『論　文』

非正規性の下での最尤因子分析法とその効率評価

山口　和範
渡邊　美智子

1. 序

多次元データの構造模型に対する推測の一助として，統計的多変量解析の諸手法の適用はひろく一般化している。しかし，それらの手法の理論的妥当性を保障する数理的仮定と実際データの現実の姿との乖離は，結果の信頼性を左右する適用上の問題点でもある。とくに，統計モデルを規定するパラメータに関しての推定手法が最尤法に基づく場合，連続値型のデータ行列は多次元正規分布に従うことを前提とする手法がほとんどであり，因子分析における従来の最尤法もその例外ではない。一方，単一変量の場合と異なり，現実データが多次元正規モデルでうまく適合された実証例は少なく，とくに，因子分析モデルに代表される一般の潜在構造モデルにあって，概念変数（潜在変数）および誤差変数ともに特定の分布型を課すことは実際的ではない。そのため，潜在基礎分布の誤規定，および，異質集団からの異常値の混入に対して，多次元正規型最尤法よりもより頑健性が期待できる推定法の構築が望まれる。

渡邊・山口（1989）は正規分布を特別な場合として含む多次元 N/I 分布族の下での最尤因子分析法を提唱し，従来の多次元正規性にのみ依存する最尤因子分析の枠を拡張している。本論文の目的は，この拡張された最尤法の実際的有用性を数値評価することである。まず，モンテカルロ法により，データに対する潜在基礎分布の誤規定が結果の推定量に及ぼす影響を MSE の観点から評

価し，種々の分布型の仮定の下で構築される各最尤法の頑健性を吟味する。次に，実データへの適用を通して，モデル選択の拡張性，および，最尤推定値導出の際に副次的に算出される統計量が，正規性に基づく因子分析モデルにおいて各個体の診断統計量として利用できることについて論じる。

2. 多次元 N/I 分布の下での最尤因子分析法

因子分析モデル：

$$\boldsymbol{Y}_j=\boldsymbol{\alpha}+\boldsymbol{\beta}\boldsymbol{Z}_i+\boldsymbol{e}_i,\quad (i=1,\cdots\cdots,\ n)$$

$\boldsymbol{Y}_i$；p次の観測ベクトル

$\boldsymbol{Z}_i$；m次の因子得点ベクトル

$\boldsymbol{\alpha}$；平均ベクトル

$\boldsymbol{\beta}$；$p\times m$の因子負荷行列

$\boldsymbol{e}_i$；p次の特殊因子ベクトル

において，以下を仮定する：

q_i を確率（密度）関数 $M(q_i)$ に従う正の確率変数とし，q_i が与えられた下で，

$$\boldsymbol{Z}_i\sim N(\boldsymbol{0},\ \boldsymbol{I}_m/q_i),\quad \boldsymbol{e}_i\sim N(\boldsymbol{0},\ \boldsymbol{\Psi}/q_i),$$

ここに，$\boldsymbol{e}_i$ と $\boldsymbol{Z}_i$ は独立とする。

すなわち，q_i の条件付きの下で，

$$\begin{pmatrix}\boldsymbol{Y}_i\\ \boldsymbol{Z}_i\end{pmatrix}\sim N\left(\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{0}\end{pmatrix},\ \boldsymbol{\Sigma}/q_i\right)$$

となる。ここに，

$$\boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YZ}\\ \boldsymbol{\Sigma}_{ZY} & \boldsymbol{\Sigma}_{ZZ}\end{pmatrix}=\begin{pmatrix}\boldsymbol{\beta}\boldsymbol{\beta}'+\boldsymbol{\Psi} & \boldsymbol{\beta}\\ \boldsymbol{\beta}' & \boldsymbol{I}_m\end{pmatrix},$$

$$\boldsymbol{\Psi}=\mathrm{diag}\{\psi_1,\ \psi_2,\ \cdots\cdots,\ \psi_p\}$$

である。

ここで，$M(q)$ を具体的に規定することでにより，$\begin{pmatrix}\boldsymbol{Y}_i\\ \boldsymbol{Z}_i\end{pmatrix}$ に対する分布型の

仮定を次のように個別化することができる。

もし

$$M(q)=\begin{cases}1-\delta & \text{if } q=1\\ \delta & \text{if } q=\lambda \qquad (\lambda\ll 1)\\ 0 & \text{otherwise,}\end{cases}$$

とすれば，$\begin{pmatrix}\boldsymbol{Y}_i\\ \boldsymbol{Z}_i\end{pmatrix}$ の分布として，いわゆる変量多コンタミネイト正規分布が仮定されたことになる。この場合の被コンタミネイト分布は多変量正規分布 $N\left(\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{0}\end{pmatrix}, \boldsymbol{\Sigma}\right)$ であり，また，コンタミネイト（汚染）分布としては，分散がもとの分布より拡大された多変量正規分布 $N\left(\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{0}\end{pmatrix}, \boldsymbol{\Sigma}/\lambda\right)$ が相当し，その汚染率は δ である。

次に，$q\nu$ が自由度 ν の χ^2 分布に従うとするとき，$\begin{pmatrix}\boldsymbol{Y}_i\\ \boldsymbol{Z}_i\end{pmatrix}$ の分布としては，自由度 ν の多変量 t 分布 $T\left(\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{0}\end{pmatrix}, \boldsymbol{\Sigma}, \nu\right)$ が仮定されたことになる。

