中央大学理工学研究所論文集　第 13 号　2007年（2008. 3. 31）
Journal of the Institute of Science and Englineering. Chuo University

高次元における独立性の検定と頑健性

藤越康祝 *，青木　誠 **，櫻井哲朗 *，杉山高一 *

Tests for Independence in High-dimension and Their Robustness

Yasunori Fujikoshi*, Makoto Aoki**,
Tetsuro Sakurai*, Takakazu Sugiyama*

abstract

This paper examines tests for independence of p variables which were proposed under normality, focussing on a high-dimensional case. The tests considered are based on (1) likelihood ratio test with large sample approximation, (2) the sum of squarred correlations with large sample approximation, (3) the sum of squarred correlations with high-dimensional approximation, and (4) the sum of squarred covariances with high-dimensional approximation. First, by numerical experiments we point some tendency on whether the actual rejection probabilities of the tests are near to the nominal rejection probabilities. In particular, the actual rejection probabilities of the test (2) are almost near to the nominal rejection probabilites for large sample situations as well as high-dimensional situations. Next we examine whether such property of the test (2) holds for discrete data. It is shown that the test (2) is fairly robust, though the result depends on the type of discrete distributions. Finally we extend the test (2) to one for independence of the variables after the effects of the other variables are removed.

1 はじめに

最近,マイクロアレイデータやファイナンスデータ,画像データなどにおいて, 変数の数 p が標本数 n より大きい高次元データの分析に関心が寄せられている．例えば，Ledoit and Wolf [4], Fujikoshi [2], Schott [5], Srivastava [7], Srivastava and Fujikoshi [8], Srivastava [9] などにおいて, 高次元の場合の多変量推測問題が取り扱われている．

この論文では，高次元の場合に焦点を当てながら p 個の変数が互いに独立であるという (完全) 独立性検定問題を扱う．この問題に対して，正規性を仮定したもとで種々の検定法が提案されている．p が n より小さければ尤度比統計量 $w_{n,p}$ が提案されている．この統計量の仮説のもとでの分布は，次元数を固定し，標本数を限りなく大きくする大標本漸近理論のもとで，自由度 $m=p(p-1)/2$ の χ^2 分布によって近似できることが知られている．この他の統計量として，相関係数の 2 乗和に基づく Hsu [3] の検定統計量

* 中央大学理工学研究科，** イーピーエス株式会社

$v_{n,p}$ がある．この統計量は，p が n よりも大きくても利用できる．最近，Schott [5] は，標本数と次元数がともに限りなく大きくなるという高次元漸近理論のもとで，検定統計量 $v_{n,p}$ の仮説のもとでの分布が正規分布に近づくことを示している．高次元の場合に利用できる検定統計量として，Srivastava [7] は共分散の平方和に基づく検定法を提案している．この検定統計量の分布は，標本数と次元数をそれぞれ別個に限りなく大きくしたもとでの高次元近似法に基づいている．

上記の 4 つの検定法；(1) 尤度比統計量（大標本近似），(2) 相関係数平方和（大標本近似），(3) 相関係数平方和（高次元近似），(4) 共分散平方和（高次元近似）と表す．この論文では，まず，これらの検定法について有意水準を 5%と定めたとき，実際の有意水準がどのようになっているかをシミュレーションによって検証する．高次元の状況を含めた種々の p,n に対して，実際の有意水準が示される．とくに，検定法 (3) については，高次元のみならず大標本の状況においても妥当な有意水準になっていることが示される．

実際のデータ分析においては，マイクロアレイデータやファイナンスデータ，画像データなどを離散データとして取扱う場合も多い．このため，上記の検定法 (3) が離散データの場合にも適用できるかどうか，すなわち，非正規性に対して頑健（ロバスト）であるかどうかを検証することは重要である．そこで，本論文においては，種々の離散分布を想定してシミュレーションによって実際の棄却確率を調べた．その結果，想定される離散分布に依存するが，ある程度満足のいく棄却確率が得られることを指摘する．

