科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１１　日現在

機関番号：　３２６６５
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７４００８５

研究課題名（和文）高次元データについての検定法の非正規分布に対する頑健性の研究

研究課題名（英文）A study of robustness for testing problem under non-normality for high-dimensional data

研究代表者
山田　隆行（YAMADA, Takayuki）

日本大学・工学部・助教

研究者番号：６０５１０９５６

交付決定額（研究期間全体）：（直接経費）　　1,900,000円、（間接経費）　　570,000円

研究成果の概要（和文）：高次元データの統計的検定問題を扱っている．複数の群の平均がすべて等しいかどうかを調べる検定について母集団分布に正規分布を仮定したもとで提案されているものが多変量ｔ分布に拡張した場合に使用できないことをシミュレーション実験より示した．正規分布を含むような一般化した母集団分布のもとで使用できるような検定統計量を提案した．加えて，母集団分布に正規分布を仮定してよいかどうかを調べる検定法を提案した．シミュレーション実験を通し標本サイズと次元数が大きくなるにつれ提案した検定の精度が良くなることを確かめた．

研究成果の概要（英文）：This study is concerned with the statistical testing problem for high-dimensional data. Firstly, we dealt with the problem for testing homogeneity of mean vectors, that is, testing that all mean vectors are equal. Under the condition that the population distribution is multivariate normal, testing criterion was being proposed. Through the simulation study, I confirmed that the precision of the testing criterion gets worth when the population distribution is multivariate t. I proposed other testing statistic under a generalized population distribution which contains multivariate normal. In addition, we proposed a testing criterion for testing multivariate normality for high-dimensional data. Through simulation, we validate that the precisions of the proposals become well as the sample size and the dimension are large.

研究分野：　数物系科学

科研費の分科・細目：　数学・数学一般（含確率論・統計科学）

キーワード：　多変量解析　統計的推測論　高次元データ　統計数学

１．研究開始当初の背景

近年の情報収集と保存技術の進歩に合わせて，観測項目の数 p と標本サイズ n がともに大きい高次元データの統計解析法の研究が脚光をあびている．この際，高次元枠組み: $p \to \infty$, $n \to \infty$, $p/n \to c \in (0, \infty)$ のもとでの漸近理論が解析法の構成に有力であることが指摘されている．

正規分布を仮定したもとでの研究として，分散共分散行列の検定問題については Lodoit and Wolf (2002, Ann. Statist.)；平均ベクトルの検定については Srivastava and Du (2007, J. Multivariate Anal.)；一元配置多変量分散分析については Schott (2007, J. Multivariate Anal.)；多変量線形モデルにおける多変量線形仮説検定問題については Fujikoshi et al. (2004, J. Japan Statist. Soc.), Srivastava and Fujikoshi (2006, J. Multivariate Anal.)等がある．これらの研究では高次元枠組みのもとでの検定統計量の漸近分布が導出されている．

２．研究の目的

遺伝子マイクロアレイデータや環境データのように明らかに正規性を満たさないデータが存在する．実際には正規分布に従っていないデータに正規性の仮定のもとで提案された手法を適用した場合，推定量が外れ値の影響を受けたり，検定のサイズが非常に大きくなったりと深刻な問題を引き起こす．これを解決するために検定法の補正が行われてきた．その多くは統計量のエッジワース展開によって行われる．しかし，高次元データに対してはこの補正法がうまくいかない．

非正規分布を仮定したもとでの研究としては，共分散行列の検定についての成果はまだない．平均ベクトルの 2 標本問題を扱っているものとして Srivastava (2009, J. Multivariate Anal.) と Chen and Qin (acceptad in Anals of Statistics)がある．これらの研究では誤差ベクトルの成分間に独立性を仮定し高次元枠組みのもとでの漸近分布を導出している．

本研究は母集団分布に正規分布を仮定したもとで得られた検定規準が正規分布を仮定していない場合と等しいかどうか，つまり分布に関して頑健であるかどうかを調べる．頑健でなかったものに関してはその補正を行う．

３．研究の方法

非正規分布を仮定したもとでの高次元漸近分布論の研究は平均ベクトルの 2 標本問題について前出の 2 つの研究成果がある.まず，形モデルの誤差ベクトルの成分が独立であるという仮定でのもの(Srivastava, 2009, J. Multivariate Anal.)と，誤差ベクトルの成分の 8 次までのモーメントがそれぞれの周辺のモーメントの積に分解可能という仮定の下でのもの(Chen and Qin, 2010, Anals of Statistics)である．これらの誤差ベクトルの仮定をより弱めたもとで研究を行う．