上記のモデルに対し，目的パラメータの MLE は次の E-step と M-step の反復演算により導出される：

E-step：

$$E(q_i|\boldsymbol{Y}_i)=w_i, \quad (w_i \text{ の具体形については後述する。})$$

$$\begin{aligned}E(q_i\boldsymbol{Z}_i|\boldsymbol{Y}_i)&=E\{q_iE(\boldsymbol{Z}_i|q_i,\boldsymbol{Y}_i)|\boldsymbol{Y}_i\}\\ &=w_iE(\boldsymbol{Z}_i|\boldsymbol{Y}_i)\\ &=w_i\hat{\boldsymbol{Z}}_i,\end{aligned}$$

$$\begin{aligned}E(q_i\boldsymbol{Z}_i\boldsymbol{Z}'_i|\boldsymbol{Y}_i)&=E\{q_iE(\boldsymbol{Z}_i\boldsymbol{Z}_i'|q_i,\ \boldsymbol{Y}_i)|\boldsymbol{Y}_i\}\\ &=E\{q_i(\hat{\boldsymbol{Z}}_i\hat{\boldsymbol{Z}}_i'+\mathrm{Cov}(\boldsymbol{Z}_i|q_i,\ \boldsymbol{Y}_i))|\boldsymbol{Y}_i\}\\ &=w_i\hat{\boldsymbol{Z}}_i\hat{\boldsymbol{Z}}_i'+\boldsymbol{\Sigma}^*_{zz},\end{aligned}$$

ここに，

$$\hat{\boldsymbol{Z}}=\boldsymbol{\Sigma}_{ZY}\boldsymbol{\Sigma}^{-1}{}_{YY}(\boldsymbol{Y}_i-\boldsymbol{\alpha}),$$

$$\boldsymbol{\Sigma}^*{}_{ZZ}=\boldsymbol{\Sigma}_{ZZ}-\boldsymbol{\Sigma}_{ZY}\boldsymbol{\Sigma}^{-1}{}_{YY}\boldsymbol{\Sigma}_{YZ}$$

である。

M-step：

$$\hat{\boldsymbol{\alpha}}=\overline{\boldsymbol{Y}}-\hat{\boldsymbol{\beta}}\,\overline{\boldsymbol{Z}},$$

$$\hat{\boldsymbol{\beta}}=\boldsymbol{C}_{YZ}\boldsymbol{C}^{-1}{}_{ZZ},$$

$$\hat{\boldsymbol{\Psi}}=\mathrm{Diag}(\boldsymbol{C}_{YY}-\boldsymbol{C}_{YZ}\boldsymbol{C}^{-1}{}_{ZZ}\boldsymbol{C}_{ZY})w_0/n,$$

ここに，

$$\begin{pmatrix}\boldsymbol{C}_{YY} & \boldsymbol{C}_{YZ}\\ \boldsymbol{C}_{ZY} & \boldsymbol{C}_{ZZ}\end{pmatrix}=\begin{pmatrix}\boldsymbol{S}_{YY}-\overline{\boldsymbol{Y}}\overline{\boldsymbol{Y}}' & \boldsymbol{S}_{YZ}-\overline{\boldsymbol{Y}}\overline{\boldsymbol{Z}}'\\ \boldsymbol{S}_{ZY}-\overline{\boldsymbol{Z}}\overline{\boldsymbol{Y}}' & \boldsymbol{S}_{ZZ}-\overline{\boldsymbol{Z}}\overline{\boldsymbol{Z}}'\end{pmatrix},$$

$$\begin{pmatrix}\boldsymbol{S}_{YY} & \boldsymbol{S}_{YZ}\\ \boldsymbol{S}_{ZY} & \boldsymbol{S}_{ZZ}\end{pmatrix}=\begin{pmatrix}\sum w_i\boldsymbol{Y}_i\boldsymbol{Y}_i'/w_0 & \sum w_i\boldsymbol{Y}_i\hat{\boldsymbol{Z}}_i/w_0\\ \sum w_i\hat{\boldsymbol{Z}}_i\boldsymbol{Y}_i'/w_0 & \sum\{w_i\hat{\boldsymbol{Z}}_i\hat{\boldsymbol{Z}}_i'+\boldsymbol{\Sigma}^*{}_{ZZ}\}/w_0\end{pmatrix},$$

$$\overline{\boldsymbol{Y}}=\sum w_i\boldsymbol{Y}_i/w_0,$$

$$\overline{\boldsymbol{Z}}=\sum w_i\hat{\boldsymbol{Z}}_i/w_0,$$

$$w_0=\sum w_i$$

である。

[w_i の定式化]

多変量コンタミネイト正規モデルの場合，q_i の条件付き期待値 w_i は，

$$w_i=\frac{1-\delta+\delta\lambda^{1+p/2}\exp\{(1-\lambda)d_i^2/2\}}{1-\delta+\delta\lambda^{p/2}\exp\{(1-\lambda)d_i^2/2\}}$$

で与えられる。ここに，

$$d_i^2=(\boldsymbol{Y}_i-\boldsymbol{\alpha})'\boldsymbol{\Sigma}^{-1}{}_{YY}(\boldsymbol{Y}_i-\boldsymbol{\alpha})$$

である。

また，多変量 t モデルの場合は，

$$w_i=\frac{\nu+p}{\nu+d_i^2}$$

となる。

3. 頑健性の評価

本節において，応答データ $\boldsymbol{Y}_i$ および潜在因子得点 $\boldsymbol{Z}_i$ に，多変量正規分布モデル，多変量 t 分布モデル，多変量コンタミネイト正規分布モデルを仮定した場合に得られる各最尤推定量の効率比較を行なう。とくに，潜在基礎分布に対して誤規定が生じた場合の影響も含めて，シミュレーションにより頑健性を数値評価する。