最後に，p 個の変数に加え，これに関連する q 個の変数があるとき，最初の p 個の変数から，残りの変数の影響を除いたときの独立性検定問題を考える．ここでは，この検定問題に対して，検定法 (3) を拡張できることを示す．

2　独立性の検定と有意水準近似の精度

p 個の変数を $x_1,\ldots,x_p$ 考え，これらを確率ベクトル $\boldsymbol{x}=(x_1,\ldots,x_p)$ と表す．このとき，$\boldsymbol{x}$ は (母) 平均ベクトル $\boldsymbol{\mu}$，(母) 共分散行列 $\Sigma=(\sigma_{ij})$ をもつ p 次元多変量正規分布に従うと仮定する．$\boldsymbol{x}$ についての大きさ $N=n+1$ (実際の標本数は $n+1$ であるが，この論文では n を標本数とよんでいる) の無作為標本 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N$ が与えられ，これらの標本に基づく標本平均ベクトルと標本共分散行列をそれぞれ，$\bar{\boldsymbol{x}}$，S，すなわち

$$\bar{\boldsymbol{x}}=\frac{1}{N}\sum_{\alpha=1}^{N}\boldsymbol{x}_\alpha,\quad S\equiv\frac{1}{n}V=\frac{1}{n}\sum_{\alpha=1}^{N}(\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha-\bar{\boldsymbol{x}})'$$

とする．標本共分散行列 S の (i,j) 成分を s_{ij} で表すと，x_i と x_j の標本相関係数は

$$r_{ij}=\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

として与えられる．また，x_i との母相関係数を p_{ij} とすると，母相関行列 P と標本相関行列 R はそれぞれ，$P=(p_{ij})$，$R=(r_{ij})$ として定義される．

p 個の変数 $x_1,\ldots,x_p$ が互いに独立であるという仮説は，これらの変数が多変量正規分布をするという仮定のもとで，母相関行列が単位行列である，すなわち，$H_{01}: P=I_p$ と表せる．このとき，尤度比検定統計量は

$$w_{n,p}=\left\{n-\frac{1}{6}(2p+5)\right\}\log|R|$$

で与えられ，これは仮説のもとで自由度 $p(p-1)/2$ の χ^2 分布に収束することが知られている．この近似は次元数を固定し標本数を大きくしたもとでの近似である．また，明らかに，この検定統計量では次元数が標本数を超える場合，すなわち，$p>n$ の場合，$|R|=0$ となるため利用できない．これを検定法 (1) とする．

次に，相関係数の 2 乗和に基づく検定統計量を考える．Hsu [3] は検定統計量

$$v_{n,p}=\frac{n\sum_{i<j}^{p}r_{ij}^2-\frac{1}{2}p(p-1)}{\sqrt{p(p-1)}}$$

について，これが仮説のもとで平均 0，分散 1 の正規分布に収束することを用いた検定法を提案している．この近似は次元数を固定し標本数を大きくしたもとでの近似である．この検定は p が n より大きくても定義されることを注意したい．これを検定法(2)とする．Schott [5] は，高次元の場合を考慮して検定統計量

$$t_{n,p}=\frac{\sum_{i=2}^{p}\sum_{j=1}^{i-1}r_{ij}^2-\frac{1}{2n}p(p-1)}{\sigma_{t_{n,p}}} \tag{2.1}$$

が仮説のもとで平均 0，分散 1 の正規分布で近似できるとした検定法を提案している．ここに，

$$\sigma_{t_{n,p}}^2=\frac{p(p-1)(p-1)}{n^2(n+2)} \tag{2.2}$$

この検定法は，$\lim(p/n)=\gamma_1\in(0,\infty)$ という高次元漸近的枠組のもとで，$t_{n,p}$ が平均 0，分散 1 の正規分布に収束することを示すことによって提案されたものである．これを検定法(3)とする．

p 次元確率ベクトルが多変量正規分布に従う場合には，仮説 H_{01} は共分散行列 Σ が対角行列であるという仮説 H_{02}: $\sigma_{ij}=0$ と同値である．したがって，共分散の 2 乗和に基づく検定が考えられる．Srivastava [7] は，高次元の場合にも利用できる検定統計量として

$$u_{n,p}=\frac{n(c/b-1)}{2\sqrt{1-a/\left(pb^2\right)}}$$

を提案している．ここで

$$a=\frac{1}{p}\sum_{i=1}^{p}s_{ii}^4,\ b=\frac{n}{p(n+2)}\sum_{i=1}^{p}s_{ii}^2,$$

$$c=\frac{n^2}{p(n-1)(n+2)}\left[\mathrm{tr}S^2-\frac{1}{n}(\mathrm{tr}S)^2\right]$$

