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　 ７ 月　 ７ 日現在

機関番号：３２６２９
研究種目：若手研究(B)
研究期間：2015～2019
課題番号：１５Ｋ１５９５３

研究課題名（和文）多変量解析における正規性および高次元モデル選択の推測と応用

研究課題名（英文）Multivariate normality test, and inference for variable selection in high-dimensional model and its applications

研究代表者

榎本　理恵（Enomoto, Rie）

成蹊大学・理工学部・助教

研究者番号：３０７１１７６７

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：歪度と尖度を用いたJB統計量タイプの多変量正規性検定についての研究を行った．数値実験はHenze-Zirkler統計量との比較を行い，モンテカルロシミュレーションによって，検定統計量の第一種の過誤とさまざまな分布による検出力を計算した．
成長曲線モデルのもとでのモデル選択規準に基づく変数選択の一致性について導出した．取り扱う規準量はAIC，BIC，Cpやそれらを修正した規準量である．特性については高次元データの枠組みのもとと，グループ数と標本数の比率が１より小さい一定値に近づくという枠組みのもとで導出を行った．

研究成果の学術的意義や社会的意義
昨今，データが一般に手に入りやすくなったこともあり，分析手法に正規性の仮定が必要である場合にはまず分析をする前にデータを吟味する必要がある．また，変数の次元が大きい場合の統計的方法や多量の変数の中から有用な変数を選び出す変数選択方法には高い需要が生まれてきている．これらはデータ分析における重要な課題であることから基礎研究に限らず，その応用研究にも取り組むことで統計科学における理論や応用に対する貢献が期待される．

研究成果の概要（英文）：In this study, tests for multivariate normality (MVN) of Jarque-Bera type, based on skewness and kurtosis, have been considered. Simulation results have been compared to the Henze-Zirkler's test. In the Monte Carlo simulations, we calculated empirical Type I errors of tests under consideration, and empirical power against different alternative distributions.
In this study, we derived consistency properties for variable selection methods based on model selection criteria in growth curve model. The model selection criteria treated includes AIC criterion, BIC criterion and Cp criterion, or their modified criteria. The properties were derived under a high-dimensional asymptotic framework and a large-(q;n) asymptotic frame work such that the ratio of the number of group to the sample size tends to afixed number less than 1.

研究分野：多変量解析

キーワード：統計科学

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

昨今，データサイエンスが分野を問わず色々な場面で取り上げられている．さまざまな統計モデルや分析手法は“正規分布”を仮定していることが多い．それは正規分布が取り扱いし易く，標本数が十分に確保できる状況下で，データは漸近的に正規分布に従うと仮定できるという理由からである．しかし，実際に分析するデータが正規分布であるかどうかを吟味する分析者は多くはない．

一変量の場合，視覚的な判断から正規分布に従うかどうかの判断をすることが可能である．しかし多変量の場合，Q-Q プロットによる視覚的な判断が不可能であり，かつ特性関数に基づく統計量というのは使い手にとって取扱いし難いといった背景がある．そこで使い手が容易に検定できるよう積率に基づく検定方法が従来から提案されている．特に，大標本理論（標本数が十分大きい）に対して，Mardia や Srivastava が定義した歪度と尖度が有名であり，これは３次と４次積率に基づいて定義されている．

これまで研究代表者は，Srivastava による歪度と尖度に関する統計量の提案を行ってきた．先行研究で提案されてきた統計量は，歪度のみや尖度のみを使用し，一方の特性しか見ていない統計量であった．そこで申請者は歪度と尖度を同時に用い，一度の検定で歪度と尖度両方の特性に注目することができる簡便な統計量を提案した．Srivastava による定義は主成分スコアが基になっているためにデータの一部分しか用いていないという欠点があった．一方で，Mardia による定義は一変量の定義の自然な拡張ではあるものの，次元数が大きくなると統計量の分布の自由度が巨大になる欠点があり，次元に対しては脆弱な統計量であることが周知されている．したがって，積率に基づく統計量は簡便ではあるが上記のような問題，また分布の特性全てを反映できないゆえに検定の面（仮説検定結果が保留となった場合の解釈）で正確性に欠ける問題があった．そこで，簡便な統計量と実際に使用される統計量との吟味および，既存の統計量の改良などについて目指す．