４．研究成果

(1) 共分散行列の関数についての推定とその応用：

高次元の場合の平均ベクトルの検定の検定統計量の漸近分散は、データベクトル $\boldsymbol{x}$ に関する正則変換に関して不変ではないため，漸近分散が共分散行列の 2 乗のトレース $\mathrm{tr}\,\boldsymbol{\Sigma}^2$ の関数となっている．他の高次元データの多変量解析でも統計量の漸近分散が共分散行列の関数になることがある．本研究では多変量線型モデル $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}$ の下で$(\alpha_2,\alpha_1^2,\kappa)=(\mathrm{tr}\boldsymbol{\Sigma}^2/p,(\mathrm{tr}\boldsymbol{\Sigma}/p)^2,\kappa_{11})$の不偏推定量を見出した．ここで，$\boldsymbol{\mu}$ は平均ベクトル，$\boldsymbol{z}$ を誤差ベクトル，$\kappa_{ij} = \mathrm{E}[\boldsymbol{z}'\boldsymbol{\Sigma}^i\boldsymbol{z}\,\boldsymbol{z}'\boldsymbol{\Sigma}^j\boldsymbol{z}]-2\mathrm{tr}\boldsymbol{\Sigma}^{i+j}-\mathrm{tr}\boldsymbol{\Sigma}^i\,\mathrm{tr}\boldsymbol{\Sigma}^j$, $\boldsymbol{z}$ は誤差ベクトルの確率分布に従うある変数ベクトルとする．この不偏推定量を導くにあたり，$(\mathrm{tr}\,\boldsymbol{\Sigma}^2,(\mathrm{tr}\boldsymbol{\Sigma})^2)$に関してはその共分散行列を標本共分散行列 $\boldsymbol{S}$で置き換えた自然な推定量，κ に関してはその自然な推定量として $Q = \sum_{i=1}^{N}\{(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\}^2/(N-1)$ を使った.ここで Nは標本サイズとする．これらの推定量は偏りを持ったものになってしまうため，これら 3 つの 1 次結合によって不偏推定量を導出した．$(\alpha_2,\alpha_1^2,\kappa)$の不偏推定量を$(a_2,a_1^2,\hat{\kappa})$ で表すとき，

$a_2 = c\{(N-1)(N-1)\,\mathrm{tr}\boldsymbol{S}^2+(\mathrm{tr}\boldsymbol{S})^2-NQ\}/p$,

$a_1^2 = c\{2\,\mathrm{tr}\boldsymbol{S}^2+(N^2-3N+1)\,(\mathrm{tr}\boldsymbol{S})^2-NQ\}/p$,

$\hat{\kappa} = -\{2(N-1)^2\,\mathrm{tr}\boldsymbol{S}^2+(N-1)^2(\mathrm{tr}\boldsymbol{S})^2-N(N+1)Q\}/\{(N-2)(N-3)\}$,

ここで $c = (N-1)/\{N(N-2)(N-3)\}$ とする．

次に $(a_2,a_1^2,\hat{\kappa})$ について高次元漸近枠組み：A1：$p \to \infty$, $n \to \infty$, $p/n \to c \in (0, \infty)$のもとでの一致性を調べた．一致性を調べるにあたり，共分散行列の構造に関して次を仮定する．

A2: $\alpha_i = \mathrm{tr}\boldsymbol{\Sigma}^i/p \to \alpha_{i0} \in (0,\infty)$.

a_2 については，誤差ベクトルの分布の条件 C1:$\mathrm{E}[(\boldsymbol{z}_1'\boldsymbol{\Sigma}\boldsymbol{z}_2)^4] = o(p^4)$, $\kappa_{22} = o(p^3)$ の下一致性を示した．a_1^2については，誤差ベクトルの分布の条件 C2: $\kappa_{11} = o(p^3)$ の下一致性を示した．ここで $\boldsymbol{z}_1$ と $\boldsymbol{z}_2$ は独立で誤差ベクトルの分布に従う確率変数ベクトルとする．$\hat{\kappa}$ に関しては，誤差ベクトルの成分が独立の場合とそうでない場合で一致性を担保する条件が異なる.独立の場合は $\kappa_{11} = O(p)$を仮定することにより$\hat{\kappa}/p$ が κ/p の一致推定量であることが示される．それ以外の場合は，誤差ベクトルの分布に $\kappa/p^2 = O(1)$ が仮定できる場合に，$\hat{\kappa}/p^2$ が κ/p^2 の一致推定量であることが示される．以上の一致性のための条件は正規分布の場合には自然に成り立つ．　この一致推定量の応用として，$\hat{\kappa}/p$ を使った多次元正規性の検定を提案した．Mardia(1970, Biometrika)では多変量尖度パラメータの推