これは，検定統計量 $u_{n,p}$ が仮説 H_{02} のもとで平均 0，分散 1 の正規分布で近似できるとした検定である．この漸近正規性は，n と p，および，$a_i=(\mathrm{tr}\Sigma^i)/m, i=1,\cdots,8$ について

(A)：$n=O(p^{\delta})$，$0<\delta\le 12$,
(B)：$p\to\infty$ のとき，$a_i\to a_{i,\mathrm{o}}$，$0<a_{i,\mathrm{o}}<\infty, i=1,\cdots,8$

を仮定して示されたものである．なお，この正規近似は,次元と標本数を同時ではなく，別個に限りなく大きくしたもとでの近似である．これを検定法(4)とする．

上記の 4 つの検定法の有意水準はいずれも漸近近似によるものであって，ここでは，これらの近似の良さをシミュレーションを用いて調べる．なお，検定法(3)については，近似の精度は調べられているが([5])，他の検定法の比較のためここでも検討している．シミュレーションでは，各検定方式に基づく帰無仮説のもとでの棄却確率，すなわち，有意水準 α を 0.05 として，実際の棄却確率を調べた．n と p は 4, 8, 16, 32, 64, 128 の場合を調べた．また，このシミュレーションでは，$\boldsymbol{x}$ の分布について正規性を仮定している．

まず，検定法(1)における $w_{n,p}$ のシミュレーション結果が表 1 である．$p>n$ の場合には，検定統計量が定義されないので実際の有意水準は空欄になっている．この検定では，n が p の 8 倍以上ないと近似が悪いことがわかる．

次に，検定法(2)における $v_{n,p}$ のシミュレーション結果が表 2 である．表より，$p>n$ の場合と p が n

に近いときに近似が悪くなっていることがわかる．

検定法(3)においては，検定統計量は $u_{n,p}$ であるが，高次元漸近的枠組みのもとでの近似を利用している．統計量 $t_{n,p}$ のシミュレーション結果が表3である．全ての場合において，目標の有意水準 α=0.05 に等しく，良い結果を与えている．

最後に，検定法(4)における $u_{n,p}$ のシミュレーション結果を表4に与えている．表より $p>n$ のときと p が n に近いときに近似が悪くなっていることがわかる．これは高次元の場合に対して利用できるものとして提案されたものであるが，$t_{n,p}$ の場合と比べそれほど近似がよくないのは，標本数と次元数を別々に無限大に近づけているためであると思われる．

表1: $w_{n,p}$ の実際の有意水準(α =0.05の場合).

$p\backslash n$	4	8	16	32	64	128
4	0.4552	0.1257	0.0738	0.0596	0.0542	0.0519
8		0.4537	0.1523	0.0807	0.0626	0.0558
16			0.4512	0.1446	0.0745	0.0598
32				0.4444	0.2143	0.0867
64					0.4399	0.2928
128						0.4370

表2: $v_{n,p}$ の実際の有意水準(α =0.05の場合).

$p\backslash n$	4	8	16	32	64	128
4	0.1618	0.1509	0.0958	0.0799	0.0744	0.0718
8	0.4025	0.1663	0.1011	0.0795	0.0706	0.0669
16	0.7730	0.2736	0.1367	0.0910	0.0732	0.0649
32	0.9996	0.6303	0.2559	0.1316	0.0878	0.0693
64	1.0000	0.9981	0.6219	0.2569	0.1289	0.0846
128	1.0000	1.0000	0.9944	0.6287	0.2588	0.1280

表3: $t_{n,p}$ の実際の有意水準(α =0.05の場合).