3.1 シミュレーションの計画

人工データ生成の基礎となる因子分析モデルとして，Ihara & Okamoto (1985) により使用された数値モデルを採用した。ここに，応答データの次数 p は 9，共通因子数 m は 3 で，平均ベクトル，因子負荷行列および残差分散行列に関して次の数値を設定した。

$$\boldsymbol{\alpha}=\boldsymbol{0}$$

$$\boldsymbol{\beta}'=\begin{bmatrix} 0.7 & 0.7 & 0.5 & 0.8 & 0.7 & 0.8 & 0.7 & 0.4 & 0.8 \\ 0.3 & 0.2 & 0.3 & -0.3 & -0.3 & -0.4 & 0.4 & 0.3 & 0.4 \\ 0.1 & -0.2 & -0.2 & -0.1 & -0.2 & 0.1 & -0.1 & 0.5 & 0.0 \end{bmatrix}$$

$$\boldsymbol{\Psi}=\mathrm{diag}(0.41,\ 0.43,\ 0.62,\ 0.26,\ 0.38,\ 0.19,\ 0.34,\ 0.50,\ 0.20)$$

この数値モデルは，Emmett's (1949) のデータに対して Lawley and Maxwell (1971) により導かれた最尤因子解に基づいている。

また，人工データ生成に際して，因子得点 $\boldsymbol{Z}_i$ および誤差項 $\boldsymbol{e}_i$ に対する分布として次の 4 種類を採用した。

1) 多変量正規分布

2) 自由度10の多変量 t 分布

3) 自由度 4 の多変量 t 分布

4)　多変量コンタミネイト正規分布：

$$0.9\cdot N\left(\begin{pmatrix}\alpha\\ 0\end{pmatrix},\ \Sigma\right)+0.1\cdot N\left(\begin{pmatrix}\alpha\\ 0\end{pmatrix},\ \Sigma/0.0767\right)$$

上述のモデルにより発生された潜在基礎分布の仮定の異なる4種類の人工データ各々に対して，次の4種類の MLE を計算した。

a)　多変量正規分布を仮定した下での MLE（正規型 MLE）

b)　自由度10の多変量 t 分布を仮定した下での MLE（T10型 MLE）

c)　自由度4の多変量 t 分布を仮定した下での MLE（T4型 MLE）

d)　多変量コンタミネイト正規分布を仮定した下での MLE（コンタミネイト型 MLE）

ここで，一回の実験において，合計16種の推定値が因子負荷行列 β と残差分散 Ψ に関して算出されることになる。これら各推定値の精度に対し，次の評価基準を計算した。

因子負荷行列 β に関する平均自乗誤差の平方根；

$$\{\sum_{i=1}^{p}\sum_{j=1}^{m}(\hat{\beta}_{ij}-\beta_{ij})^2/pm\}^{1/2}$$

残差分散 Ψ に関する平均自乗誤差の平方根；

$$\{\sum_{i=1}^{p}(\hat{\psi}_i-\psi_i)^2/p\}^{1/2}$$

サンプルサイズ n を 50, 100, 200 と変化させて，シミュレーションを行った。シミュレーションサイズは200である。

3.2　結果と考察

表3.1から3.8にかけて，因子負荷行列および特殊因子の分散に関する推定量の RMSE を与えている。行方向に，人工データ発生に際して使用した乱数型を，列方向に，最尤法構築に際し仮定した分布型を配している。従って，表中の対角セルにおいては，潜在基礎分布型に関する仮定が正しい下での各最尤法の効率が示されている。一方，非対角セルにおいては，潜在基礎分布に関する仮定を誤った場合の効率が示されている。また，（　）内の数値は，各行に

表 3.1：多変量正規分布及びコンタミネイト多変量正規分布に対する Root Mean Squared Error (×1000)

因子負荷行列		
$n=50$		
	最尤法構築の際の仮定分布	
発生データ	Normal	Contam.
Normal	133(100)	134(101)
Contam.	346(257)	134(100)
Mean	(179)	(101)
$n=100$		
	最尤法構築の際の仮定分布	
発生データ	Normal	Contam.
Normal	97(100)	98(101)
Contam.	313(333)	94(100)
Mean	(217)	(101)
$n=200$		
	最尤法構築の際の仮定分布	
発生データ	Normal	Contam.
Normal	66(100)	67(101)
Contam.	288(443)	65(100)
Mean	(272)	(101)

表 3.2：多変量正規分布及びコンタミネイト多変量正規分布に対する Root Mean Squared Error (×1000)

特殊因子の分数		
$n=50$		
	最尤法構築の際の仮定分布	
発生データ	Normal	Contam.
Normal	154(100)	154(100)
Contam.	335(224)	150(100)
Mean	(162)	(100)
$n=100$		
	最尤法構築の際の仮定分布	
発生データ	Normal	Contam.
Normal	110(100)	110(100)
Contam.	435(406)	107(100)
Mean	(253)	(100)
$n=200$		
	最尤法構築の際の仮定分布	
発生データ	Normal	Contam.
Normal	76(100)	76(100)
Contam.	465(612)	76(100)
Mean	(366)	(100)

おいて，対角セルの数値を 100 とした場合の相効比率を表わしている。最終行 (Mean) の（ ）内の数値は，これら相効比率の各列にわたる平均を示している。相対効率は，その値が小さい程，分布型の誤規定に対しての頑健性が優れていることを表現している。

表 3.1 は，とくに，多変量正規分布とコンタミネイト多変量正規分布を比較

した場合の因子負荷行列の推定に関する RMSE 表である。多変量正規モデルに従う人工データに対して，正規型 MLE の効率とコンタミネイト型 MLE の効率はほぼ一致する。一方，コンタミネイト多変量正規モデルに従う人工デ

表 3.3：多変量正規分布及び多変量 t 分布に対する Root Mean Squared Error（×1000）

因子負荷行列

$n=50$

	最尤法構築の際の仮定分布		
発生データ	Normal	T(df 10)	T(df 4)
Normal	133(100)	135(101)	136(102)
T(df 10)	173(122)	142(100)	137(96)
T(df 4)	303(206)	167(114)	147(100)
Mean	(143)	(105)	(99)