成長曲線モデルというのは，例えば，女子と男子の成長（身長）が１次直線，２次曲線と予測できるようなモデル化のことである．したがって，各個体の時間変化に対して適当な曲線を当てはめ，更に各個体の繰り返し測定に未知の分散共分散行列をもつ多変量正規性などを想定するモデルである．AIC 規準は真のモデルを含む多項式モデルを正しく選択する規準として古くから使われている．Kobayashi，Satoh and Fujikoshi は成長曲線モデルに対する大標本理論の下での規準量を提案しており，研究代表者はこれを高次元データに適用した場合，誤ったモデルを選択することを確認している．更に，高次元データに対して新たに提案した規準量は，大標本データおよび高次元データの両方に対して正しい変数選択することを確認している．

AIC 規準量は大標本理論の下で一致性を持たないことが問題とされている．「**一致性がある**」とは，標本数を増やしていった場合に必ず真のモデルを選択する確率が１になることである．実データを取り扱う研究者は，候補のモデルの当てはまりの良さについて AIC 規準を用いて確かめることが多々ある．しかし，AIC に一致性がないこと，真のモデルより大きいモデルを選ぶことを意識していない．また，提案されている多くの規準量に実際の値から順位（優劣）をつけることは難しい．そのため，多くの規準量について特性を議論し，世間に広めていくことは重要な課題である．そして自然な拡張として，観測する「グループ数の個数と標本数が共に増加する」場合に着目し，いくつかの仮定を付与した場合に一致性があることを確認している．さまざまな規準量の特性や一致性についての議論は重要であり更なる拡張を考えより実用的な枠組みでの議論を目指す．

２．研究の目的

研究代表者は，正規性検定と多項式モデルにおける変数選択に関する研究を主とする．これまでに，標本数が次元数（説明変数）よりも大きい大標本理論の下で，データが正規分布に従うかどうかを調べる正規性検定について研究を行ってきた．特に，積率に基づいた統計量の提案と実用性を確認してきた．研究の背景で述べたように，積率に基づく統計量は簡便ではあるが問題点も存在する．そのため，正規性検定で用いられる積率に基づく簡便な統計量と，実際に使用される統計量との近似精度や有用性について議論することで問題点を明らかにすることを目標とする．更に，積率に基づく既存の統計量における近似精度の改良などを目指す．

また，各個体について経時的に繰り返し測定されるデータのモデル化の一つである成長曲線モデルに対して，データから適切なモデルを選ぶための変数選択規準の構成や特性について研究代表者はこれまで取り組んできた．殊に大標本理論や高次元データの下での議論を行い，提案した情報量規準がどちらの仮定の下でも有用であることに言及した．しかし，成長曲線モデルにおける議論は多変量回帰とは異なる傾向を持ち，高次元データに対しても一致性を持たないことが見出された．そこで，先行研究より枠組みを拡張し，より実用的な枠組みで規準量の特性を検討することを目的とする．

３．研究の方法

正規性検定に対しては，積率に基づく簡便な統計量と特性関数に基づく Henze-Zirkler（HZ）

統計量との比較を行う．HZ 統計量は検出力が高いことで知られている．簡便な統計量である歪度と尖度を用いた統計量と比較して，さまざまな分布に対する検出力比較をし，どのような状況のもとでそれぞれの統計量が高い検出力を持つかを検討する．他には Mardia による尖度の定義を用いた統計量について，先行研究で提案されている統計量よりも近似精度の良い統計量の提案を行う．先行研究で提案されている統計量は平均と分散を用いて標準化した統計量であり標本数が小さいときには近似精度が良くないことが分かっている．提案する統計量は３次積率を用いて統計量の分布を改良することを目指す．使用する正規化変換の方法は Konishi によるものであり，数値実験では標準化した統計量と Wilson-Hilferty（WH）統計量との比較も行う．WH 統計量も３次積率を用いた統計量であり，それぞれの統計量に対して近似精度や検出力比較を行う．