$p\backslash n$	4	8	16	32	64	128
4	0.0465	0.0431	0.0438	0.0449	0.0458	0.0462
8	0.0452	0.0460	0.0469	0.0464	0.0462	0.0462
16	0.0466	0.0459	0.0485	0.0489	0.0493	0.0488
32	0.0474	0.0461	0.0484	0.0496	0.0499	0.0497
64	0.0472	0.0460	0.0492	0.0493	0.0503	0.0501
128	0.0472	0.0459	0.0487	0.0499	0.0502	0.0505

表4: $u_{n,p}$の実際の有意水準(α =0.05の場合).

$p\backslash n$	4	8	16	32	64	128
4	0.0674	0.0504	0.0466	0.0464	0.0461	0.0466
8	0.1783	0.0736	0.0528	0.0481	0.0468	0.0468
16	0.7491	0.1709	0.0768	0.0555	0.0502	0.0490
32	1.0000	0.5899	0.1702	0.0783	0.0564	0.0514
64	1.0000	1.0000	0.5458	0.1699	0.0783	0.0566
128	1.0000	1.0000	0.9958	0.5290	0.1705	0.0788

3 検定統計量 $t_{n,p}$ の頑健性

検定法(3)の有意水準は，検定統計量 $t_{n,p}$ の高次元漸近的枠組での正規近似に基づくものであるが，第2節で見てきたように大標本や高次元のみならずすべての場合において，目標としている有意水準にほぼ等しくなっている．このような性質は，正規性の仮定のもとで示されたのであるが，ここでは離散データの場合にも成り立つのかどうかをシミュレーションによって調べた．

離散データとしては，まず一様な2値データ，3値データ，4値データ，5値データの場合を扱った．一様分布から生成した2値データによる，各変数は互いに独立であるとした $t_{n,p}$ についてのシミュレーション結果が表5である．また同様に3値データによるシミュレーション結果が表6，4値データによるシミュレーション結果が表7，5値データによるシミュレーション結果が表8である．これらは $p \gg n$ のときを除き，いずれもよい近似を与えている．

また，2値データの場合において，0と1の出現確率を変えた場合についても考察した．まず，1が出る確率が10%のときのシミュレーション結果が表9である．また同様に1が出る確率が20%のときのシミュレーション結果が表10，1が出る確率が30%のときのシミュレーション結果が表11，1が出る確率が40%のときのシミュレーション結果が表12である．

これらの表から，離散データの場合において，1の出現確率が0.5に近づくほど近似がよくなることがわかった．また，表5～表8の結果を総合すれば，それぞれの値の出る確率が一様である場合がもっとも近似がよいと思われる．

表5: $t_{n,p}$の実際の有意水準–2値データ(α =0.05の場合).

$p\backslash n$	4	8	16	32	64	128
4	0.1227	0.0487	0.0499	0.0490	0.0474	0.0474
8	0.3842	0.0599	0.0552	0.0520	0.0492	0.0478
16	0.6888	0.0749	0.0580	0.0555	0.0522	0.0505
32	0.9284	0.1058	0.0585	0.0561	0.0526	0.0513
64	0.9974	0.1740	0.0596	0.0561	0.0532	0.0514
128	1.0000	0.3191	0.0604	0.0562	0.0533	0.0523

表6: $t_{n,p}$の実際の有意水準−3値データ(α=0.05の場合).

$p\backslash n$	4	8	16	32	64	128
4	0.0716	0.0457	0.0471	0.0466	0.0465	0.0469
8	0.1520	0.0497	0.0508	0.0489	0.0479	0.0469
16	0.2866	0.0514	0.0532	0.0519	0.0507	0.0499
32	0.5157	0.0545	0.0537	0.0532	0.0515	0.0506
64	0.7989	0.0597	0.0541	0.0531	0.0519	0.0512
128	0.9723	0.0710	0.0538	0.0532	0.0514	0.0510

表7: $t_{n,p}$の実際の有意水準−4値データ(α=0.05の場合).

