統計数理 第35巻 第1号（1987）

潜在クラス分析における不完全データからの最尤推定値の導出法とその評価

九州大学理学部附属基礎情報学研究施設　渡　辺　美智子

（1987年2月　受付）

要　約

潜在クラスモデルを一種の多項分布モデルととらえ，潜在確率の最尤推定値を得るための一つの反復推定法を与える．とくに，この推定法は不完全な応答データがある場合にも適用可能である．また，推定値の漸近分散の評価法を与え，不完全な応答データを利用することの漸近分散の観点からの有効性を数値例で示す．

さらに，漸近分散による評価が妥当でない小標本に対して，最尤推定値の平均自乗誤差をシミュレーションにより評価し，不完全な応答データを利用することの有効性が変わらないことを示す．

1. 序

潜在クラス分析は，予め用意された k 個の質問項目に対する被験対象集団の応答データを用いて，集団を構成する T 個の異質な潜在クラスを規定することを目的とした分析手法である．具体的に，集団内での各潜在クラスの構成率及び k 個の質問項目各々に対してクラス内応答確率の推定値を与える．この分析法は，Lazarsfeld (1950)により最初に提唱され，以後，潜在確率の推定手法とその評価をめぐって多くの研究がなされてきた．例えば，潜在確率と顕在確率間の構造方程式系を代数的に解く観点から，Anderson (1954), Green (1951), Gibson (1955) 等が，各々独自の推定法を提唱している．しかし，これらの推定法では不適解の発生が問題となることが，Okamoto and Isogai (1978), Morita (1980) 等によって報告されている．最尤法に関しても，Newton-Raphson 法にもとづく反復法が，Lazarsfeld and Henry (1968) によって与えられているが，やはり，不適解の発生が避けられないことが，Morita (1984), Sato, Sugawa and Kawaguchi (1980)に示されている．最尤法における不適解対策として，Formann (1978)は，潜在確率の変換法を提案した．また，Goodman (1979)は，別の観点から，最尤推定値を適解の範囲で導く一つの反復推定法を与え，適当な条件の下で，推定量が漸近一致性，漸近有効性，漸近正規性をもつことを示している．

他方，潜在確率の推定値の不定性(Instability)は，推定手法の相違に加えて，潜在クラスモデルの構造自身にも大きく起因することが，Sato, Sugawa and Kawaguchi (1980) によって示されている．従って，Goodman の示した最尤推定量の漸近的性質も，どの程度に実用性をもつか疑問である．

本稿の主要な目的は，Goodman の推定法を，顕在変量に対し不完全な応答を示した個人の頻度にも利用できるように拡張して，その結果得られる推定量の漸近分散の評価法を与え，不完全な応答データ利用の有効性を検討することである．また，本論文では，若干の具体例にもと

づくシミュレーションによって漸近理論が適用できる標本数についての考察を行う．本論文の第2節で，潜在クラスモデルの概要を与え，第3節で拡張された潜在確率の推定法を示す．第4節で漸近分散・共分散行列の評価法を与え，漸近分散の観点から，不完全な応答データ利用の有効性を示す数値例を与える．第5節では，シミュレーションにより，小標本での最尤推定量の平均自乗誤差を潜在クラス数，質問項目数及び標本数を変えて評価する．同時に，小標本でも，平均自乗誤差基準に関して，不完全応答データ利用の有効性が成立することを示す．

2. 潜在クラスモデル

一般に，k 個の多肢変量：$I_1, I_2,\cdots,I_k$ が存在して，各々，$C_1, C_2,\cdots,C_k$ 個のカテゴリーをもつとする．また，これらの多肢変量以外に，T 個のカテゴリーをもつ多肢変量 Z の存在を仮定する．対象集団中の各個人は，上記の全ての変量に関して応答値を示すが，実際に観測されるのは最初の k 個の変量に関する応答値のみで，Z に関する応答値は現実に観測することができない．この意味で，Z を潜在変量，また，応答値が観測される k 個の変量を顕在変量と呼ぶ．

多くの場合，顕在変量としては，質問項目，反応試験等が対応し，潜在変量としては，採られた顕在変量群間の連関関係を説明する潜在概念が対応する．そして，その潜在概念に関して，何らかの尺度で異質とみなされる各クラスが，潜在変量の T 個のカテゴリーを形成する．

この変量間の関係は，変量が従う同時多項分布における生起確率間の関係式として以下のように表現される：

$$\mathrm{P}(\boldsymbol{i}, t)\equiv \mathrm{P}(Z=t)\cdot \mathrm{P}(\boldsymbol{i}\,|\,Z=t)=\mathrm{P}(Z=t)\cdot \prod_{\alpha=1}^{k} \mathrm{P}(I_\alpha=i_\alpha\,|\,Z=t), \tag{2.1}$$

