成蹊大学理工学研究報告
J. Fac. Sci.Tech., Seikei Univ.
Vol.50 No.2 (2013)pp.79-82

多変量線形モデルにおける高次元漸近理論

姫野　哲人*1

High-dimensional asymptotic theory for multivariate linear model

Tetsuto HIMENO*1

ABSTRACT：When a statistic with a complicated distribution is dealt, the asymptotic distribution is often used. Even if the exact distribution is complicated, the asymptotic distribution generally becomes simple form such as normal distribution and chi square distribution. There are also previous studies which derive the asymptotic correction due to improve the approximation. However it is empirically known that the classical asymptotic approximations become worse as the dimension becomes large. So we derive some high-dimensional asymptotic results for the multivariate linear model. These results have better approximations in spite of the size of dimension. These results not only derive better approximation but also clarify asymptotic properties of some test statistics.

Keywords：high-dimension, asymptotic theory, multivariate linear model

(Received September 20, 2013)

1．はじめに

統計学の分野では，扱う統計量の分布が複雑である場合や，その分布が未知である場合は数多くあり，このような場合，一般的に（パラメトリックな手法で）データを分析することは困難である。しかし，このような場合であっても，その近似分布を得ることができれば，様々な統計的分析が可能となる。代表的な近似手法としては，中心極限定理や最尤推定量の漸近正規性，尤度比統計量や適合度検定のカイ二乗近似などがよく知られている。これらの近似を使うことによって，これまでに数多くの検定手法が提案されている。しかし，近年のコンピューターの発達により，我々が扱うデータの中にはマイクロアレイデータや画像データ，時系列データなどの高次元データも増えてきており，このような高次元データに対し，従来の古典的な漸近理論はうまく適用できないことが知られている。これは，古典的な漸近理論では，サンプルサイズ N とパラメータ数 p に対し，p/N のような項は 0 に収束する項として扱われるが，高次元データではこの比率が無視できない程度の大きさになることが原因である。そこで，本報告では多変量線形モデルにおけるパラメータの線形仮説に対する検定統計量としてよく知られている尤度比検定統計量，Lawley-Hotellingトレース規準，Bartlett-Nanda-Pillaiトレース規準，Dempsterトレース規準に対し，高次元漸近理論を適用し，漸近分布を導出する。また，これらの漸近分布を用い，検出力（帰無仮説が正しくない場合に帰無仮説を棄却する確率）の漸近比較を行うことにより，様々な状況下での最適な検定統計量を選ぶ規準を与える。

*：情報科学科助教 (t-himeno@st.seikei.ac.jp)

2．多変量線形モデル

まず，多変量線形モデルについて説明する。p 次元の観測データが $\boldsymbol{y}_1,\cdots,\boldsymbol{y}_N$ のように N 個あるとする（ここで，$\boldsymbol{y}_i=(y_{i1},\cdots,y_{iN})'$ とし，「'」は転置を表す記号とする）。このとき，これらのデータが以下の線形モデル

$$\boldsymbol{y}_i=\Theta'\boldsymbol{x}_i+\boldsymbol{\varepsilon}_i\ \ (i=1,\cdots,N)$$

に従うとする。ここで，Θ は $k\times p$ の未知のパラメータ行列，$\boldsymbol{x}_i$ は k 次元の既知の説明変数ベクトル，$\boldsymbol{\varepsilon}_i$ は p 次元の誤差ベクトルであり，それぞれ独立に正規分布 $N(\boldsymbol{0}_p,\Sigma)$ に従うとする（ $\boldsymbol{0}_p$ は成分が全て 0 である p 次

元ベクトルとし，Σ は $p\times p$ の正定値行列とする)。このモデルは，$Y=(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_N)'$，$X=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)'$, $E=(\boldsymbol{\varepsilon}_1,\cdots,\boldsymbol{\varepsilon}_N)'$ と置くことで，

$$Y=X\Theta+E$$

と表すことができる (rank(X)=k と仮定する)。ここでパラメータに対する線形仮説

$$H_0:C\Theta=O$$

を考える (rank(C)=q とする)。ここで，このモデルと仮説がどのような状況を表せるのか考えてみる。例えば $N=N_1+\cdots+N_{q+1}$ のようにデータを $q+1$ 個の群に分け，$\Theta=(\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_{q+1})'$とし，$j$ 番目の群に属する i に対し，$\boldsymbol{x}_i$ を第 j 成分が 1 で他は 0 となるベクトルとし，

$$C=\begin{pmatrix}1 & & O & -1\\ & \ddots & & \vdots\\ O & & 1 & -1\end{pmatrix}$$

