多次元データの正規性検定について

東京理科大・理・院　小泉 和之
東京理科大・理　岡本 直也
東京理科大・理　瀬尾　隆

多次元データが多変量正規分布に従っているかを調べる多変量正規性検定には様々な手法が提案されているが, その中に多変量標本歪度, 多変量標本尖度 (以下, それぞれ歪度, 尖度と呼ぶ) を用いた方法がある．本報告では, Srivastava (1984, *Statist. Probab. Lett.*) の定義による歪度, 尖度を用いた正規性検定統計量について議論した.

$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ を互いに独立な平均 $\boldsymbol{\mu}$, 分散共分散行列 Σ をもつ分布に従う p 次元確率ベクトルとする． Σ は, 直交行列 $\Gamma = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_p)$ と $D_\lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ ($\lambda_i,\ i = 1, 2, \ldots, p$ は Σ の固有値) を用いて, $\Sigma = \Gamma D_\lambda \Gamma'$ と表される．また, 標本平均ベクトルと標本共分散行列をそれぞれ $\overline{\boldsymbol{x}}$, $S = HD_\omega H'$ とする．ここに, $H = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_p)$ は直交行列, $D_\omega = \mathrm{diag}(\omega_1, \omega_2, \ldots, \omega_p)$ ($\omega_i,\ i = 1, 2, \ldots, p$ は S の固有値) である．Srivastava (1984) は, $y_{ij} = \boldsymbol{h}_i' \boldsymbol{x}_j (j = 1, 2, \ldots, N)$, $\overline{y}_i = N^{-1} \sum_{j=1}^{N} y_{ij}$ を用いて, 以下のように歪度, 尖度を定義した.

$$b_{1,p}^2 \equiv \frac{1}{N^2 p} \sum_{i=1}^{p} \left\{ \omega_i^{-\frac{3}{2}} \sum_{j=1}^{N} (y_{ij} - \overline{y}_i)^3 \right\}^2,$$

$$b_{2,p} \equiv \frac{1}{Np} \sum_{i=1}^{p} \omega_i^{-2} \sum_{j=1}^{N} (y_{ij} - \overline{y}_i)^4.$$

本報告では, 歪度, 尖度の両方を考慮した総括的な検定統計量を提案するとともに, 非正規のもとで, 歪度 $b_{1,p}^2$ に対する期待値を求めた.

Srivastava (1984) は, 正規性の仮定のもとで, 歪度, 尖度それぞれを展開することで以下の Theorem を得た.

Theorem 1 (Srivastava (1984)) 十分大きな N に対して,

$$\frac{Np}{6} b_{1,p}^2 \sim \chi_p^2, \ \frac{\sqrt{Np}(b_{2,p} - 3)}{\sqrt{24}} \sim N(0, 1)$$

となる.

このことより, 1 変量で考えられている歪度と尖度の両方を同時に考慮した Jarque-Bera 統計量を多変量に拡張した

$$MJB \equiv Np \left\{ \frac{b_{1,p}^2}{6} + \frac{(b_{2,p} - 3)^2}{24} \right\}$$

を提案した．この統計量は漸近的に自由度 $p + 1$ の χ^2 分布に従う.

また, 期待値と分散をそれぞれ改良した以下の Lemma が得られた.

Lemma 1 十分大きな N に対して,

$$\begin{aligned}
E(b_{1,p}^2) &= \frac{6(N-2)}{(N+1)(N+3)},\\
Var(b_{1,p}^2) &= \frac{1}{p}\frac{72N(N-2)(N^3+37N^2+11N-313)}{(N+1)^2(N+3)^2(N+5)(N+7)(N+9)},\\
E(b_{2,p}) &= \frac{3(N-1)}{N+1},\\
Var(b_{2,p}) &= \frac{24}{p}\frac{N(N-2)(N-3)}{(N+1)^2(N+3)(N+5)}
\end{aligned}$$

である.

Lemma 1 を用いることにより, もう一つの総括的な検定統計量 (MJB' 統計量) として,

$$MJB' \equiv \frac{pb_{1,p}^2}{E(b_{1,p}^2)} + \frac{(b_{2,p}-E(b_{2,p}))^2}{Var(b_{2,p})}$$

を提案した. この統計量も MJB と同様に漸近的に自由度 $p+1$ の χ^2 分布に従う. これらの提案した統計量に対し, 期待値, 分散, 上側パーセント点のシミュレーションを行った.

次に, 母集団分布に楕円分布を仮定したもとで $b_{1,p}^2$ の期待値の計算を行った. つまり, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ は互いに独立に $\mathrm{E}_p(\boldsymbol{\mu}, \Lambda)$ に従う. このとき, $\boldsymbol{x}_j$ の特性関数は, $E(i\boldsymbol{t}'\boldsymbol{\mu})\Psi(\boldsymbol{t}'\Lambda\boldsymbol{t})$ であり, $Cov(\boldsymbol{x}_j) = -2\Psi'(0)\Lambda \equiv \Sigma$ である. 楕円分布のもとでは, 一般に $\overline{\boldsymbol{x}}$ と S は独立ではないため, Iwashita (1997, *J. Statist. Plann. Inference*) によって漸近展開された同時密度を用いた.

ここで, 対角化した標本共分散行列は,

$$D_\omega = \frac{1}{N-1}\sum_{j=1}^{N}(\boldsymbol{y}_j - \overline{\boldsymbol{y}})(\boldsymbol{y}_j - \overline{\boldsymbol{y}})'$$

と表されることに注意し, $\boldsymbol{y}_\alpha$, $\boldsymbol{y}_\beta$, $\overline{\boldsymbol{y}}$, D_ω の従属性をさけるために次を定義する.

$$\begin{aligned}
\overline{\boldsymbol{y}}^{(\alpha,\beta)} &= \frac{1}{N-2}\sum_{j\neq\alpha,\beta}^{N}\boldsymbol{y}_j,\\
D_\omega^{(\alpha,\beta)} &= \frac{1}{N-3}\sum_{j\neq\alpha,\beta}^{N}(\boldsymbol{y}_j - \overline{\boldsymbol{y}}^{(\alpha,\beta)})(\boldsymbol{y}_j - \overline{\boldsymbol{y}}^{(\alpha,\beta)})'.
\end{aligned}$$

また, 一般性を失うことなく, $\boldsymbol{\mu} = \boldsymbol{0}$, $\Sigma = I_p$ としてよいから,

$$\begin{aligned}
\sqrt{N-2}\,\overline{\boldsymbol{y}}^{(\alpha,\beta)} &= \boldsymbol{z} \equiv (z_1, z_2, \ldots, z_p)',\\
\sqrt{N-2}(D_\omega^{(\alpha,\beta)} - I_p) &= Z \equiv \{z_{ab}\}
\end{aligned}$$

とおく. それにより $b_{1,p}^2$ を摂動展開することで, 次の定理を得た.

Theorem 2 標本がパラメータ $\boldsymbol{\mu}$, Λ の楕円分布に従うとき, 歪度 $b_{1,p}^2$ の期待値は,

$$E(b_{1,p}^2) = \frac{1}{N}(6 + 15\kappa_{(3)} - 18\kappa_{(2)}) + o(N^{-1})$$

となる. ここに, $\kappa_{(k)}$ は k 次のモーメントパラメータで, $\kappa_{(k)} = \Psi^{(k)}(0)/\{\Psi'(0)\}^k - 1$ である.