平均および分散共分散行列のロバスト推定法

関西大学 経済学部　渡　邊　美智子*
立教大学 社会学部　山　口　和　範

多くの多変量解析において，まず標本平均および標本共分散行列が計算され，それらに基づいてパラメータの推定や検定が行われる．この場合，一般にデータに対して多変量正規性が仮定されているが，外れ値の存在するデータに対して正規性に依存する推測法が効率的でないことはよく知られている．外れ値が存在する場合の平均ベクトルと分散共分散行列のロバストな推定法として，Rubin (1983) や Little (1988) は，多変量 t 分布や混淆多変量正規分布等の正規分布の尺度混合分布族に基づく最尤法を提唱している．ここでは，正規分布よりも裾の重い(尖度の大きい)分布をデータに適合させることで外れ値の影響を受けないパラメータの推定を意図している．一方，多変量 t 分布等の楕円分布では各周辺分布の尖度が等しいため，現実のデータへの適合に際して不十分である場合も多い．実際，Cook and Johnson (1981) や Kano et al. (1990) に周辺分布の尖度に一様性を仮定できない実際例が挙げられている．

本研究では，Dempster et al.(1980), Little (1988) 等で使用されている正規分布の尺度混合分布族を拡張する観点からある分布族を考え，その下での平均ベクトルと分散共分散行列の最尤推定量を計算するアルゴリズムを与える．

平均 $\boldsymbol{\mu}$，分散 $\boldsymbol{\Sigma}$ の p 次の多変量正規分布に従う確率ベクトル $\boldsymbol{x}$ は，互いに独立な標準正規変数を要素にもつ確率ベクトル $\boldsymbol{e}$ (i.e., $\boldsymbol{e}\sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_p)$) の線形変換により得られる：$\boldsymbol{x}=\boldsymbol{A}^{-1}\boldsymbol{e}+\boldsymbol{\mu}$，ただし，$\boldsymbol{A}^{-1}(\boldsymbol{A}^{-1})'=\boldsymbol{\Sigma}$ で，$\boldsymbol{A}$ は対角成分が正である上三角行列とする．多変量 t 分布などの正規分布より尖度の大きい正規分布の尺度混合分布族は，一般に，ある正の確率変数 q が与えられた下で，$\boldsymbol{x}$ の条件付分布が $\boldsymbol{x}\sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/q)$ であると仮定することにより得られる．このことは，q が与えられた下で，$\boldsymbol{e}$ の条件付分布を $\boldsymbol{e}\sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_p/q)$ とおくことと等しい．ここでは，各変数に対して共通の確率変数 q を用いているため各変数の周辺分布の尖度は等しくなる．そこで，周辺分布に多様性をもたせるため，確率変数 q を互いに独立な p 個の正の確率変数からなる確率ベクトル $\boldsymbol{q}=(q_1, q_2, \ldots, q_p)'$ に置き換えて得られる分布族を考える．即ち，ある正の確率ベクトル $\boldsymbol{q}$ が与えられた下での $\boldsymbol{e}$ の条件付分布が $\mathrm{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1})$，ここに，$\boldsymbol{Q}=\mathrm{diag}\{q_1, q_2, \ldots, q_p\}$ である場合に得られる分布を対象とする．一般に，$E(\boldsymbol{x})=\boldsymbol{\mu}$, $\mathrm{Cov}(\boldsymbol{x})=\boldsymbol{A}^{-1}\boldsymbol{Q}^*\boldsymbol{A}^{-1\prime}$，ただし，$\boldsymbol{Q}^*$ は $\int_0^\infty q_j^{-1}M_j(q_j)dq_j$ を (j, j) 成分にもつ対角行列である．また，観測値が得られた下での q_j^m の条件付期待値は，q_j の確率（密度）関数を $M_j(q_j)$ としたとき，

$$E(q_j^m \mid \boldsymbol{x};\ \boldsymbol{\mu},\ \boldsymbol{A})=\frac{\int_0^\infty q_j^{m+1/2}\exp(-q_je_j^2/2)M_j(q_j)dq_j}{\int_0^\infty q_j^{1/2}\exp(-q_je_j^2/2)M_j(q_j)dq_j} \tag{1}$$

となる．ここに，e_j は $\boldsymbol{e}=\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})$ の第 j 成分である．とくに，q_j が退化した確率変数(定数)でなければ，$E(q_j|\boldsymbol{x};\ \boldsymbol{\mu},\ \boldsymbol{A})$ は e_j^2 の非増加関数である．

