama Takatoshi、Seo Takashi	4．巻 52
2．論文標題 Test for equality of standardized generalized variance with different dimensions under high-dimensional settings	5．発行年 2022年
3．雑誌名 Hiroshima Mathematical Journal	6．最初と最後の頁 217-233
掲載論文のDOI（デジタルオブジェクト識別子） 10.32917/h2021025	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 Masashi Hyodo, Takahiro Nishiyama, Hiromasa Hayashi	4．巻 TR21-03
2．論文標題 High-dimensional multiple comparison procedures among mean vectors undercovariance heterogeneity	5．発行年 2021年
3．雑誌名 Technical Report, Statistical Research Group	6．最初と最後の頁 1-23
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 無
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 －

〔学会発表〕　計10件（うち招待講演　1件／うち国際学会　5件）

１．発表者名
Takahiro Nishiyama, Masashi Hyodo, Tatjana Pavlenko

２．発表標題
A two sample Behrens-Fisher problem for factor models in high dimensions

３．学会等名
International Symposium on New Developments of Theories and Methodologies for Large Complex Data（招待講演）（国際学会）

４．発表年
2021年

１．発表者名
Takahiro Nishiyama, Masashi Hyodo

２．発表標題
On the multiple comparison procedures among mean vectors for high-dimensional data under covariance heterogeneity

３．学会等名
International Conference on Econometrics and Statistics（国際学会）

４．発表年
2021年

１．発表者名
中川智之，渡邉弘己，兵頭昌

２．発表標題
ユークリッド距離に基づく判別分析の変数選択について

３．学会等名
2021年度応用統計学会年会

４．発表年
2021年

１．発表者名
兵頭昌，渡邉弘己，中川重和

２．発表標題
Normalized transformation of Dempster type statistic in high-dimensional setting

３．学会等名
日本計算機統計学会第34回大会

４．発表年
2020年

１．発表者名
米口貴誠，首藤信通，兵頭昌

２．発表標題
楕円母集団から得られた2-step単調欠測データに基づく平均ベクトルの尤度比検定と検出力について

３．学会等名
日本計算機統計学会第34回シンポジウム

４．発表年
2020年

１．発表者名
兵頭昌，西山貴弘，渡邉弘己，中川智之，田畑耕治

２．発表標題
Tests for the equality of covariance matrices under a low dimensional factor structure

３．学会等名
日本計算機統計学会 第37回シンポジウム

４．発表年
2023年

１．発表者名
兵頭 昌，西山 貴弘，中川 智之，田畑 耕治，渡邉 弘己

２．発表標題
高次元枠組みにおける分散共分散行列の同等性

３．学会等名
統計関連学会連合大会

４．発表年
2023年

１．発表者名
Tomoyuki Nakagawa, Hiroki Watanabe, Masashi Hyodo

２．発表標題
Kick-one-out-based variable selection method for Euclidean distance-based classifier in high-dimensional settings

３．学会等名
6th International Conference on Econometrics and Statistics (EcoSta 2023)（国際学会）

４．発表年
2023年

１．発表者名
Takahiro Nishiyama, Masashi Hyodo

２．発表標題
On a general linear hypothesis testing problem for latent factor models in high dimensions

３．学会等名
International Symposium on Recent Advances in Theories and Methodologies for Large Complex Data（国際学会）

４．発表年
2023年

１．発表者名
Takahiro Nishiyama, Masashi Hyodo

２．発表標題
Linear hypothesis testing on mean vectors for factor model in high-dimensional settings

３．学会等名
International Conference on Econometrics and Statistics（国際学会）

４．発表年
2023年

〔図書〕　計1件

１．著者名	４．発行年
兵頭 昌、中川 智之、渡邉 弘己	2022年
２．出版社	５．総ページ数
共立出版	208
３．書名	
よくわかる！Rで身につく 統計学 入門	

〔産業財産権〕

〔その他〕

https://scholar.google.com/citations?user=r7r9T-AAAAAJ&hl=en

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
研究協力者	西山　貴弘 (Nishiyama Takahiro)		
研究協力者	渡邉　弘己 (Watanabe Hiroki)		
研究協力者	中川　智之 (Nakagawa Tomoyuki)		

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関