ここに，$\boldsymbol{i}=(i_1,i_2,\cdots,i_k)$ は顕在変量への応答パターン，$\mathrm{P}(\boldsymbol{i}, t)$ は顕在変量と潜在変量への同時応答パターンが $(\boldsymbol{i}, t)$ となる確率，$\mathrm{P}(\boldsymbol{i}\,|\,Z=t)$ は，とくに，$Z=t$ の下で応答パターン $\boldsymbol{i}$ を示す条件付確率を示す．$\mathrm{P}(Z=t)$ は，各個人が潜在的にクラス t $(t=1,\cdots,T)$ に属する確率，または母集団上でのクラス t の構成率である．また，$\mathrm{P}(I_\alpha=i_\alpha\,|\,Z=t)$ は，潜在クラス t において，項目 I_α への応答がカテゴリー i_α となる条件付確率を表す．このとき，(2.1) 式の第2番目の等号は，潜在クラスが与えられた下での顕在変量間の条件付独立性を示している．これは広義の潜在構造モデルにおける局所独立の公理に対応するもので，各顕在変量のカテゴリー別応答確率は，クラス毎で特徴的に定まり，しかも同一クラス内における2つ以上の顕在変量の同時応答確率は，それぞれの変量において対応するカテゴリーの応答確率の積として単純に表現されることを表している．このことは，クラスに関する情報を与えない状態で顕在変量間に存在する連関は見かけ上のものであり，それは，異質の応答確率を有するクラスの混在に起因して生じていること，そして，もしクラスの情報が与えられれば，その連関は完全に消失することを意味している．

解析の目的は，顕在変量に関する対象集団の応答頻度データから，すべての潜在確率，$\mathrm{P}(I_\alpha=i_\alpha\,|\,Z=t)$ 及び，$\mathrm{P}(Z=t)$ を条件付独立性の仮定と制約条件：

$$\sum_{i_\alpha=1}^{C_\alpha} \mathrm{P}(I_\alpha=i_\alpha\,|\,Z=t)=1, \qquad \sum_{t=1}^{T} \mathrm{P}(Z=t)=1$$

の下に推定することである．そして，得られた推定値にもとづき，データとモデルとの適合度検定，各潜在クラスの解釈，また，各顕在変量の潜在クラスへの寄与度の評価，対象集団のクラスター分け等がなされる．

代表的な解析例に，林（1974）の〝政党支持の構造分析〟が挙げられる．ここでは，各個人

に対して，同一質問（政党支持について，3つのカテゴリー：1．自民党支持，2．支持なし，わからない，3．社会党支持，から一つを選択させる形式）を複数回（年毎に3年間）適用し，回答を導くに至る個人の意見の潜在的な安定性，強固性，頑健性の構造が潜在クラスモデルの下に分析されている．

3．潜在確率の推定アルゴリズム

いま，$n(\boldsymbol{i})$を，k個の顕在変量に応答パターン$\boldsymbol{i}$を示した個人の顕在頻度，$n(\boldsymbol{i}, t)$を顕在変量と潜在変量に関して，応答パターン$(\boldsymbol{i}, t)$を示した個人の潜在頻度とする．このとき，潜在頻度は潜在確率の最尤推定に関して十分統計量となっている．即ち，潜在頻度による潜在確率の最尤推定量は，

$$\begin{aligned}&\hat{\mathrm{P}}(Z=t)=\Big[\sum_{\{Z\}} n(\boldsymbol{i}, t)\Big] \Big/ N,\\ &\hat{\mathrm{P}}(I_a=i_a \mid Z=t)=\Big[\sum_{\{Z, I_a\}} n(\boldsymbol{i}, t)\Big] \Big/ \{N\cdot\hat{\mathrm{P}}(Z=t)\}\end{aligned} \tag{3.1}$$

によって求められる．ここに，Nは標本数，$\sum_{\{\ \}}$は，{　}内の変量以外の全ての多肢変量に関して，各々対応するカテゴリーについての和の演算を意味する．