$n=100$

	最尤法構築の際の仮定分布		
発生データ	Normal	T(df 10)	T(df 4)
Normal	97(100)	99(102)	102(105)
T(df 10)	131(130)	101(100)	100(99)
T(df 4)	272(264)	120(117)	103(100)
Mean	(165)	(106)	(101)

$n=200$

	最尤法構築の際の仮定分布		
発生データ	Normal	T(df 10)	T(df 4)
Normal	66(100)	68(103)	73(111)
T(df 10)	95(140)	68(100)	69(101)
T(df 4)	239(336)	85(120)	71(100)
Mean	(192)	(108)	(104)

表 3.4：多変量正規分布及び多変量 t 分布に対する Root Mean Squared Error（×1000）

特殊因子の分散

$n=50$

	最尤法構築の際の仮定分布		
発生データ	Normal	T(df 10)	T(df 4)
Normal	154(100)	160(105)	167(109)
T(df 10)	189(115)	164(100)	167(102)
T(df 4)	334(207)	184(114)	161(100)
Mean	(141)	(106)	(104)

$n=100$

	最尤法構築の際の仮定分布		
発生データ	Normal	T(df 10)	T(df 4)
Normal	110(100)	118(107)	126(114)
T(df 10)	158(130)	122(100)	124(102)
T(df 4)	362(297)	146(120)	122(100)
Mean	(176)	(109)	(105)

$n=200$

	最尤法構築の際の仮定分布		
発生データ	Normal	T(df 10)	T(df 4)
Normal	76(100)	88(116)	100(132)
T(df 10)	132(163)	81(100)	87(107)
T(df 4)	371(452)	107(130)	82(100)
Mean	(238)	(115)	(113)

ータに対しては，同じコンタミネイト型 MLE の効率が正規型 MLE の効率を大きく上回っている。とくに，サンプル数が，50，100，200と増えるに従いその効率比は約2.5倍，3倍，4倍と大きくなる。つまり，潜在基礎分布が正規分布からコンタミネイト正規分布にずれた場合，正規型 MLE の頑健性が損なわれるのに対し，コンタミネイト型 MLE は，この場合の分布のずれに対して非常に頑健であることがわかる。この傾向は，特殊因子の分散の推定に関しても同様であるが，サンプルサイズ増加に伴う2つの MLE の頑健性の差異の増加率は，因子負荷量の推定の場合よりさらに大きくなる（表3.2参照）。

表3.3は，多変量正規分布と多変量 t 分布に関する因子負荷行列推定の場合の RMSE 表である。多変量正規モデルに従う人工データに対して，正規型 MLE の効率と自由度4および自由度10の双方の多変量T型 MLE の効率とはほとんど差がないといってよい。ただし，サンプルサイズが200と大きくなると，自由度4の多変量T型 MLE の効率は他の2つに比べて，やや下回る傾向にある。一方，自由度10の多変量T分布モデルに従う人工データに対しては，自由度10および自由度4の多変量T型 MLE の効率はいずれも高いものの，正規型 MLE の効率はこれらに比べ著しく劣る。また，自由度4の多変量 t 分布モデルに従う人工データに対して，自由度10の多変量T型 MLE の効率は低くなるが，正規型 MLE の効率は更にそれを下回る。平均的にみても，潜在基礎分布の仮定のずれに対する各 MLE の頑健性は，自由度4の多変量T型が最も強く，次いで，自由度10の多変量T型，最後に多変量正規型となり，正規型MLE の頑健性が最も悪い。つまり，最尤法構築に際し仮定される分布の裾の重さ（尖度）に比例して，結果として得られる最尤推定量の頑健性が増していることになる。またこれら一連の傾向は，サンプル数が大きくなるに従い強調される。この傾向は，特殊因子の分散に関する推定の場合も同様であるが，サンプルサイズ増加に伴う各 MLE の頑健性の差異は，因子負荷量推定の場合よりさらに大きくなる（表3.4参照）。

表3.5および表3.6は，コンタミネイト多変量正規分布と多変量 t 分布を比

較した場合の RMSE 表である。前述の結果と同様，対角セルを境に，上三角セルに相当する RMSE の値は下三角セルの RMSE より比較的小さな値をとっている。つまり，発生データの分布よりも裾の長い分布を仮定した最尤推定

表 3.5：コンタミネイト多変量正規分布及び多変量 t 分布に対する Root Mean Squared Error（×1000）

因子負荷行列

$n=50$

	最尤法構築の際の仮定分布		
発生データ	Contam.	T(df 10)	T(df 4)
Contam.	134(100)	157(117)	144(107)
T(df 10)	156(110)	142(100)	137(96)
T(df 4)	168(114)	167(114)	147(100)
Mean	(108)	(110)	(101)

$n=100$

	最尤法構築の際の仮定分布		
発生データ	Contam.	T(df 10)	T(df 4)
Contam.	94(100)	110(117)	101(107)
T(df 10)	114(113)	101(100)	100(99)
T(df 4)	129(125)	120(117)	103(100)
Mean	(113)	(111)	(102)

$n=200$

	最尤法構築の際の仮定分布		
発生データ	Contam.	T(df 10)	T(df 4)
Contam.	65(100)	78(123)	69(106)
T(df 10)	79(116)	68(100)	69(101)
T(df 4)	95(134)	85(120)	71(100)
Mean	(117)	(114)	(102)