$p\backslash n$	4	8	16	32	64	128
4	0.0556	0.0450	0.0462	0.0465	0.0469	0.0465
8	0.0832	0.0482	0.0499	0.0486	0.0476	0.0468
16	0.1347	0.0491	0.0518	0.0514	0.0502	0.0495
32	0.2338	0.0494	0.0521	0.0527	0.0517	0.0504
64	0.4152	0.0498	0.0526	0.0525	0.0511	0.0509
128	0.6815	0.0507	0.0525	0.0523	0.0512	0.0506

表8: $t_{n,p}$の実際の有意水準−5値データ(α=0.05の場合).

$p\backslash n$	4	8	16	32	64	128
4	0.0515	0.0449	0.0452	0.0461	0.0466	0.0464
8	0.0643	0.0483	0.0495	0.0484	0.0478	0.0468
16	0.0909	0.0487	0.0512	0.0511	0.0506	0.0495
32	0.1420	0.0488	0.0521	0.0516	0.0511	0.0505
64	0.2417	0.0487	0.0523	0.0518	0.0510	0.0510
128	0.4220	0.0493	0.0523	0.0520	0.0514	0.0505

表9: $t_{n,p}$の実際の有意水準−出現確率10%2値データ(α=0.05の場合).

$p\backslash n$	4	8	16	32	64	128
4	0.8474	0.5378	0.0686	0.0842	0.0688	0.0553
8	0.9921	0.8471	0.4472	0.1865	0.0784	0.0573
16	1.0000	0.9882	0.7248	0.2606	0.0966	0.0643
32	1.0000	0.9999	0.9445	0.3735	0.1062	0.0664
64	1.0000	1.0000	0.9986	0.5787	0.1131	0.0674
128	1.0000	1.0000	1.0000	0.8486	0.1252	0.0677

表10: $t_{n,p}$の実際の有意水準–出現確率20%2値データ(α =0.05の場合).

$p\backslash n$	4	8	16	32	64	128
4	0.5442	0.1207	0.0503	0.0481	0.0458	0.0457
8	0.8989	0.3890	0.0890	0.0494	0.0468	0.0461
16	0.9961	0.7203	0.1464	0.0534	0.0494	0.0492
32	1.0000	0.9547	0.2631	0.0566	0.0505	0.0499
64	1.0000	0.9993	0.4912	0.0606	0.0503	0.0499
128	1.0000	1.0000	0.8034	0.0686	0.0503	0.0499

表11: $t_{n,p}$の実際の有意水準–出現確率30%2値データ(α =0.05の場合).

$p\backslash n$	4	8	16	32	64	128
4	0.2884	0.0481	0.0442	0.0456	0.0464	0.0462
8	0.6773	0.1300	0.0485	0.0473	0.0470	0.0469
16	0.9373	0.2761	0.0547	0.0495	0.0499	0.0493
32	0.9981	0.5393	0.0649	0.0506	0.0506	0.0503
64	1.0000	0.8492	0.0849	0.0508	0.0507	0.0508
128	1.0000	0.9895	0.1314	0.0513	0.0509	0.0511

表12: $t_{n,p}$の実際の有意水準–出現確率40%2値データ(α =0.05の場合).

$p\backslash n$	4	8	16	32	64	128
4	0.1583	0.0457	0.0476	0.0477	0.0473	0.0466
8	0.4645	0.0676	0.0525	0.0505	0.0484	0.0475
16	0.7810	0.1046	0.0545	0.0533	0.0518	0.0500
32	0.9683	0.1825	0.0563	0.0539	0.0522	0.0511
64	0.9995	0.3468	0.0583	0.0546	0.0531	0.0510
128	1.0000	0.6286	0.0610	0.0548	0.0530	0.0515

4 $\boldsymbol{x}_1$ から $\boldsymbol{x}_2$ の影響を除いた変数の独立性検定

p 個の変数 $x_1,\ldots,x_p$ と関連する q 個の変数 $x_{p+1},\ldots,x_m$ があるとし,

$$\boldsymbol{x}=(\boldsymbol{x}'_1,\ \boldsymbol{x}'_2)',\ \boldsymbol{x}_1=(x_1,\ldots,x_p)',\ \boldsymbol{x}_2=(x_{p+1},\ldots,x_m)'$$