とする。すると，このモデルは，第 j 群に属するデータに対し，

$$\boldsymbol{y}_i=\boldsymbol{\mu}_j+\boldsymbol{\varepsilon}_i$$

と表すことができ，仮説は

$$\boldsymbol{\mu}_1=\cdots=\boldsymbol{\mu}_{q+1}$$

という多標本モデルに対する同質性検定を表すことができる。つまり，多変量線形モデルとその線形仮説は様々な線形モデルや仮説を含んだモデルであるといえる。

3．検定統計量

多変量線形モデルとその線形仮説に対し，

$$S_h=(C\hat{\Theta})'[C(X'X)^{-1}C']^{-1}C\hat{\Theta}$$
$$S_e=(Y-X\hat{\Theta})'(Y-A\hat{\Theta})$$
$$\hat{\Theta}=(X'X)^{-1}X'Y$$

とおく。このとき，尤度比検定統計量，Lawley-Hotelling トレース規準，Bartlett-Nanda-Pillai トレース規準はそれぞれ

$$-\log(|S_e|/|S_e+S_h|)$$
$$\mathrm{tr}S_hS_e^{-1}$$
$$\mathrm{tr}S_h(S_e+S_h)^{-1}$$

として定義される (Muirhead, 1982)。ここで，tr は行列のトレースを表す。これらの漸近分布の導出にあたり，

$$T_{\mathrm{LR}}=-\sqrt{p}\left(\frac{N-k+q}{p}\right)\times\left\{\log\frac{|S_e|}{|S_e+S_h|}+q\log\frac{N-k+q}{p}\right\}$$
$$T_{\mathrm{LH}}=\sqrt{p}\left(\frac{N-k-p+q}{p}\mathrm{tr}S_hS_e^{-1}-q\right)$$
$$T_{\mathrm{BNP}}=\sqrt{p}\left(\frac{N-k+q}{p}\right)\times\left(\frac{N-k+q}{p}\mathrm{tr}S_h(S_e+S_h)^{-1}-q\right)$$

のように規準化して考える。これらの統計量を定義する場合，S_e が正則である必要があるため，漸近分布を考えるための高次元枠組みとして，

$$(C1)\ N,p\to\infty,\ p/N\to c\in(0,1)$$

という条件が必要となる。つまり，p が N を超えてはいけない。そこで，p が N を超えても定義できる方法として提案されたものがDempsterトレース規準である。これは最初Dempster (1958, 1960) によって，一標本問題と二標本問題の場合に定義された。これを多変量線形モデルの場合に拡張したものは

$$\mathrm{tr}S_h/\mathrm{tr}S_e$$

として定義される。この統計量も規準化し，

$$T_{\mathrm{D}}=\sqrt{p}\{(\mathrm{N}-\mathrm{k})\mathrm{tr}S_h/\mathrm{tr}S_e-q\}$$

として漸近分布を考える。この検定統計量は N と p の大小に関係なく定義できるので，

$$(C2)\ N,p\to\infty,\ p/N\to c\in(0,\infty)$$

という高次元枠組みで扱うことが可能である。また，漸近分布を導出するために，

$$(A1)\ \mathrm{tr}\Sigma^i/p=O(1)\ (i=1,\cdots,4)$$

を仮定する。また，対立仮説

$$H_1:C\Theta\neq O$$

の下での漸近分布を考える場合は，非心行列を

$$\Omega=\Sigma^{-1/2}(C\Theta)'(C(X'X)^{-1}C')^{-1}C\Theta\Sigma^{-1/2}$$

とし，

$$(A2)\ \mathrm{tr}\Sigma^i\Omega/p=O(1)\ (i=1,2)$$

も仮定する。これらの仮定の下で，漸近分布の導出を行う。

4．T_{LR}，T_{LH}，T_{BNP}の比較

漸近分布を導出する方法は様々存在するが，本報告で使った手法は，検定統計量の特性関数の漸近展開を求め，特性関数の反転公式を用いることで，分布関数の漸近展開を得る。その結果，T_G (G = LR, LH, BNP) の帰無仮説の下での漸近分布は

$$P(T_G/\sigma \le z_{CF}(\alpha)) = 1-\alpha+O(p^{-3/2})$$
$$z_{CF}(\alpha) = z_\alpha + p^{-1/2}b_1(z_\alpha,G) + p^{-1}b_2(z_\alpha,G)$$