成長曲線モデルに対してはグループ数の個数と標本数が共に増加するもとでの一致性について議論を行い，さまざまな規準量での特性について検討する．取り扱う規準量は AIC，BIC の他に Mallows の規準量と多くの規準量について議論を行う．それぞれの規準量のリスクを計算し，各バイアス項が大標本理論での規準量とどのような関係にあるかについても言及する．重要なことは非心行列のオーダー項について調べることと，推定量の分布などを求めることである．他には，共分散行列に自己回帰モデルや一様共分散構造の仮定を置いた場合の規準量の一致性について議論を行う．共分散に仮定を置く前までは成長曲線モデルとしては多項式を仮定していたが，新しい一致性の証明方法として KOO 法を採用することで，より多くのモデル（全ての）に対して検討が可能になる．KOO 法とは一つの説明変数を除いたモデルの規準量を考え，さらにフルモデルに対する規準量と差を取る．その差の値によって，その説明変数が不要か必要であるかを見極めることが可能となる．本来ならば，説明変数の全ての組み合わせを考え，それぞれの規準量を導出した上で，一番値の小さいものを最適なモデルとして選択するという流れがあるが，KOO 法を用いると説明変数の個数の分だけ考えれば良いこととなる．これらに関しては規準量の一致性だけでなく，それぞれの共分散のパラメータの推定量や分布についても新たに求める必要がある．さらに本研究の性質上，最終的に数値実験により導出した規準量や他の規準量の正当性・妥当性を確認する．

４．研究成果

正規性検定に関して，積率を用いた簡便な（Mardia または Srivastava による定義を用いた）統計量と HZ 統計量と比較実験を行った結果，予想では HZ 統計量が全ての分布に対して高い検出力を持つであろうとされていたが，必ずしもそのような結果とはならなかった．簡便な統計量と HZ 統計量が同程度の検出力を持つ場合や積率を用いた統計量の方が勝っていた分布も存在した．さまざまな分布の下で検出力比較を行うことで，各統計量の特徴を吟味することが可能となった．更に帰無分布の下で第一種の過誤についても調べ，それぞれの統計量の近似精度や保守的な統計量とみなせるか否かについても評価した．経験的に取り扱うデータが正規分布とみなされる場合においても，正規性検定でデータを精査することは重要であり，データの特徴が予め分かっている状態であれば適切な統計量を用いて検定することで，より信頼できる分析結果を得ることが可能となる．

また，Mardia による尖度の定義を用いた統計量については，先に述べた簡便な統計量でも標準化した統計量が使用されていたように，平均と分散で標準化した統計量と WH 統計量がよく知られている．特に標準化した統計量に関しては，標本数が小さい場合において近似精度がよくないという問題点と，分布としての近似精度の改良を目指し新たな統計量の提案を行った．３次積率を用いて正規化変換を行うことで統計量の分布関数のオーダー項の改良が可能となった．よって，標本数が小さい場合でも統計量の分布の形状は左右対称に近づき，実際に統計量の尖度と歪度が改良されていることを確認している．また，実際に検定で用いるパーセント点や検出力についても先行研究の結果と比較し，提案した統計量の有用性を検証している．正規性検定を行う際に，例えばデータに歪みがないことが事前情報として認識されている場合には，尖度と歪度を用いた簡便な統計量よりも尖度の統計量を用いて検定を行った方が検出力は高くなる．したがって，尖度のみや歪度のみの統計量もデータ次第で有用であると言える．

正規性検定に限らず，仮説検定によって検定を行う場合には，いくつかの検定方法を組み合わせることとなる．したがって，データの特性が分からない状態での正規性検定では，積率に基づく統計量を使用して検定を行い，更に別の（例えば，特性関数に基づいた）統計量と組み合わせることでより正確な結果を得ることが可能となる．よって，別の観点で統計量を構築することは重要な課題と言える．