表 3.6：コンタミネイト多変量正規分布及び多変量 t 分布に対する Root Mean Squared Error（×1000）

特殊因子の分散

$n=50$

	最尤法構築の際の仮定分布		
発生データ	Contam.	T(df 10)	T(df 4)
Contam.	150(100)	161(107)	159(106)
T(df 10)	172(105)	164(100)	167(102)
T(df 4)	183(114)	184(114)	161(100)
Mean	(106)	(107)	(103)

$n=100$

	最尤法構築の際の仮定分布		
発生データ	Contam.	T(df 10)	T(df 4)
Contam.	107(100)	125(117)	118(110)
T(df 10)	136(111)	122(100)	124(102)
T(df 4)	153(125)	146(120)	122(100)
Mean	(112)	(112)	(104)

$n=200$

	最尤法構築の際の仮定分布		
発生データ	Contam.	T(df 10)	T(df 4)
Contam.	76(100)	92(121)	81(107)
T(df 10)	100(123)	81(100)	87(107)
T(df 4)	120(146)	107(130)	82(100)
Mean	(123)	(117)	(105)

表 3.7：多変量尖度

分布	Normal	T(df 10)	Contam.	T(df 4)
多変量尖度	99	128	383	∞

表 3.8：Root Mean Squared Error (×1000)

因子負荷行列

$n=50$

	最尤法構築の際の仮定分布			
発生データ	Normal	T(df 10)	Contam.	T(df 4)
Normal	133(100)	135(101)	134(101)	136(102)
T(df 10)	173(122)	142(100)	156(110)	137(96)
Contam.	346(257)	157(117)	134(100)	144(107)
T(df 4)	303(206)	167(114)	163(114)	147(100)
Mean	(171)	(108)	(106)	(101)

$n=100$

	最尤法構築の際の仮定分布			
発生データ	Normal	T(df 10)	Contam.	T(df 4)
Normal	97(100)	99(102)	98(101)	102(105)
T(df 10)	131(130)	101(100)	114(113)	100(99)
Contam.	313(333)	110(117)	94(100)	101(107)
T(df 4)	272(264)	120(117)	129(125)	103(100)
Mean	(207)	(109)	(110)	(102)

$n=200$

	最尤法構築の際の仮定分布			
発生データ	Normal	T(bf 10)	Contam.	T(df 4)
Normal	66(100)	68(103)	67(101)	73(111)
T(df 10)	95(140)	68(100)	79(116)	69(101)
Contam.	288(443)	78(123)	65(100)	69(106)
T(df 4)	239(336)	85(120)	95(134)	71(100)
Mean	(255)	(112)	(113)	(105)

量の方が裾の短い分布を仮定した場合の最尤推定量よりも，潜在基礎分布の誤規定に伴う頑健性への影響が，より小さいことがわかる。ただし，多変量正規分布を含めた前例と異なり，コンタミネイト分布と多変量 t 分布の間の取り違

表 3.9：Root Mean Squared Error（×1000）

特殊因子の分散

$n=50$

	最尤法構築の際の仮定分布			
発生データ	Normal	T(df 10)	Contam.	T(df 4)
Normal	154(100)	160(105)	154(100)	167(109)
T(df 10)	189(115)	164(100)	172(105)	167(102)
Contam.	335(224)	161(107)	150(100)	159(106)
T(df 4)	334(207)	184(114)	183(114)	161(100)
Mean	(162)	(107)	(105)	(104)

$n=100$

	最尤法構築の際の仮定分布			
発生データ	Normal	T(df 10)	Contam.	T(df 4)
Normal	110(100)	118(107)	110(100)	126(114)
T(df 10)	158(130)	122(100)	136(111)	124(102)
Contam.	435(406)	125(117)	107(100)	118(110)
T(df 4)	362(297)	146(120)	153(125)	122(100)
Mean	(233)	(111)	(109)	(107)

$n=200$

	最尤法構築の際の仮定分布			
発生データ	Normal	T(df 10)	Contam.	T(df 4)
Normal	76(100)	88(116)	76(100)	100(132)
T(df 10)	132(163)	81(100)	100(123)	87(107)
Contam.	465(612)	92(121)	76(100)	81(107)
T(df 4)	371(452)	107(130)	120(146)	82(100)
Mean	(332)	(117)	(117)	(112)

えはさほど大きくない。表3.7にこれら4分布の多変量尖度（Mardia (1970)）を与えている。

表3.8および表3.9は，前出の多変量正規分布，コンタミネイト多変量正規分布および多変量 t 分布に関する表をすべてまとめた RMSE 表である。これら4つの分布は，その尖度の小さい順に配置されている。総合的にみて，正規分布の下での MLE が他のより裾の長い分布の仮定の下での MLE に比べて，頑健性が劣ることがわかる。また，多変量正規分布に関しては，多変量 t 分布との取り違えよりもコンタミネイ分布との取り違えのほうが，正規型 MLE にあたえる影響の大きいことが注目される。

4. 実データへの適用

本節ではとくに，多変量解析の諸手法の例題として頻に利用されている the open/closed book data (Mardia *et al.*, 1979, Table 1.2.1) にここで提唱している因子分析法を適用して，本手法の実データ解析上での利点を論じる。

表4.1に，ここで用いた the open/closed book data のリストを与えている。これは，88人の被験者に5種類の試験（Mechanics, Vectors, Algebra, Analysis, Statistics）を受験させた結果得られた得点データである。行方向は同一の被験者で対応付けられている。5種類の試験において，Algebra, Analysis および Statistics は，テキストの参照が許された下での試験（open book examination）である。一方，Machanics と Vectors は，テキストを閉じた状態で行う試験（closed book examination）である。Mardia *et al.* (1979) には，Jöreskog (1979) により導出されたアルゴリズムを用いた下での多変量正規性の仮定の下での最尤因子分析解が計算されている。ここで，因子数に関して，2因子モデルが最もデータに適合することが結論付けられ，また，それらの抽出された2因子の解釈は，第1因子が総合能力を示し，第2因子が，open book 形式の試験能力に対して，closed book 形式の試験の能力を強調する因子を示すとされている。以下は2因子モデルの下で議論をすすめる。