定量に基づく多次元正規性の検定を提案しているが，κ は Σ が単位行列の場合に多変量尖度パラメータと一致するため，導出した推定量を使って高次元の場合の正規性の検定を提案することが可能である．帰無仮説を誤差ベクトルの分布が正規分布であるとした場合にκ/p の漸近正規性を示し，その漸近分散の一致推定量の平方根で割ったスチューデント化検定統計量に基づく検定規準を与えた．シミュレーション実験により標本サイズと次元数が大きくなるにつれ検定の精度が良くなることを確認した．また対立仮説として多変量 T 分布（自由度を固定）の場合や誤差ベクトルが独立の場合を考え，前者の場合は標本サイズと次元数が大きくなるにつれ検出力が 1 に近づいていくこと，後者の場合は次元数の増加の影響は少ないが標本サイズが大きくなるにつれ検出力が 1 に近づいていくことを確認した．Dudoit et al. (2002, JASA)で使用されている白血病患者の DNA マイクロアレイデータに対して正規性を検証したところ，急性リンパ芽球性白血病患者のものに対しては p-値が 0.663, 急性骨髄性白血病患者のものに対しては p-値が 0.838 となった．

(2) 多群の平均の検定に関しての修正：
高次元データに関して 2 つ以上の複数個の群に関して平均ベクトルが等しいかどうかの検定を考えた．母集団分布に正規分布を仮定した結果としては，たとえば Fuikoshi et al.(2004, JJSS)がある．この問題を正規分布を含むような一般の分布の場合に考える．
加えて，各群に共通の共分散行列を仮定せずに考えている．g 個の群に対する帰無仮説として H0: $\mu_1 = \dots = \mu_g$ とおく．帰無仮説からの乖離を $m = \Sigma_{i=1}^{g} N_i (\bar{x}_i - \bar{x})' (\bar{x}_i - \bar{x})$, $\bar{x} = (1/N) \Sigma_{i=1}^{g} N_i \bar{x}_i$, $N = N_1 + \dots + N_g$ である．1 元配置多変量分散分析における群間平方和積和行列 B と Σ_i の不偏推定量 S_i を使って乖離度の不偏推定量 T を次のように定義する:
$T = \mathrm{tr} B - \Sigma_{i=1}^{g} (1 - N_i / N) \mathrm{tr} S_i$.
この漸近帰無分布を次の高次元漸近枠組み A1 の下で与えた．
A1: $p \to \infty$, $N_i \to \infty$, $N_i/p \to c_i \in (0, \infty)$, $N_i/N \to \gamma_i \in (0, 1)$, i = 1,...,g.
観測値ベクトルに多変量線型モデル
$x_{ij} = \mu_i + \Sigma_i^{1/2} z_{ij}$, j = 1,...,$N_i$, i = 1,...,g.
を仮定して導出を行う．誤差ベクトルの確率分布 F として共分散行列の仮定 A2,A3 の下で成り立つ条件 A4～A6 を充たす多次元分布を仮定する．
A2: $\mathrm{tr}\Sigma_i^2 / p = O(1)$, i = 1,...,g, ただし少なくても一つは正の定数に収束する．
A3: $\mathrm{tr}\Sigma_i^4 / p = O(1)$, i = 1,...,g.
A4: $\kappa_1 = \sup_{1 \le i \le p} E[(z' \Sigma_i^2 z - \mathrm{tr}\Sigma_i^2)^2] = O(p^2)$, ただし z は F に従う確率変数ベクトル．
A5: $\kappa_2 = \sup_{1 \le i,j \le p} E[(z_1' \Sigma_i^{1/2} \Sigma_j^{1/2} z_2)^4] = o(p^4)$, ただし z_1 と z_2 は独立で F に従う確率変数ベクトル．
A6: $\kappa_3 = \sup_{1 \le i,j \le p} E[(z' \Sigma_i^{1/2} \Sigma_j \Sigma_i^{1/2} z)^4] = O(p^2)$, ただし z は F に従う確率変数ベクトル．
以上の仮定の下で統計量 T の漸近正規性を示した．統計量 T はデータベクトル x に関する正則変換に関して不変ではないため，漸近分散が共分散行列の 2 乗のトレース $\mathrm{tr}\Sigma_i^2$ や $\mathrm{tr}\Sigma_i\Sigma_j$ (i,j = 1,...,g, i≠j) の関数となっている．(1)の研究成果を使うことにより $\mathrm{tr}\Sigma_i^2$ の不偏推定量を与えることができる．$\mathrm{tr}\Sigma_i\Sigma_j$ の不偏推定量を標本共分散行列で置き換えた $\mathrm{tr}S_iS_j$ を使う．推定量の一致性を示すために次の仮定 A7 を用意した．
