高次元データにおける2次形式の近似について

藤本 翔太，狩野 裕
大阪大学大学院基礎工学研究科
sfujimoto@sigmath.es.osaka-u.ac.jp, kano@sigmath.es.osaka-u.ac.jp

要 約：多変量1標本問題において変数の次元が標本サイズに比して相対的に大きい場合，標本共分散行列が特異になり，Hotelling の T^2 統計量による検定を行うことができない．このような場合の検定方法は，ある条件の下で Dempster (1958, 1960)，Bai and Saranadasa (1996)，Fujikoshi (2004)，Srivastava (2007) などによって提案されている．本報告では，彼らの条件よりも現実的な条件の下で新たな検定方法を構成する．

1 背景

本報告で扱う多変量1標本問題においては，高次元データを想定する場合，標本共分散行列 $\boldsymbol{S}_{n,p}$ が特異になり，$\boldsymbol{S}_{n,p}^{-1}$ を伴う検定統計量，例えば Hotelling の T^2 統計量を構成することができない．このような場合の検定方法を最初に扱ったのは Dempster (1958, 1960) であり，これを基にして Bai and Saranadasa (1996)，Fujikoshi (2004)，Srivastava (2007) などによって研究が行われてきた．これらの手法はどれも，真の共分散行列 $\boldsymbol{\Sigma}_p$ に対して次の条件を課した下で求められている．

条件 (A)
$$\lim_{p\to\infty}\frac{\mathrm{tr}\boldsymbol{\Sigma}_p^i}{p}\in(0,\,\infty)\quad(i=1,2,3,4)$$

本報告では，条件 (A) とは異なる条件の下で検定法を構成する．これは，条件 (A) をみたす状況に限らず他の状況においても検定ができるという点で有用であると考えられる．具体的には次の条件を考える．

条件 (B) p に依存しない定数 $a_i\,(i=1,2,\cdots,p)$, $\alpha>0$, $C>0$, $m<p$ が存在して

$$\sum_{i=m+1}^{p}a_i^k=o(p^{k\alpha})\,(k=1,2,3,4)$$

$$\lambda_i=a_i+Cp^{\alpha}\ \ (i=1,2,\cdots,m),\quad \lambda_j=a_j\ \ (j=m+1,m+2,\cdots,p)$$

ここに，$\lambda_i\,(i=1,2,\cdots,p)$ は $\boldsymbol{\Sigma}_p$ の固有値である．条件 (B) は m 個の固有値が p のオーダで発散し，残りの固有値が定数であるような状況を想定している．例えば，共分散行列が $\boldsymbol{\Sigma}_p=(1-\rho)I_p+\rho\mathbf{1}_p\mathbf{1}_p^T\,(0\le\rho<1)$ の場合がそれにあたる．ここに I_p は p 次の単位行列，$\mathbf{1}_p$ は要素がすべて1の p 次元列ベクトルである．実際，固有値は $\lambda_1=1+(p-1)\rho$, $\lambda_2=\cdots=\lambda_p=1-\rho$ であるから，$\rho=0$ でない限り条件 (B) をみたす．

2 高次元データにおける多変量1標本問題

Dempster (1958, 1960)，Bai and Saranadasa (1996)，Fujikoshi (2004)，Srivastava (2007) の結果を簡単に紹介する．検定方法の性質から，2つのタイプに分けられる．以下，$\boldsymbol{X}_p^{(i)}\overset{i.i.d.}{\sim}N_p(\boldsymbol{\mu}_p,\,\boldsymbol{\Sigma}_p)\,(i=1,2,\cdots,n)$ とし，

$$\bar{\boldsymbol{X}}_{n,p}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_p^{(i)},\quad \boldsymbol{S}_{n,p}=\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_p^{(i)}-\bar{\boldsymbol{X}}_{n,p})(\boldsymbol{X}_p^{(i)}-\bar{\boldsymbol{X}}_{n,p})^T$$