実際に潜在頻度は観測されることはなく，k個の顕在変量に関する周辺頻度のみ観測される．これらの周辺頻度は，潜在変量Zについてグループ化された不完全頻度データでもある．グループ化された不完全データからの母数の最尤推定法の構築に，Dempster, Laird and Rubin (1977)によるEMアルゴリズムが適用できる．EMアルゴリズムは，期待値操作(Expectation-step)と最大化操作(Maximization-step)の反復を通して，目的である最尤推定値を導くアルゴリズムである．潜在クラスモデルの場合のE-stepは，顕在頻度と暫定的な潜在確率の推定値が与えられた下で，潜在頻度をその条件付期待値で推定することに対応している：

(3.2)　[E-step]

$$\begin{aligned}\hat{n}(\boldsymbol{i}, t)&=\mathrm{E}[n(\boldsymbol{i}, t) \mid n(i), \{\hat{\mathrm{P}}(Z=t), \hat{\mathrm{P}}(I_a=i_a \mid Z=t)\}]\\ &=n(\boldsymbol{i})\cdot\hat{\mathrm{P}}(\boldsymbol{i}, t) \Big/ \sum_{t=1}^{T} \hat{\mathrm{P}}(\boldsymbol{i}, t),\end{aligned}$$

ここに，$\hat{\mathrm{P}}(\boldsymbol{i}, t)$は，(2.1)式における第3項中の潜在確率を，各々対応する推定値で置き換えて得られたものである．

M-stepは，潜在頻度にもとづく尤度の最大化によって潜在確率の推定値を更新する段階で，上述の（3.1）式に対応する．

従って，適当な初期値から出発して，（3.2）式により潜在頻度を推定し，次に，（3.1）式によって潜在確率の推定値を更新し，また(3.2)式に戻る反復法がEMアルゴリズムにもとづく最尤推定法である．(3.2)式を直接(3.1)式に代入し，潜在頻度の項を消去することで，Goodmanの与えた反復法を得ることができる．しかし，ここでは潜在頻度の推定を意識しておく．なぜなら，そうすることで，顕在変量の幾つかが欠測した不完全な応答パターン頻度が存在しても，(3.2)式のE-stepの修正だけで，それらを完全応答パターン頻度と同様に推定に利用できるよう推定法を拡張することができるからである．

ここで，対象集団中の各個人の顕在変量I_aへの応答形態として，C_a個のカテゴリー以外に欠測を示すカテゴリー＊を新たに追加する．以下では，この欠測カテゴリー＊まで含めた場合の

添字を i_α^*, $i_\alpha^*=1,2,\cdots,C_\alpha$, $*$とし，欠測カテゴリー$*$を含まない場合の添字を i_α として，互いを区別する．応答パターン $\boldsymbol{i}$, $\boldsymbol{i}^*$ は各々，i_α, i_α^*, $\alpha=1,\cdots,k$ で構成される．

この場合の顕在頻度 $n(\boldsymbol{i}^*)$ は，潜在頻度が潜在変量及び欠測した顕在変量に関してグループ化された形の不完全データとみなすことができる．従って，この場合もEMアルゴリズムは適用可能で，このとき，M-stepは変らず，E-stepのみ次式となる．

$$\hat{n}(\boldsymbol{i},\ t)=\mathrm{E}[n(\boldsymbol{i},\ t)\mid\{n(\boldsymbol{i}^*)\},\ \{\hat{\mathrm{P}}(Z=t),\ \hat{\mathrm{P}}(I_\alpha=i_\alpha\mid Z=t)\}] \tag{3.3}$$
$$=\hat{\mathrm{P}}(\boldsymbol{i},\ t)\left[\sum_{\boldsymbol{i}^*}\delta(\boldsymbol{i},\ \boldsymbol{i}^*)n(\boldsymbol{i}^*)\Big/\sum_{t=1}^{T}\sum_{\substack{i_\alpha=1,\\ \alpha\in\{\alpha;\ i_\alpha^*=*\}}}^{C_\alpha}\hat{\mathrm{P}}(\boldsymbol{i},\ t)\right],$$

ここに，$\delta(\boldsymbol{i},\ \boldsymbol{i}^*)$ は，欠測カテゴリーを除いて，応答パターン $\boldsymbol{i}^*$ と $\boldsymbol{i}$ が一致したとき，1となり，それ以外のときは0となる変数である．

潜在頻度の推定値として，(3.3)式を(3.2)式の代わりに用いることにより，不完全な応答頻度も利用する最尤推定法が得られる．手順は，次のようになる．

（i） 潜在確率に適当な初期値：$\mathrm{P}(Z=t)^{(0)}$, $\mathrm{P}(I_\alpha=i_\alpha\mid Z=t)^{(0)}$ を与える．