A7: $\kappa_{22} = \sup_{1 \le i \le p} \{ E[(z_1' \Sigma_i^2 z_1)^2] - 2\mathrm{tr}\Sigma_i^4 - (\mathrm{tr}\Sigma_i^2)^2 \} = o(p^3)$, $\sup_{1 \le i \le p} E[(z_1' \Sigma_i z_2)^4] = o(p^4)$, ただし z_1 と z_2 は独立で F に従う確率変数ベクトル．
仮定 A1～A3,A5,A7 の下で $\mathrm{tr}\Sigma_i^2$ の不偏推定量の一致性は示される．さらに仮定 A1～A3,A5, A6 の下で $\mathrm{tr}\Sigma_i\Sigma_j$ の不偏推定量の一致性が示される．以上から T のスチューデント化統計量の漸近正規性が仮定 A1～A7 の下で成り立つことが分かった．
シミュレーション実験で今回の検定統計量が頑健であることを確かめるために，母集団分布として多変量正規分布，多変量 T 分布，誤差ベクトルの成分が独立同一分布に従う場合を扱った．提案した検定規準と，正規分布を仮定した場合に提案されている検定規準(Fuikoshi et al.(2004, JJSS)の結果)を比較した．2 群の場合を扱い，共通の共分散行列を仮定し，複数回発生させた標本から計算した実際の過誤確率を調べたところ，母集団分布が多変量 T 分布の場合に正規分布を仮定した場合に提案された検定規準は p の値が大きくなるにつれ設定した第一種過誤確率の値よりはるかに小さくなり，過大評価を与えるようになっていくことがわかった．いっぽう提案規準は，設定した第一種過誤確率の値より実際の過誤確率は大きくなってしまうが，p の値が大きくなるにつれ設定した第一種過誤確率の値に近づいていくことがわかった．母集団分布が正規分布の場合や誤差ベクトルの成分が独立同一分布に従う場合は，提案規準と正規分布の下で得られた規準 2 つの精度の違いは p の値が大きくなるにつれなくなっていくことがわかった．

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 1 件）

Himeno, T. and Yamada, T., Estimations for some functions of covariance matrix in high dimension under non-normality and

its applications, *Journal of Multivariate Analysis*, **査読有**, To appear, http://dx.doi.org/10.1016/j.jmva.2014.04.020

〔学会発表〕（計 4 件）

① **山田隆行**, **姫野哲人**, Testing homogeneity of mean vectors under heteroscedasticity in high-dimension. 2013 **年度統計関連学会　連合大会**.

② Yamada, T. and Himeno, T., Test for assessing multivariate normality available for high-dimensional data. Special Session on Perspectives on High Dimensional Data Analysis(organized by Srivastava,M.S.), 22nd International Workshop on Matrices and Statistics, August 12-15, 2013, Toronto, Canada.

③ Himeno, T. and Yamada T., Estimations for some functions of covariance matrix in high dimension under non-normality and its application. 2012 **年度統計関連学会　連合大会**.

④ Yamada, T. and Himeno, T. Test for mean vector in high-dimension under non-normality. Poster presentation in 8th World Congress in Probability and Statistics, July 9-14, 2012, Istanbul, Turkey.

６．研究組織

(1)研究代表者

山田　隆行（YAMADA, Takayuki）
日本大学・工学部・助教
研究者番号： 60510956