とする．また，帰無仮説 H_0 および対立仮説 H_1 を以下のように設定する．

$$H_0:\boldsymbol{\mu}_p=\mathbf{0}\quad\text{versus}\quad H_1:\boldsymbol{\mu}_p\neq\mathbf{0}$$

2.1 F 近似を用いる方法

高次元データにおいては $\boldsymbol{S}_{n,p}$ が特異になるため，Hotelling の T^2 統計量 $T_H = n\bar{\boldsymbol{X}}_{n,p}^T \boldsymbol{S}_{n,p}^{-1} \bar{\boldsymbol{X}}_{n,p}$ は定義されない．Dempster (1958, 1960) は $\boldsymbol{S}_{n,p}$ を $\mathrm{tr}\boldsymbol{S}_{n,p}$ で置き換えた次の統計量を提案した．

$$T_D = \frac{n\bar{\boldsymbol{X}}_{n,p}^T \bar{\boldsymbol{X}}_{n,p}}{\mathrm{tr}\boldsymbol{S}_{n,p}}$$

データ行列を適当な直交行列で変換することによって，T_D は帰無仮説 H_0 の下，

$$T_D = \frac{(n-1)Q_p^{(1)}}{Q_p^{(2)} + \cdots + Q_p^{(n)}}$$

と表せる．ここに，$Q_p^{(i)} = \boldsymbol{Y}_p^{(i)T}\boldsymbol{Y}_p^{(i)}$, $\boldsymbol{Y}_p^{(i)} \overset{i.i.d.}{\sim} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_p)$ $(i = 1,2,\cdots,n)$ である．T_D の分布を求めるために，$Q_p^{(i)} \overset{i.i.d.}{\sim} m\chi_r^2$ $(i = 1,2,\cdots,n)$ のように近似できれば，$T_D \sim F_{r,(n-1)r}$ となる．Dempster (1958, 1960) は 3 つの方法で r を推定し，$T_D \sim F_{\hat{r},(n-1)\hat{r}}$ として検定を行うことを提案した．しかしながら，r の 3 つの推定量はどれも直交変換に依存しており，検定結果の解釈が難しいという問題点がある．そのため，本報告ではこれ以上扱わないことにする．詳細は Dempster (1958, 1960) を参照されたい．

ところで，$Q_p^{(i)} \overset{i.i.d.}{\sim} m\chi_r^2$ の近似を 2 次モーメントまで一致するように行えば，$r = (\mathrm{tr}\boldsymbol{\Sigma}_p)^2/\mathrm{tr}\boldsymbol{\Sigma}_p^2$ となる．このとき，Srivastava (2007) は次の定理を示し，r の ratio-consistent な推定量を求め，その推定量を用いて上記の検定を行うことを提案している．この推定量は直交変換に依存しない．

定理 2.1

条件 (A) を仮定する．このとき任意の $\varepsilon > 0$ に対して

$$\lim_{n\to\infty} \sup_p P\left(\left|\frac{\hat{a}_1}{\mathrm{tr}\boldsymbol{\Sigma}_p} - 1\right| > \varepsilon\right) = 0, \quad \lim_{n\to\infty} \sup_p P\left(\left|\frac{\hat{a}_2}{\mathrm{tr}\boldsymbol{\Sigma}_p^2} - 1\right| > \varepsilon\right) = 0, \quad \lim_{n\to\infty} \sup_p P\left(\left|\frac{\hat{r}}{r} - 1\right| > \varepsilon\right) = 0$$

が成立つ．ここに，

$$\hat{a}_1 = \mathrm{tr}\boldsymbol{S}_{n,p}, \quad \hat{a}_2 = \frac{(n-1)^2}{(n+1)(n-2)}\left[\mathrm{tr}\boldsymbol{S}_{n,p}^2 - \frac{1}{n-1}(\mathrm{tr}\boldsymbol{S}_{n,p})^2\right], \quad \hat{r} = \frac{\hat{a}_1^2}{a_2}$$