成長曲線モデルにおける規準量については主に，共分散構造を仮定した場合と仮定しない場合についての結果を得ている．構造に仮定を置かない場合，グループ数の個数と標本数が共に増加する枠組みのもとで，AIC，BIC，Mallows の規準量と各規準量を改良した規準量についての一致性や特性について言及した．対数尤度に基づく規準量を一般の形で表し，モデルの非心行列のオーダー項についての仮定を置くことで一致性について証明を行った．また，それぞれの規準量との関係性について表し，追加で仮定を置くことで一致性をもつ状況を示した．これらの結果は数値実験でも確認をしている．更なる発展として実際のデータとして考えられる，次数や次元についても十分大きい仮定を追加した場合の議論などが考えられる．

また，高次元データの下で共分散に自己回帰モデルを仮定した場合，パラメータの漸近分布や特性について導出し，数値実験によって近似精度の確認を行っている．一様共分散構造を仮定した場合も同様に高次元データの下で，まずは適合性について議論を行った．更に，パラメータの漸近分布を導出し，一致性についての議論を行っている．以上の規準量の一致性については多項式モデルを仮定していたが，一様共分散構造の一致性については KOO 法を用いて証明を行った．KOO 法により多項式モデルに限定せずに説明変数の全ての組み合わせを考慮し一致性の証明を行うことが可能となった．数値実験ではパラメータの近似精度の確認を行っており，一致性についての数値実験は今後の課題となっている．

５．主な発表論文等

〔雑誌論文〕　計4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　2件）

１．著者名 Rie Enomoto, Zofia Hanusz, Ayako Hara and Takashi Seo	４．巻 49
２．論文標題 Multivariate normality test using normalizing transformation for Mardia ' s multivariate kurtosis	５．発行年 2020年
３．雑誌名 Communications in Statistics - Simulation and Computation	６．最初と最後の頁 684-698
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1080/03610918.2019.1661476	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 Zofia Hanusz, Rie Enomoto, Takashi Seo and Kazuyuki Koizumi	４．巻 47
２．論文標題 A Monte Carlo comparison of Jarque-Bera type tests and Henze-Zirkler test of multivariate normality	５．発行年 2018年
３．雑誌名 Communications in Statistics - Simulation and Computation	６．最初と最後の頁 1439-1452
掲載論文のDOI（デジタルオブジェクト識別子） http://dx.doi.org/10.1080/03610918.2017.1315771	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 Tetsuro Sakurai, Rie Enomoto and Yasunori Fujikoshi	４．巻 17
２．論文標題 High-Dimensional Asymptotic Distributions of Simplified MLEs in Growth Curve Model with an Autoregressive Covariance Structure	５．発行年 2017年
３．雑誌名 Technical Report, Statistical Research Grourp, Hiroshima University	６．最初と最後の頁 1-16
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 オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Rie Enomoto, Tetsuro Sakurai and Yasunori Fujikoshi	４．巻 51
２．論文標題 Consistency properties of AIC, BIC, Cp and their modifications in the growth curve model under a large-(q,n) framework	５．発行年 2015年
３．雑誌名 SUT Journal of Mathematic	６．最初と最後の頁 59-81
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3件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 榎本理恵
２．発表標題 一様共分散構造をもつ成長曲線モデルのモデル選択規準について
３．学会等名 2018年度 統計関連学会連合大会
４．発表年 2018年

１．発表者名 榎本 理恵
２．発表標題 一様共分散構造をもつ成長曲線モデルに対する次数選択について
３．学会等名 2015年度 統計関連学会連合大会
４．発表年 2015年

１．発表者名 榎本 理恵，櫻井 哲朗，藤越 康祝
２．発表標題 一様および自己回帰共分散構造をもつ成長曲線モデルに関する高次元推測
３．学会等名 2015年度 統計関連学会連合大会
４．発表年 2015年

〔図書〕　計1件

１．著者名 杉山高一，藤越康祝(監修)，榎本理恵，その他著者	４．発行年 2020年
２．出版社 勉誠出版	５．総ページ数 272
３．書名 R・Python による 統計データ科学	

〔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