（ii） 潜在確率の第 m 回目の推定値を求める（$m=1,2,\cdots$）：$\mathrm{P}^{(m)}(Z=t)$, $\mathrm{P}^{(m)}(I_\alpha=i_\alpha\mid Z=t)$ と，観測された顕在頻度 $\{n(\boldsymbol{i}^*)\}$ から(3.3)式により，潜在頻度の第 m 回目の推定値 $n^{(m)}(\boldsymbol{i},\ t)$ を全ての応答パターン $\boldsymbol{i}$ に対して計算する．

（iii） 潜在頻度の第 m 回目の推定値から（3.1）式にもとづき，第（$m+1$）回目の潜在確率の推定値 $\mathrm{P}^{(m)}(Z=t)$ と $\mathrm{P}^{(m+1)}(I_\alpha=i_\alpha\mid Z=t)$ を算出する．

（iv） 潜在確率の推定値が所与の精度で収束するまで，上記の（ii）と（iii）を繰り返す．

この推定法は，多項分布においてグループ化された頻度にもとづくパラメータの最尤推定をEMアルゴリズムにより具体化したものである．従って，適当な条件の下で，推定量が漸近一致性，漸近有効性，漸近正規性をもつことは，完全応答パターンの頻度のみ利用する場合と同じである（Sundberg (1976)，参照）．

4. 漸近分散・共分散行列の評価法

4.1 評価法の導出

Louis (1982) は，一般に，不完全データにもとづく観測情報量 $I_Y(\theta)$ として，

$$I_Y(\theta)=\mathrm{E}_\theta[B(X,\ \theta)\mid X\in R]-\mathrm{E}_\theta[S(X,\ \theta)S^T(X,\ \theta)\mid X\in R] \tag{4.1}$$

を与えている．ここに，

X	：観測できない完全データ
$Y=Y(X)$	：X の代わりに観測された不完全データ
$S(X,\ \theta)$	：X の対数尤度の θ に関する1階偏導関数ベクトル
$B(X,\ \theta)$	：X の対数尤度の θ に関する2階偏導関数行列の負値
$R=\{x:\ y(x)=y\}$	：観測値 y をとり得る X の標本値集合

である．

とくに，完全データ X が，多項分布から得られる N 個の独立な指標ベクトル，X_1, X_2,$\cdots$, X_N であるとき，(4.1)式は，

$$I_Y(\hat{\theta})=\sum_{j=1}^{N}S(\hat{X}_j,\ \hat{\theta})S^T(\hat{X}_j,\ \hat{\theta}) \tag{4.2}$$

となる．ここに，

$$\hat{X}_j = \mathrm{E}_\theta[X_j \mid X_j \in R]. \tag{4.3}$$

さて，潜在クラスモデルにおいて，潜在頻度は，$M \times T$ 個のセルをもつ多項分布に従うと考えてよい．ここに，M は顕在変量に対する応答パターンの総数：$M = \prod_{\alpha=1}^{k} C_\alpha$，$T$ は潜在クラス数である．

いま，$\boldsymbol{i}^{(1)}, \boldsymbol{i}^{(2)}, \cdots, \boldsymbol{i}^{(M)}$ を顕在変量に関する M 個の応答パターンとする．ここに，$\boldsymbol{i}^{(m)} = (i_1^{(m)}, \cdots, i_k^{(m)})$，$i_\alpha^{(m)} = 1, 2, \cdots, C_\alpha$，$\alpha = 1, 2, \cdots, k$，$m = 1, 2, \cdots, M$ である．また，(4.2) 式の X_j は，対象集団中の第 j 番目の個体が示す潜在的な応答指標である．ここに，$X_j = (x_j(1,1), \cdots, x_j(1,T), x_j(2,1), \cdots, x_j(m,t), \cdots, x_j(M,T))$ で，これは，顕在変量及び潜在変量に関する第 j 番目の個体の応答パターンが $(\boldsymbol{i}^{(m)}, t)$ であるとき，それに対応するセルの要素 $x_j(m,t)$ のみ 1 で，それ以外の要素 $x_j(m', t')$ ($m' \neq m$ または，$t' \neq t$) は 0 となる変数である．しかし，実際には，この X_j は観測されないため，これを (4.3) 式に従い，観測された顕在変量への応答パターンの下での条件付期待値で推定する．

もし，顕在変量への応答パターンに欠測がないとき，応答パターン $\boldsymbol{i}^{(q)}$ を示す第 j 番目の個体の指標は，

$$\hat{X}_j = \{\hat{x}_j(m,t)\}, \tag{4.4}$$

によって推定される．ここに，$\gamma(m,q)$ を，$m = q$ のとき 1，それ以外では 0 となる変数と定義するとき，$\hat{x}_j(m,t)$ は，

$$\hat{x}_j(m,t) = \gamma(m,q)\hat{\mathrm{P}}(\boldsymbol{i}^{(m)}, t) \Big/ \sum_{t=1}^{T} \hat{\mathrm{P}}(\boldsymbol{i}^{(m)}, t),$$