する．ここに，$m=p+q$ である．$\boldsymbol{x}$ の平均ベクトル，共分散行列をそれぞれ $\boldsymbol{\mu},\Sigma$ とし，大きさ $N=n+1$ の標本に基づく標本共分散行列を S とする．また，$\boldsymbol{x}$ の分割に対応して,

$$\boldsymbol{\mu}=\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix},\quad \Sigma=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix},\quad S=\begin{pmatrix}S_{11} & S_{12}\\ S_{21} & S_{22}\end{pmatrix}$$

と分割する．このとき，$\boldsymbol{x}_1$ から $\boldsymbol{x}_2$ の影響を除いた(残差)変数は

$$\boldsymbol{y} = \boldsymbol{x}_1 - \boldsymbol{\mu}_1 - \Sigma_{12}\,\Sigma_{22}^{-1}\,(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$$

によって定義される．$\boldsymbol{y}$ の母共分散行列は

$$\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\,\Sigma_{22}^{-1}\,\Sigma_{21} = (\sigma_{ij\cdot p+1,\ldots,m})$$

で与えられ，また，標本共分散行列は

$$S_{11\cdot 2} = S_{11} - S_{12}S_{22}^{-1}S_{21} = (s_{ij\cdot p+1,\ldots,m})$$

で与えられる．さらに，y_i と y_j の母相関係数，標本相関係数はそれぞれ

$$\rho_{ij\cdot p+1,\ldots,m} = \frac{\sigma_{ij\cdot p+1,\ldots,m}}{\sqrt{\sigma_{ii\cdot p+1,\ldots,m}}\sqrt{\sigma_{jj\cdot p+1,\ldots,m}}}$$

$$r_{ij\cdot p+1,\ldots,m} = \frac{s_{ij\cdot p+1,\ldots,m}}{\sqrt{s_{ii\cdot p+1,\ldots,m}}\sqrt{s_{jj\cdot p+1,\ldots,m}}}$$

で与えられる．

さて，$\boldsymbol{x}$ が m 次元正規分布従うとして，$y_1,\ldots,y_p$ が互いに独立あるという仮説の検定について考える．このとき，$\boldsymbol{y}$ も p 次元正規分布に従い，独立性の仮説は $H_{03}:\rho_{ij\cdot p+1,\ldots,m} = 0\,(i>j)$ として書ける．これらについての詳しい解説は，例えば Anderson [1] を参照されたい．さらに，$nS_{11\cdot 2}$ はウィシャート分布 $W_p(n-q,\Sigma_{11\cdot 2})$ に従うことが知られている．これらのことから Schott [5] の結果が適用できることがわかる．実際，検定法は n を $n-q$ に変えることによって作られる．つまり，検定統計量は

$$t^*_{n-q,p} = \frac{\sum_{i=2}^{p}\sum_{j=1}^{i-1} r^2_{ij\cdot p+1,\ldots,m} - \dfrac{p(p-1)}{2(n-q)}}{\sigma_{t_{n-q,p}}}$$

で与えられる．ここに

$$\sigma^2_{t^*_{n-q,p}} = \frac{p(p-1)(n-q-1)}{(n-q)^2(n-q+2)}$$

である．高次元漸近的枠組における条件は

$$\lim \frac{p}{n-q} = \gamma_2 \in (0,\ \infty)$$

である．なお，このとき

$$\lim \sigma^2_{t^*_{n-q,p}} = \lim \frac{p(p-1)(n-q-1)}{(n-q)^2(n-q+2)} = \gamma_2^2$$

となっている．これらのことから，次の定理を得る．

定理 4.1　$\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2')'$ は $m = p+q$ 次元正規分布に従うとする．$\boldsymbol{x}_1$ から $\boldsymbol{x}_2$ の影響を除いた変数 $\boldsymbol{y}$ の独立性検定統計量 $t^*_{n-q,p}$ 仮説のもとでの分布は，$\lim\{p/(n-q)\} = \gamma_2 \in (0,\ \infty)$ のとき標準正規分布に収束する．