4.1 多変量t分布の下での最尤因子分析モデル

表4.2は，自由度を8から14まで変化させたときの多変量t分布に基づく最尤因子分析モデル及び多変量正規分布に基づく最尤因子分析モデルを各々適合させたときの対数尤度を与えている。その中で最大対数尤度を示した自由度11

Table 4.1 Marks in open-book and closed-book examination out 100†

Mechanics(C)	Vectors(C)	Algebra(O)	Analysis(O)	Statistics(O)
77	82	67	67	81
63	78	80	70	81
75	73	71	66	81
55	72	63	70	68
63	63	65	70	63
53	31	72	64	73
51	67	65	65	68
59	70	68	62	56
62	60	58	62	70
64	72	60	62	45
52	64	60	63	54
55	67	59	62	44
50	50	64	55	63
65	63	58	56	37
31	55	60	57	73
60	64	56	54	40
44	69	53	53	53
42	69	61	55	45
62	46	61	57	45
31	49	62	63	62
44	61	52	62	46
49	41	61	49	64
12	58	61	63	67
49	53	49	62	47
54	49	56	47	53
54	53	46	59	44
44	56	55	61	36
18	44	50	57	81
46	52	65	50	35
32	45	49	57	64
30	69	50	52	45
46	49	53	59	37
40	27	54	61	61
31	42	48	54	68
36	59	51	45	51
56	40	56	54	35
46	56	57	49	32
45	42	55	56	40
42	60	54	49	33
40	63	53	54	25
23	55	59	53	44
48	48	49	51	37
41	63	49	46	34
46	52	53	41	40

†O indicates open-book, C indicates closed book.

Table 4.1 (Continued)

Mechanics(C)	Vectors(C)	Algebra(O)	Analysis(O)	Statistics(O)
46	61	46	38	41
40	57	51	52	31
49	49	45	48	39
22	58	53	56	41
35	60	47	54	33
48	56	49	42	32
31	57	50	54	34
17	53	57	43	51
49	57	47	39	26
59	50	47	15	46
37	56	49	28	45
40	43	48	21	61
35	35	41	51	50
38	44	54	47	24
43	43	38	34	49
39	46	46	32	43
62	44	36	22	42
48	38	41	44	33
34	42	50	47	29
18	51	40	56	30
35	36	46	48	29
59	53	37	22	19
41	41	43	30	33
31	52	37	27	40
17	51	52	35	31
34	30	50	47	36
46	40	42	29	17
10	46	36	47	39
46	37	45	15	30
30	34	43	46	18
13	51	50	25	31
49	50	38	23	9
18	32	31	45	40
8	42	48	26	40
23	38	36	48	15
30	24	43	33	25
3	9	51	47	40
7	51	43	17	22
15	40	43	23	18
15	38	39	28	17
5	30	44	36	18
12	30	32	35	21
5	26	15	20	20
0	40	21	9	14

の多変量 t 分布の下での最尤因子解及び多変量正規分布の下での最尤因子解を表 4.3 に示している。自由度11の多変量 t 分布モデルでの対数尤度は，多変量正規モデルの対数尤度より，約13程大きくなっている。仮に，多変量 t 分布における自由度が未知であるとして，自由度に関するパラメータの増量を考慮し

ても，この尤度の改善は統計的に有意である。つまり，従来の多変量正規因子分析モデルに比べて，多変量 t 分布の仮定の下での因子分析モデルの方がよりデータとの適合性が高いことが結論付けられる。

4.2 外れ値に関する考察

前節に於て，対数尤度の観点から，the open/closed data に対しては，多変量 t 分布による因子分析モデルが従来の多変量正規分布による因子分析モデルよりも，より適合性が高いことが示されたが，このことから，潜在因子得点及び誤差項の母集団分布モデルとして，多変量正規分布よりも多変量 t 分布の方が好ましいと一概に解釈するのは早急である。むしろ，現データセット中に何等かの観点から他の大勢を占めるデータから外れる観測個体が存在することの可能性の示唆ととらえる方が自然であろう。この場合，データセット中からその外れ値を検出するための方法論が必要となるが，第2節中の E-step で得られる w_i の収束値は外れ値検出のための有効な統計量として利用できる。

w_i は，各個体の観測値 $\boldsymbol{Y}_i$ が与えられた下での q_i の条件付期待値 $\mathrm{E}(q_i|\boldsymbol{Y}_i)$

表 4.2　多変量 t 分布の各自由度に対する対数尤度

自由度	8	10	11	12	14	∞(Normal)
対数尤度	-1278.13	-1277.63	-1277.61	-1277.69	-1278.00	1290.77

表 4.3　最尤因子分析解

Normal			T(df 11)		
因子負荷行列		特殊因子の分散	因子負荷行列		特殊因子の分散
.628	.373	.466	.627	.315	.507
.695	.312	.419	.703	.298	.418
.899	-.050	.189	.897	-.041	.194
.780	-.201	.352	.780	-.157	.368
.727	-.200	.431	.727	-.217	.425
対数尤度		-1290.77			-1277.61