2.2 漸近分布を用いる方法

Bai and Saranadasa (1996)，Fujikoshi (2004) は条件 (A) を仮定して，帰無仮説 H_0 の下で $\bar{\boldsymbol{X}}_{n,p}^T\bar{\boldsymbol{X}}_{n,p}$, $\mathrm{tr}\boldsymbol{S}_{n,p}$ に関する漸近正規性を示した．

定理 2.2

条件 (A) を仮定し，$\lim_{p\to\infty} n = \infty$ とする．このとき，帰無仮説 H_0 の下で

$$T_{BS} = \frac{1}{\sqrt{2\mathrm{tr}\boldsymbol{\Sigma}_p^2}}\left(n\bar{\boldsymbol{X}}_{n,p}^T\bar{\boldsymbol{X}}_{n,p} - \mathrm{tr}\boldsymbol{S}_{n,p}\right) \xrightarrow{d} N(0,1) \quad \text{as } p \to \infty$$

$$T_F = \frac{\mathrm{tr}\boldsymbol{\Sigma}_p}{\sqrt{2\mathrm{tr}\boldsymbol{\Sigma}_p^2}}(T_D - 1) \xrightarrow{d} N(0,1) \quad \text{as } p \to \infty$$

この定理により，$\mathrm{tr}\boldsymbol{\Sigma}_p$, $\mathrm{tr}\boldsymbol{\Sigma}_p^2$ が既知の場合は正規分布を用いて検定できるが，未知の場合はそれぞれ定理 2.1 の $\hat{a}_1$, $\hat{a}_2$ で置き換えて検定を行うことを提案している．

3 主結果

第 2 節で示した 2 種類の検定方法は条件 (A) をみたすことを前提としている．しかしながら，条件 (A) は Cauchy-Schwarz の不等式の等号成立条件からわかるように，$\boldsymbol{\Sigma}_p$ が単位行列に近い場合でしか成立たない．そこで，条件 (A) をより現実的な $\boldsymbol{\Sigma}_p$ を許すような条件に置き換えることを考える．

F 近似を用いる方法においては，次の定理が示される．

定理 3.1 定理 2.1 は次の条件を仮定しても成立つ．

条件 (C)

$$^{\exists}\delta_i \geq 0\ (i=1,2,4)\ \text{ s.t. }\ \delta_1 \geq \delta_2 \geq \delta_4\ \text{ and }\ \lim_{p\to\infty} \mathrm{tr}\left(\frac{\boldsymbol{\Sigma}_p}{p^{\delta_i}}\right)^i \in (0,\infty)\ \ (i=1,2,4)$$

この定理から，条件 (A) よりも弱い条件 (C) を仮定しても，F 近似による検定を行うことができる．また，条件 (B) は条件 (C) より強いので，条件 (B) の場合も F 近似による検定を行うことができる．

漸近分布を用いる方法においては，次の定理が示される．

定理 3.2

条件 (B) を仮定し，$\lim_{p\to\infty} n = \infty$ とする．このとき，帰無仮説 H_0 の下で

$$T_{BS} = \frac{1}{\sqrt{\mathrm{tr}\boldsymbol{\Sigma}_p^2}}\left(n\bar{\boldsymbol{X}}_{n,p}^T\bar{\boldsymbol{X}}_{n,p} - \mathrm{tr}\boldsymbol{S}_{n,p}\right) \xrightarrow{d} \frac{\chi_m^2 - m}{\sqrt{m}} \quad \text{as } p\to\infty$$

$$T_F = \frac{\mathrm{tr}\boldsymbol{\Sigma}_p}{\sqrt{\mathrm{tr}\boldsymbol{\Sigma}_p^2}}(T_D - 1) \xrightarrow{d} \frac{\chi_m^2 - m}{\sqrt{m}} \quad \text{as } p\to\infty$$