として得られる。ここで，P は確率を表し，z_α は標準正規分布の上側 100α %点を表す。係数 $b_1(z_\alpha,G)$ と $b_2(z_\alpha,G)$ に関しては，Himeno (2007a) を参照のこと。ここで，σ は G に依存しないので，これら 3 つの検定統計量の極限分布は等しく，これらの違いは $p^{-1/2}$ のオーダーでしかないことに注意する。また，仮定 (A2) と対立仮説の下での漸近分布も同様に導出することができ，その漸近分布を用いることにより，T_{LR}, T_{LH}, T_{BNP} の漸近的な検出力の差が

$$\frac{c_1}{\sqrt{p}}\varphi\left(z_\alpha - \frac{1}{\sigma\sqrt{p}}\mathrm{tr}\Omega\right)$$
$$\times\left\{\frac{1}{\sigma p}\mathrm{tr}\Omega^2 + \frac{1}{\sigma pq}(\mathrm{tr}\Omega)^2 - \frac{2z_\alpha}{q\sqrt{p}}\mathrm{tr}\Omega\right\}$$

として得られる。ここで，φ は標準正規分布の密度関数とし，c_1 は T_{LR} のときは $-p/(2(N-k+q))$，T_{LH} のときは 0，T_{BNP} のときは $-p/(N-k+q)$ という値をとるものとする。つまり，T_{LR} の検出力は常に他の二つの間の値となり，T_{LH} と T_{BNP} の検出力は上記の式の 2 行目の符号によって決まることが分かる。

5．T_Dとその他の検定の比較

T_D の帰無仮説の下での漸近分布はHimeno (2007b) で述べられているが，他の検定統計量との検出力の比較を行う際には極限分布の結果のみで十分なので，ここでは極限分布の結果についてのみ触れることにする。

条件 (C2) と仮説 (A1) が成り立つとする。このとき，帰無仮説の下で

$$T_D/\sigma_D \xrightarrow{d} N(0,1)$$

が成り立つ。ここで，$\xrightarrow{d}$ は分布収束を示し，

$$\sigma_D = \frac{\sqrt{2q\mathrm{tr}\Sigma^2/p}}{\mathrm{tr}\Sigma/p}$$

であるとする。また，他の検定統計量の場合と同様に仮定 (A2) と対立仮説の下での極限分布を導出することが可能であり，これらの結果を使い，T_D を用いた際の検出力の極限 P_D は

$$P_D = \Phi\left(\frac{\mathrm{tr}\Sigma\Omega}{\sqrt{2q\mathrm{tr}\Sigma^2}} - z_\alpha\right)$$

として得られる。ここで，Φ は標準正規分布の分布関数とする。一方，条件 (C1) の下，T_{LR}, T_{LH}, T_{BNP} の検出力の極限 P_G は

$$P_G = \Phi\left(\frac{\mathrm{tr}\Omega}{\sqrt{2q(N-k+q)}} - z_\alpha\right)$$

となる。この二つの検出力の極限を条件 (C1) の下で比較すると，以下のことが分かる (Fujikoshi et al., 2004)。

(1) Σ が単位行列の定数倍であれば，$P_G<P_D$となる。
(2) Σ の最大固有値と最小固有値の差が小さければ，$P_G<P_D$ となる。
(3) Σ の最大固有値と最小固有値の差が大きく，p が小さければ，$P_G>P_D$ となる。

6．まとめ

本報告では，多変量線形モデルにおける線形仮説に対し，複数の検定統計量の高次元漸近理論に基づく漸近分布の結果を紹介した。また，漸近分布の結果を用いた検出力の比較を行い，状況に応じた最適な検定統計量の選択法を提案した。複雑な分布の漸近分布を導出することは，分布そのものの近似を得るだけでなく，その統計量自身の性質を調べるためにも有効な手段となる。

また，今回の漸近理論では，データが正規分布に従うことを仮定した。しかし，一般的にデータが正規分布に従っているとは限らず，高次元データが正規分布に従うかどうかを調べることは困難である。したがって，データが正規分布に従うという仮定無しでの漸近理論もいろいろ提案されている。だが，これらの手法の多くはかなり強い仮定が必要であるため，現在かなり緩い条件（楕円分布を含むクラス）の下での高次元漸近理論の研究を行っている。

参考文献

[1] Dempster, A. P. (1958). A high dimensional two sample significance test, *Ann. Math. Statist.*, **29**, 995-1010.

[2] Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples, *Biometrics*, **16**, 41-50.

[3] Fujikoshi, Y., Himeno T., and Wakaki, H. (2004). Asymptotic results of a high dimensional MANOVA test and power comparison when the dimension is large compared to the sample size, *J. Japan Statist. Soc.*, **34**, 19-26.

[4] Himeno, T. (2007a). A discriminant condition for the test of greatest power in the MANOVA model when the dimension is large compared to the sample size, *International Journal of Pure and Applied Mathematics*, **40**, 89-102.

[5] Himeno, T. (2007b). Asymptotic expansions of the null distributions for the Dempster trace criterion, *Hiroshima Math. J.*, **37**, 431-454.

[6] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, New York.