科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１３　日現在

機関番号：　２２７０１
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７００３４３

研究課題名（和文）非正則な多次元データにおける統計理論の研究

研究課題名（英文）Research on statistical theory in non-regular multi-dimensional data

研究代表者
小泉　和之（KOIZUMI, KAZUYUKI）

横浜市立大学・総合科学部・助教

研究者番号：７０５４８１４８

交付決定額（研究期間全体）：（直接経費）　　1,800,000円、（間接経費）　　540,000円

研究成果の概要（和文）：欠測値を含む統計解析理論と得られた多次元データが正規分布に従うかを調べる多変量正規性検定理論の研究を行った。本研究において欠測値を含む理論研究では数多く存在する多変量解析理論の中でも判別分析に注目し、その中でも欠測値を含む場合には分散パラメータの推定に問題に関して、特に欠測のパターンが単調である場合に限ってではあるが陽に計算できる最尤推定量を導出し、それを用いて線形判別関数の期待誤判別率の漸近近似を与えた。
また、正規性検定では近年注目されている高次元データにおける検定統計量の漸近近似の改良を行った。さらに分散パラメータの推定についてはgLasso推定を用いた方法の改良も行った。

研究成果の概要（英文）：We consider the multivariate statistical theory with missing data and the multivariate normality test. In particular we focus on discriminant analysis with missing data. We derive an estimator of variance parameter with monotone missing data. And by using this estimator, we improve the asymptotic approximation for the EPMC of linear discriminant function.
Further, we give two types of improved statistic for assessing multivariate normality in high-dimensional data. And we propose a new estimator for variance-covariance matrix in high-dimensional data by using AIC and gLasso estimator.

研究分野：　総合領域

科研費の分科・細目：　情報学・統計科学

キーワード：　多変量解析　漸近論

１．研究開始当初の背景
ここ数年では、遺伝子データ、画像データなどにあげられるように情報の多様化が急速に進んでいき、データの多次元化が進み多変量解析の手法にはこれまでよりも注目が集まっていた。また、それに付随してデータの一部が欠落する（欠測値という）現象もしばしば起こっている。それらの問題は未だに不十分な部分が多く、実用上も早い解決が望まれていた。

２．研究の目的
背景にあげたような経緯から本研究では以下の２点に絞って研究を進めることにした。
（１）各標本はすべてのデータがそろっているという前提を外し、欠測値を含むもとでも適用可能な統計解析手法の開発を行う。

（２）一般的な多変量解析手法ではデータが正規分布しているという正規性の仮定が置かれていることが多いが、背景にもあげたように情報が多様化している状況ではデータはそもそも正規分布するのかも怪しい。すなわち多次元データが正規分布しているかを調べる多変量正規性検定の理論研究を行う。

３．研究の方法
目的にあげた（１）（２）のどちらかを満たしていないデータを非正則データと呼ぶが、これら非正則なデータにおいて（１）を満たしていない、つまり欠測値が含まれている状況下で検定統計量の研究を行う。具体的には本研究で扱う統計量（分散パラメータの推定量）は母集団の共分散構造に仮定を置くものが多い。そこでその共分散構造の仮定が妥当であるかを調べるために欠測値を含まない状況では通常の尤度比検定を行うことによってなされている。本研究では欠測値を含む場合に同様の議論が可能であるかを検証し、より適用範囲の広い方法を理論、シミュレーションの両面から行っていく。さらに、検定理論だけに限定せず、必要とされている様々な多変量解析手法についても欠測値を含む場合にも使える方法や漸近近似の導出も行う。

また、目的（２）にあげた多次元データの正規性検定については以下のように行う。情報の多様化の影響の１つとしてデータの次元が大きくなる「データの高次元化」がある。それらの研究はここ数年で数多くの成果が出ており必要とされている。そこで本研究ではまずは高次元データ（ただし次元数は得られた標本数を超えない）のもとでの正規性検定理論の研究を行う。これは従来の代表本理論の成果では次元数と標本数が比較的近いときには数値実験の結果からも近似精度が悪くなっているという傾向がみられるためである。多くある多変量正規性検定統計量の中でもまずは Koizumi et al. (2009)で提案されている MJB 統計量に注目し、得られた結果を他の統計量との比較も行っていく。ここからはこれまで記載した内容次第ではあるが次元数が標本数を超えるような高次元データについても理論研究を進めていきたいと考えている。具体的にはこのような高次元データでは分散共分散行列の逆行列が存在しないという問題点を解決することが必要である。その逆行列の存在性を保証するためにさまざまなアプローチがあるがそれらのなかでも妥当なものを正規性検定理論に適用するもしくは新たな方法を提案することも考えている。