の推定値である。この場合，結果的に得られる因子解が多変量 t 分布の下での最尤推定値となるために，q_i は自由度11の χ^2 分布に従うことが仮定されている。もしq_iの従う分布が常に1の値しかとらない一点分布とすれば，第2節の推定法はそのまま多変量正規分布の下での最尤因子解を導く方法となる。つまり，各個体に関して w_i の値が1にちかいほど，その個体は多変量正規性の下での因子分析モデルと適合するデータと解釈できる。逆に，1より小さな値をとるほど多変量正規型因子分析モデルから外れたデータ，いわゆる外れ値の危険性があると解釈できる。このことは，本稿で提唱されている推定アルゴリズムが，もし，各個体の因子得点が観測された下では，反復重み付最小自乗（Iteratively Reweighted Least Squares）アルゴリズムと同等とみなせること，またその際，w_j は各個体に課せられた重み（weight）として作用することからも容易に理解できよう。

表4.4は，the open/closed book data に対して得られた w_i の収束値である。行方向に向かって個体番号の若い順に，合計88個の個体の各 w_i の値を配置している。表から，0.5以下の w_i 値を示しているのは81番目の個体だけである。81番目の個体のオリジナルデータをみてみると，closed book 形式の Mechanics と Vectors の得点が非常に低いのに対して（この2科目での合計点では88個体の中で最低），他の open book 形式の3科目の得点は平均的で

表 4.4　自由度11の多変量t分布の下での88個体に対する w_i 値

	1	2	3	4	5	6	7	8	9	0
	0.75	0.76	0.84	1.04	1.11	0.95	1.14	1.10	1.03	1.06
10	1.28	1.18	1.07	1.08	0.96	1.19	1.12	1.12	0.98	1.03
20	1.19	0.90	0.73	1.11	1.20	1.02	1.20	0.64	0.91	0.99
30	0.97	1.22	0.72	0.89	1.27	0.95	1.21	1.16	1.27	1.03
40	1.05	1.31	1.21	1.33	1.14	1.27	1.24	1.08	1.12	1.30
50	1.20	0.94	1.17	0.62	1.01	0.67	0.94	1.07	0.92	1.20
60	0.64	1.06	1.25	0.86	1.16	0.75	1.18	0.95	0.98	1.01
70	0.95	0.85	0.75	1.03	0.81	0.84	0.78	0.83	0.90	0.94
80	0.49	0.72	0.96	1.09	0.86	0.98	0.59	0.63		

表 4.5　因子得点（多変量正規モデル）

個体番号	因子得点		個体番号	因子得点	
	第1因子	第2因子		第1因子	第2因子
1	2.1804	.2738	45	.0177	.8173
2	2.4676	-.5590	46	.1348	.2641
3	2.1319	-.1051	47	-.1298	.3510
4	1.4912	-.3023	48	.0961	-.4567
5	1.4733	-.3330	49	-.0018	.2415
6	1.5865	-.8190	50	.0517	.7047
7	1.3753	-.4669	51	.0141	-.0295
8	1.5566	.0217	52	.0311	-.6658
9	1.0896	-.2013	53	-.0605	.9820
10	1.2750	.5768	54	-.1712	1.2701
11	1.0282	-.1228	55	-.1472	.5642
12	1.0154	.2561	56	-.3671	.1812
13	.8716	-.6280	57	-.6463	-.5606
14	.9307	.6721	58	-.1104	.0116
15	.6380	-1.0658	59	-.6755	.3726
16	.8036	.6263	60	-.4063	.3277
17	.6388	.2598	61	-.6433	1.3221
18	.9027	.1110	62	-.5873	.3035
19	.7431	-.1263	63	-.3306	-.1341
20	.5937	-1.2472	64	-.6753	-.2615
21	.4891	-.0519	65	-.5825	-.2150
22	.5153	-.7114	66	-.6087	1.8822
23	.5073	-1.5042	67	-.6748	.5220
24	.2876	-.1107	68	-.8209	.6969
25	.4494	-.0031	69	-.4031	-0.653
26	.1869	.2057	70	-.5183	-.6027
27	.4433	-.0716	71	-.5746	.8179
28	-.0939	-1.5993	72	-1.0668	-.4362
29	.6813	-.0857	73	-.7667	.9016
30	-.0336	-.9190	74	-.8913	-.0845
31	.2781	.1560	75	-.6303	.1333
32	.2426	-.1474	76	-.8080	1.6780
33	-.0541	-1.4151	77	-1.4427	-.5006
34	-.1485	-1.0009	78	-.8844	-.3802
35	.1746	.0566	79	-1.1847	.0303
36	.2614	-.0812	80	-1.1642	-.1731
37	.4112	.2950	81	-1.2628	-2.0494
38	.1648	-.3990	82	-1.1184	.5063
39	.3217	.3686	83	-1.1914	.2989
40	.3125	.4446	84	-1.3478	.2284
41	.2807	-.6383	85	-1.3506	-.5771
42	.0140	.1638	86	-1.7291	-.1531
43	.1441	.6066	87	-2.7243	.3513
44	.1471	.3339	88	-2.4228	.8718

ある。このことは，抽出された2因子に関する因子得点をみれば明らかである（表4.5参照）。81番目の個体の第2因子に関する因子得点は −2.05 で，全体的にみても極めて低い値を示している。

表4.6は，81番目めの個体のデータを積極的に除外した下での多変量正規分布を仮定した最尤因子解である。表4.3の2組の最尤因子解と比較すると，その値がやや多変量 t 分布の下での因子解にちかいことが見受けられる。

外れ値検出に対する w_i 値の利用の有効性をさらに明らかにするために，76

表 4.6　多変量正規分布の下での最尤因子解
（81番目の個体を取り除いた87個のデータに対して）

因子負荷行列		特殊因子の分散
.618	.295	.530
.717	.265	.416
.910	-.006	.172
.780	-.184	.357
.728	-.212	.425