帰無仮説のもとでの $t^*_{n-q,p}$ の近似の良さをシミュレーションを用いて調べた．具体的には，棄却確率を α=0.05 として，帰無仮説のもとで，実際の棄却確率を調べた．このとき，$q=4$ として，n と p は 8, 16, 32, 64, 128 の場合を調べた．ここで $p=4$ の場合を除いているが，これは q が n 以上であると，$S_{11\cdot 2}$

の分布がウィシャート分布に従わないからである．$t^*_{n-q,p}$ のシミュレーション結果は表 13 に与えられている．すべての場合について，実際の棄却確率は目標の棄却確率 α=0.05 にほぼ等しくなっている．

表13: $t^*_{n-q,p}$ の実際の有意水準 $-q=4$ ($\alpha=0.05$の場合).

$p\backslash n$	8	16	32	64	128
8	0.0467	0.0432	0.0445	0.0457	0.0460
16	0.0465	0.0475	0.0482	0.0481	0.0481
32	0.0469	0.0481	0.0493	0.0497	0.0498
64	0.0472	0.0479	0.0497	0.0500	0.0498
128	0.0473	0.0482	0.0491	0.0498	0.0499

5 結論

本論文では，高次元における独立性の検定について調べた．第 2 節で述べている 4 つの検定法 (1)～(4) を取り上げ，これらの実際の棄却確率が目標とする棄却確率 α=0.05 になっているかどうかについて，シミュレーションで調べた．実際には，高次元の状況のみならず大標本の状況においても数値実験を行なった．その結果，検定法 (3) は，次元数 p と標本数 n の様々な値に対して，目標の棄却確率のよい近似になっていることが確認された．また，他の検定法については，目標の棄却確率のよい近似になる場合は，かなり限られることもわかった．

次に，正規性を仮定したもとで，妥当な棄却確率になっている検定法 (3) について，離散データの場合の影響を数値的に調べた．その結果，離散分布のタイプにも依存するが，概ねその影響は少なく，頑健性を持つことがわかった．

最後に，p 個の変数に加え，これに関連する q 個の変数があるとき，最初の p 個の変数から，残りの変数の影響を除いたときの独立性検定問題を考えた．この検定問題に対して，検定法 (3) を拡張した検定法を提案し，妥当な棄却確率になっていることを数値的に確認した．

謝辞

査読者の方には貴重なコメントを頂きました．ここに記して，お礼を申し上げます．また本研究は理工学研究所，共同研究第 2 類「多変量高次元データ解析の理論と応用」(2007 年度) から研究助成を受けております．

参考文献

[1] ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd ed.). John Wiley & Sons, New York, NY.

[2] FUJIKOSHI, Y. (2004). Multivariate analysis for the case when the dimension is large compared to the sample size. *J. Koren Statist. Soc.*, **33**, , 1-24.

[3] Hsu, P. L. (1949). The limiting distribution of functions of sample means and applications to testing hypothesis, Proceedings of the First Berkeley Symposium of Mathematical Statistics and Probability (ed. J. Neyman), Univ. of California Press, Berkeley and Los Angeles, 359-402.

[4] LEDOIT, O. and Wolf, M.(2002). Some hypotheses tests for the covariance matrix when the dimension is large compared to the sample size. *Ann*. *Statist*. **30**, 1081-1102.

[5] SCHOTT, J. R.(2005). Testing for complete independence in high dimensions. *Biometrika*, **92**, 951-956.

[6] SCHOTT, J. R.(2007). Some high-dimensional tests for a one-way MANOVA. *J. Multivariate Anal*. 98, 1825-1839.

[7] SRIVASTAVA, M. S.(2005). Some tests concerning the covariance matrix in high dimensional data. *J. Japan Statist. Soc*. **37**, 251-272.

[8] SRIVASTAVA, M. S. and FUJIKOSHI, Y.(2006). Multivariate analysis of variance with fewer observations than the dimension. *J. Multivariate Anal*. **97**, 1927-1940.

[9] SRIVASTAVA, M. S.(2007). Multivariate theory for analyzing high dimensional data. *J. Japan Statist. Soc*. **37**, 53-86.