４．研究成果
欠測値を含む統計解析の研究成果から記載する。こちらは検定理論や共分散構造の箇所ではとくに学術論文としての結果までという大きな成果は得られていないが、この問題に関しては特に学会発表の①にある成果が大きい。Fujikoshi and Seo (1998)で与えられている結果は高次元データにおける判別分析の理論研究として非常に学術的価値の高いものではあるのだが欠測値を含んだ高次元データのもとではこれらの方法は用いることが出来なくなってしまう。これは分散パラメータの推定が行えなくなるという問題があるからである。これに対する一般的な解答はもちろん難しいのであるが本研究においてはこれらの結果を欠測値のパターン（メカニズムは一般に MAR を仮定する）をもっとも単純な 2step 単調欠測という場面に限って、陽に計算することができる最尤推定量の導出を行い、判別分析に広く用いられている Fisher の線形判別関数の誤って判別してしまう確率の期待値（期待誤判別確率、EPMC という）の漸近近似を与えることに成功した。これを従来の完全データに用いられている Fujikoshi and Seo (1998)の結果との比較も行い数値的にも改良されていることを示した。これに関してはより一般の k-step 単調への拡張を現在は進めている段階である。また、これらの漸近近似の根底にあるのは分布関数の近似公式の導出を行うことによってなされるのであるがこの導出の際には通常、高次のモーメント計算が必要であるがこちらの計算はある程度の次数になってしまうとかなり煩雑なものになってしまうという傾向になり近似精度の改良という観点からだと厳しい方法となることも確認できた。それにより、今後の学会などで成果を公表する予定ではあるが高次のモーメント計算を必要としない微分演算子なるものを用いた漸近展開公式の導出も行っている。これにより必要とする近似精度が得られない場合でもさらなる改良が得られるのではないかという期待も大きい。

また、もう一つの大きな研究内容としては多変量データの正規性検定である。こちらはこ

れまでにも様々な方法が提案されており、統計解析において重要な問題であることがわかる。応用上も通常の統計理論のみならず時系列解析などにも使われており幅の広い研究内容である。そういった観点からも検定に用いる統計量は近似精度が高いものである必要はもちろんあるのだが、実装のしやすさという観点も必要となる。そこで本研究においてはモーメントをもとにした検定理論に注目した。特に分布の歪みを表現すると言われている３次のモーメント（歪度）、分布の尖り具合を表現する４次のモーメント（尖度）の２つを用いた検定理論の研究を精力的に行った。これら２つのモーメントは多次元データでなければ通常は１つの定義しかなくそれらを用いるのが通例である。しかし、べき乗計算がいろいろ考えられる多次元データにおいてはこれらの定義もやはり様々である。有名なところでは本研究でも取り扱う Mardia (1970)、Srivastava (1984)で提案されているものなどがある。Mardia (1970)の歪度、尖度を用いた統計量と Srivastava (1984)のものを用いた統計量では検出力の観点からは Mardia (1970)のものの方が優れている。正規性検定は適合度検定であるので検出力の高さは評価基準としては妥当である。しかし、Mardia (1970)のものは次元が 30 を超えた辺りから計算コストの影響か計算に時間がかかりシミュレーションを行う上で大きな支障をきたす。それに対して Srivastava (1984)のものは主成分スコアを元にした指標であるため次元数にそれほど影響せず比較的演算時間は少なくて済むというメリットもある。そこで本研究ではこれら２つの歪度、尖度を用いた。雑誌論文の①ではこれら２つ（尖度のみだが）の共通する性質というものを理論的に絞り出すことに成功し、これら２つの尖度を包括したような新たな多変量尖度を提案し、それを用いた統計量の漸近分布の導出に成功している。この統計量は Mardia (1970)も Srivastava (1984)の尖度も含まれているので検出力という観点からも良いパフォーマンスを出してくれるのではないかと期待できる。
また、歪度は歪み具合、尖度は尖り具合を表すものであると述べたが従来はこれら２つを別々に検定に用いていたという背景がある。しかし、分布は複数の観点からの判断が重要であるため Koizumi et al. (2009)で提案されている総括的な検定統計量の改良を試みた。これは雑誌論文③、④で扱われている内容であるが、注目している多次元データは次元数が標本数を超えることはないが次元はある程度大きい場合も想定している。特に Koizumi et al. (2009)で提案している統計量は漸近分布への収束が遅く、少ない標本しか得られないような状況下では近似精度が良くないことが数値的に示されている。そこで高次元のもとでもかつ標本数がある程度少ないときにでも近似精度の改良を試みる。具体的には③においては Cornish-Fisher 展開を用いた正規化変換を用いることによって漸近バイアスを補正した検定統計量の提案をしている。それにより収束する速さが改良され標本数がそれほど多くなくても近似の精度が改良されることを示している。しかし、次元数と標本数がある程度近い場合にはその近似精度はそれほど改良されてはおらずその改良法として統計量の期待値を導出することによって極限分布であるカイ二乗分布で近似するのではなく、F 分布を用いた近似法を提案している。さらに④では分布の意味で収束の早いと言われている Wilson-Hilferty 変換という正規化変換を用いた方法を提案している。それにより④の方法はさらに③の改良となっていることも数値的に示されている。