この定理により，$\mathrm{tr}\boldsymbol{\Sigma}_p$, $\mathrm{tr}\boldsymbol{\Sigma}_p^2$ が既知の場合は χ^2 分布を用いて検定できる．未知の場合には定理 3.1 より，未知の部分を $\hat{a}_1$, $\hat{a}_2$ で置き換えて検定を行うことができる．

4 数値実験

各統計量の近似の精度をみるために，Monte Carlo 実験を行う．$\boldsymbol{X}_p^{(i)} \overset{i.i.d.}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma}_p)\ (i=1,2,\cdots,n)$, $\boldsymbol{\Sigma}_p = (1-\rho)I_p + \rho\mathbf{1}_p\mathbf{1}_p^T\ (0\leq\rho<1)$ とし，有意水準 $\alpha = 0.05$，繰り返し回数 500 とする．例えば F 近似を行う方法の場合は，

$$\frac{\#\{T_D > F_{\hat{r},(n-1)\hat{r}}(\alpha)\}}{500}$$

を計算し，どれだけ $\alpha = 0.05$ に近いかをみる．条件 (A) は $\rho = 0$ の場合でのみ成立ち，条件 (B) および条件 (C) は $\rho = 0$ の場合を除いて成立つ．次の頁に数値実験の結果を記載する．T_S は Srivastava (2007) の方法を表し，T_B^N, T_F^N はそれぞれ T_B, T_F を正規近似して検定を行う方法を表す．さらに，T_B^χ, T_F^χ はそれぞれ T_B, T_F を χ^2 近似して検定を行う方法を表す．

表 1: Monte Carlo 実験

ρ	n	p	T_S	T_F^N	T_{BS}^N	T_F^χ	T_{BS}^χ
0	40	40	0.050	0.066	0.08	0.060	0.048
	40	100	0.032	0.060	0.056	0.046	0.042
	40	200	0.054	0.066	0.084	0.032	0.036
	80	100	0.048	0.046	0.052	0.040	0.038
	80	200	0.046	0.052	0.070	0.026	0.026
	150	200	0.062	0.042	0.066	0.036	0.020
0.3	40	40	0.068	0.066	0.074	0.042	0.064
	40	100	0.070	0.066	0.070	0.056	0.044
	40	200	0.076	0.092	0.060	0.054	0.058
	80	100	0.054	0.058	0.048	0.056	0.048
	80	200	0.082	0.100	0.062	0.066	0.062
	150	200	0.056	0.06	0.082	0.030	0.042
0.5	40	40	0.056	0.062	0.072	0.066	0.068
	40	100	0.038	0.062	0.076	0.06	0.062
	40	200	0.056	0.082	0.058	0.074	0.042
	80	100	0.062	0.064	0.066	0.052	0.026
	80	200	0.048	0.064	0.056	0.044	0.040
	150	200	0.052	0.064	0.078	0.058	0.038
0.8	40	40	0.052	0.086	0.074	0.058	0.044
	40	100	0.058	0.066	0.062	0.054	0.058
	40	200	0.050	0.086	0.060	0.048	0.066
	80	100	0.056	0.056	0.078	0.058	0.060
	80	200	0.060	0.076	0.068	0.058	0.054
	150	200	0.044	0.076	0.080	0.064	0.052

参考文献

[1] Bai, Z. and Saranadasa, H. (1996), Effect of high dimension : by an example of a two sample problem, *Statistica Sinica*, **6**, 311–329.

[2] Dempster, A. P. (1958), A high dimensional two sample significance test, *The Annals of Mathematical Statistics*, **29**, 995–1010.

[3] Dempster, A. P. (1960), A significance test for the separation of two highly multivariate small samples, *Biometrics*, **16**, 41–56.

[4] 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, *Journal of the Japan Statistical Society*, **34**, 19–26.

[5] Srivastava, M.S. (2007), Multivariate theory for analyzing high dimensional data, *Journal of the Japan Statistical Society*, **37**, 53–86.