次に，この分布の仮定の下で，$\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の最尤推定量を得るためのアルゴリズムを導く．ここでは，直接 $\boldsymbol{\Sigma}$ を推定する代わりに $\boldsymbol{A}$ を推定する．$\boldsymbol{x}$ の分布型を具体的に規定し，その下で直接に尤度関数を評価し，最尤推定量を求めるアルゴリズムを構築する方法も考えられるが，アル

* 現 東洋大学 経済学部

ゴリズムの具体化は煩雑であろう．しかし，$\{\boldsymbol{x}_i, \boldsymbol{q}_i;\ i=1, 2, \ldots, n\}$ を完全データ，$\{\boldsymbol{x}_i;\ i=1, 2, \ldots, n\}$ を $\{\boldsymbol{q}_i;\ i=1, 2, \ldots, n\}$ が欠測した不完全データとみなして EM アルゴリズムを適用することにより，比較的容易に最尤推定量を求めるアルゴリズムを構築できる．それぞれの step は以下のとおりである．

E-step: $\boldsymbol{W}_i=E(\boldsymbol{Q}_i|\boldsymbol{x}_i;\ \boldsymbol{\mu}, \boldsymbol{A})$ を計算する．

M-step: 次の方程式を解く．

$$\sum_{i=1}^{n} \boldsymbol{W}_i\boldsymbol{A}(\boldsymbol{x}_i-\boldsymbol{\mu})=\boldsymbol{0}, \tag{2}$$

$$\mathrm{diag}\{a_{11}^{-1}, a_{22}^{-1}, \ldots, a_{pp}^{-1}\}-(1/n)\sum_{i=1}^{n} \boldsymbol{W}_i\boldsymbol{A}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})'=\boldsymbol{0} \tag{3}$$

(2) と (3) からも明らかなように，w_{ij} は i 番目の観測個体の誤差ベクトル $\boldsymbol{e}_i$ の j 変量に対する重みの役割を果たしていて，観測値と期待値のかい離が大きくなるほど小さな値をとる．従って，q_j の分布に含まれるパラメータは，データの外れ具合に応じてどの程度の重みを与えるかの調整を行っている．実際のデータ解析において，これらのパラメータはあらかじめ規定できないのが普通である．よって，これらのパラメータを所与のデータから推定する必要がある．一般に，$M_j(q_j;\ \boldsymbol{\theta})$ に未知パラメータ $\boldsymbol{\theta}$ が含まれている場合，E-step において観測値とパラメータの暫定値が与えられた下での $\log M_j(q_j;\ \boldsymbol{\theta})$ の条件付期待値を計算し，M-step で $\boldsymbol{\theta}$ に関する $\log M_j(q_j;\ \boldsymbol{\theta})$ の条件付期待値の最大化を行う．この繰り返しで，パラメータの最尤推定量を導出できる．しかし，混合多変量正規分布の場合，$\boldsymbol{\theta}$ で微分可能でないため上記の方法で最尤推定量を導くことはできない．そこで，Yamaguchi (1990) はこのような分布のための一般化 EM アルゴリズムを与えている．

参 考 文 献

Cook, R.D. and Johnson, M.E. (1981). A family of distributions for modelling non-elliptically symmetric multivariate data, *J. Roy. Statist. Soc. Ser. B*, **43**, 210-218.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1980). Iteratively reweighted least squares for linear regression when errors are normal/independent distributed, *Multivariate Analysis* (ed. P.R. Krishnaiah), **5**, 35-57, North-Holland, Amsterdam.

Kano, Y., Berkane, M. and Bentler, P.M. (1990). Covariance structure analysis with heterogeneous kurtosis parameters, *Biometrika*, **77**, 575-585.

Little, R.J.A. (1988). Robust estimation of the mean and covariance matrix from data with missing values, *Applied Statistics*, **37**, 23-38.

Rubin, D.B. (1983). Iteratively reweighted least squares, *Entry in Encyclopedia of the Statistical Sciences*, **4** (eds. S. Kotz, N.L. Johnson and C.B. Read), Wiley, New York.

Yamaguchi, K. (1990). Generalized EM algorithm for models with contaminated normal error terms, *Statistical Methods and Data Analysis* (ed. N. Niki), 107-114, Scientist Inc. Tokyo.