最後に②であるがこれは高次元データにおける歪度、尖度に関する成果であるがこれまでと違うのは次元の数は標本数よりも多いような状況を考えているという点である。これらの成果はまだ少なく②によって新たに提案している。そこでは共分散構造にブロック対角構造なるものを仮定することによって逆行列が存在しなくなるという問題点を解決している。しかし、実際にこの歪度、尖度を用いるためには真の共分散構造がどのようなブロック対角構造なのかを知る必要が出てくる。そこで本研究では Pavlenko et al. (2012)で提案されている gLasso 推定を用いた方法を適用することを考えた。しかし、gLasso 推定だけではブロック対角構造には推定してくれずただのスパース行列を与えることになる。そこで得られたスパース行列から候補の構造（モデル）を選び出し、それら候補のモデルの中からリスクの意味で最も適している共分散構造を情報量基準である AIC を使うことによってブロック対角構造を選び出す方法の提案をしている。

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　４　件）
①Miyagawa, C., Koizumi,K. and Seo,T. A new multivariate kurtosis and its asymptotic distribution, SUT Journal of Mathematics, 査読有, 47, 2011, 55-71
②Sumikawa,T.,Koizumi,K. and Seo,T. Measures of multivariate skewness and kurtosis in high-dimensional framework, Hiroshima Statistical Research Group Technical Report, 査読無, 2013
③澄川琢磨、小泉和之、瀬尾隆 正規化変換を用いた総括的な多変量正規性検定統計量の改良、計算機統計学、査読有、26、2013、33-41
④Koizumi,K.,Hyodo,M. and Pavlenko,T.

Modified Jarque-Bera type tests for multivariate normality in a high-dimensional framework, Journal of Statistical Theory and Practice,査読有,8, 2014, 382-399

〔学会発表〕（計　8　件）
①小泉和之、首藤信通
Asymptotic approximation for the EPMC of the linear discriminant function based on two-step monotone missing data under a high-dimensional framework、2011 年度統計関連学会連合大会、2011 年 9 月
②齋藤めい、澄川琢磨、小泉和之、瀬尾隆
On some tests for assessing multivariate normality based on sample skewness and kurtosis、経験尤度法と判別・分類解析の理論と応用、2011 年 12 月
③澄川琢磨、小泉和之、瀬尾隆
新たな多変量尖度を用いた正規性検定統計量の提案、2012 年度統計関連学会連合大会、2012 年 9 月
④小泉和之、澄川琢磨
標本積率を用いた多変量正規性検定について、科研費シンポジウム「統計科学における深化と横断的展開」、2012 年 10 月
⑤澄川琢磨、小泉和之、瀬尾隆
Cornish-Fisher 展開を用いた多変量 JB 型統計量の改良、日本計算機統計学会第 26 回シンポジウム、2012 年 11 月
⑥小泉和之
共分散構造の近似法、第 11 回統計研究会、2013 年 3 月
⑦脇村真隆、三ツ井誠、小泉和之、瀬尾隆
高次元データに対する多変量歪度及び尖度について、日本計算機統計学会第 27 回シンポジウム、2013 年 11 月
⑧脇村真隆、三ツ井誠、小泉和之、瀬尾隆
ブロック対角化構造法による多変量歪度及び尖度について。科研費シンポジウム「統計的推測の新展開とその応用」、2013 年 12 月

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　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　0　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
小泉和之（KOIZUMI, Kazuyuki）
横浜市立大学・国際総合科学部・助教
研究者番号：70548148

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：