表 4.7　多変量t分布の各自由度に対する対数尤度
（タイプミスを含むデータセットに対して）

自由度	5	9	10	11	12	(Normal)
対数尤度	-1285.8	-1282.2	-1282.1	-1282.2	-1282.4	-1300.6

表 4.8　最尤因子分析解（タイプミスを含むデータセットに対して）

Normal			T(df 10)		
因子負荷行列		特殊因子の分散	因子負荷行列		特殊因子の分散
.666	.446	.357	.645	.276	.507
.698	.194	.474	.712	.274	.418
.901	-.117	.174	.897	-.069	.190
.761	-.221	.372	.769	-.159	.383
.646	-.059	.579	.681	-.131	.519
対数尤度		-1300.6			-1282.1

表 4.9 Conditional Expectaions of q(Multivariate t(df 10) model)
（タイプミスを含むデータセットに対して）

	1	2	3	4	5	6	7	8	9	0
	0.75	0.74	0.84	1.04	1.11	0.93	1.14	1.11	1.04	1.05
10	1.30	1.19	1.07	1.08	0.95	1.19	1.13	1.13	0.98	1.00
20	1.20	0.89	0.70	1.12	1.23	1.02	1.23	0.62	0.91	0.99
30	0.97	1.25	0.69	0.89	1.30	0.96	1.23	1.18	1.28	1.04
40	1.05	1.33	1.21	1.35	1.15	1.29	1.25	1.08	1.13	1.30
50	1.22	0.92	1.15	0.61	1.02	0.69	0.94	1.10	0.95	1.23
60	0.64	1.05	1.28	0.85	1.18	0.70	1.19	0.97	0.97	1.01
70	0.92	0.84	0.73	1.04	0.79	0.35	0.77	0.82	0.90	0.93
80	0.46	0.70	0.95	1.08	0.86	0.98	0.58	0.61		

番目の個体の Statistics の得点 9 を90にタイプミスしたと仮定した下で同様の解析を行なってみた。表 4.7 は，自由度を 5 から12まで変化させたときの多変量 t 分布に基づく最尤因子分析モデル及び多変量正規分布に基づく最尤因子分析モデルを各々適合させたときの対数尤度を与えている。ここでも，自由度10の多変量 t 分布を仮定した際の対数尤度は多変量正規分布を仮定したときの対数尤度よりも約18程高くなっている。最大対数尤度を示した自由度10の多変量 t 分布の下での最尤因子解及び多変量正規分布の下での最尤因子解を表 4.8 に与えている。タイプミスのデータの影響は，多変量正規分布の下での最尤因子解における第 2 因子の Mechanics, Algebra, Statistics の各々の因子負荷の変化にあらわれている。一方，自由度10の多変量 t 分布モデルの下での最尤因子解は，タイプミスの含まれないオリジナルのデータを使用したときの因子解とほぼ同様の値を示している。これは，多変量正規分布の下での最尤因子分析法よりも多変量 t 分布を仮定する最尤因子分析法の方が異常値の混入に対して頑健性が高いことを示す一つの実例といえよう。表 4.9 は，この場合の w_i の値の表である。76 番目の個体の w_i の値は 0.35 と一番小さく，この個体の異常性を示唆している。

5. 結　語

本稿において，因子得点および誤差項に多変量正規分布を仮定した従来の最尤因子分析法と比較して，多変量t分布または多変量コンタミネイト分布を仮定した下での最尤因子分析法が，潜在基礎分布の誤規定に関してより頑健であることをシミュレーション実験により検証した。ここで，尖度の大きい分布を仮定した際の最尤因子分析法は，仮にデータの潜在基礎分布が仮定と反しても，その尖度が仮定分布の尖度よりも小さければ推定効率はさほど低下しないことがわかった。逆に，データが従う潜在基礎分布の尖度が最尤法の仮定分布の尖度より大きければ，つまり，異常値もしくは外れ値の混入が考えられる場合は，最尤推定の効率は著しく低下することが示された。このことは，多変量t分布及び多変量コンタミネイト正規分布に比べてその尖度の小さい多変量正規分布の下での最尤因子分析法が，分布の仮定の崩れに対して頑健性を持たないことを示している。一方，できるだけ尖度の大きい分布を仮定した最尤因子分析法は分布の仮定の崩れに対して，推定効率が変化しないという意味でロバストな推定法と結論できる。

本稿では更に，多変量t分布の下での最尤因子分析法の有用性を実データへの適用を通して確認した。この際，異常値の検出法に対しても議論を行った。

ここでは，多変量t分布の場合は自由度を，また，多変量コンタミネイト正規分布に関しては汚染率及び分散撹乱パラメータをともに既知とした下での最尤推定法を取り扱ったが，これらのパラメータは実データの解析においては未知であると考える方が自然である。従って，これらのパラメータも同時に推定できる手法の開発は今後の研究課題として重要であろう。

参 考 文 献

[1] Emmett, W. G. (1949), *Factor Analysis by Lawley's Method of Maximum Likelihood,* Brit. J. Psychol. Statist. 290-97.

[2] Ihara, M. and Okamoto, M. (1985), *Experimental Comparison of Least Squares and Maximum Likelihood Methods in Factor Analysis,* Statistics & Probability Letters 3 287-293.

[3] Jöreskog, K. G. (1977), *Factor Analysis by Least Squares and Maximum Likelihood Methods,* Statistical Methods for Digital Computers, Vol. 3 (Wiley) 125-153.

[4] Lawley, D. N. and A. E. Maxwell (1971), *Factor Analysis as a Statistical Method* (Butterworth, London, 2nd ed.)

[5] Mardia, K. V. (1970), *Measures of Multivariate Skewness and Kurtosis with Applications,* Biometrika, 57, 519-530.

[6] Mardia, K. V., Kent, J. T. and Bibby, M. (1979), *Multivariate Analysis,* Academic Press.

[7] 渡邊美智子，山口和範（1989）「非正規性の下での最尤法」關西大學經濟論集第39巻第1号，101-113.