によって与えられる．

もし，顕在変量への応答パターンが欠測カテゴリーを含んだ $\boldsymbol{i}^{*(q)}$ であれば，

$$\hat{x}_j(m,t) = \delta(\boldsymbol{i}^{(m)}, \boldsymbol{i}^{*(q)})\hat{\mathrm{P}}(\boldsymbol{i}^{(m)}, t) \Big/ \sum_{t=1}^{T} \sum_{\substack{i_\alpha = 1, \\ \alpha \in \{\alpha;\, i_\alpha^* = *\}}}^{C_\alpha} \hat{\mathrm{P}}(\boldsymbol{i}^{(m)}, t) \tag{4.5}$$

となる．

潜在応答指標，$X_j = \{x_j(m,t)\}$ が与えられた下での対数尤度 L は，

$$L \propto \sum_{m=1}^{M} \sum_{t=1}^{T} x_j(m,t) \left[\log \mathrm{P}(Z=t) + \sum_{\alpha=1}^{k} \log \mathrm{P}(I_\alpha = i_\alpha \mid Z=t) \right], \tag{4.6}$$

となる．

従って，潜在確率 $\theta = [\{\mathrm{P}(Z=t')\},\ t' = 1, 2, \cdots, (T-1),\ \{\mathrm{P}(I_\alpha = i_\alpha' \mid Z=t)\},\ i_\alpha' = 1, 2, \cdots, (C_\alpha - 1),\ t = 1, 2, \cdots, T,\ \alpha = 1, 2, \cdots, k]$ に対して，1 階偏導関数 $S(X_i, \theta)$ は，以下で与えられる：

$$\left\{ \begin{array}{l} \sum_{m=1}^{M} \{x_j(m,1)/\mathrm{P}(Z=1) - x_j(m,T)/\mathrm{P}(Z=T)\} \\ \quad \vdots \\ \sum_{m=1}^{M} \{x_j(m,t')/\mathrm{P}(Z=t') - x_j(m,T)/\mathrm{P}(Z=T)\} \\ \quad \vdots \\ \sum_{m=1}^{M} \{x_j(m,T-1)/\mathrm{P}(Z=T-1) - x_j(m,T)/\mathrm{P}(Z=T)\} \\ \quad \vdots \\ \sum_{m=1}^{M} \{\delta(i_1^{(m)}, 1)\, x_j(m,1)/\mathrm{P}(I_1=1 \mid Z=1) \end{array} \right.$$

$$(4.7)\qquad S(X_j,\theta)=\begin{vmatrix} -\delta(i_1^{(m)},C_1)\,x_j(m,1)/\mathrm{P}(I_1=C_1\mid Z=1)\} \\ \vdots \\ \sum_{m=1}^{M}\{\delta(i_1^{(m)},C_1-1)\,x_j(m,1)/\mathrm{P}(I_1=C_1-1\mid Z=1) \\ -\delta(i_1^{(m)},C_1)\,x_j(m,1)/\mathrm{P}(I_1=C_1\mid Z=1)\} \\ \vdots \\ \sum_{m=1}^{M}\{\delta(i_a^{(m)},i_a)\,x_j(m,t)/\mathrm{P}(I_a=i_a\mid Z=t) \\ -\delta(i_a^{(m)},C_a)\,x_j(m,t)/\mathrm{P}(I_a=C_a\mid Z=t)\} \\ \vdots \\ \sum_{m=1}^{M}\{\delta(i_k^{(m)},C_k-1)\,x_j(m,T)/\mathrm{P}(I_k=C_k-1\mid Z=T) \\ -\delta(i_k^{(m)},C_k)\,x_j(m,T)/\mathrm{P}(I_k=C_k\mid Z=T)\} \end{vmatrix}$$

第 j 番目の個人が顕在変量に対して実際に示した応答パターンにもとづき，(4.4) 式，もしくは，(4.5) 式を潜在確率の最尤推定値 θ を用いて評価し，その X_j を (4.7) 式に代入して，$S(X_j,\theta)$ とする．観測情報量 $I_Y(\theta)$ は，

$$(4.8)\qquad I_Y(\theta)=\sum S(X_j,\theta)\,S^T(X_j,\theta)=\sum n(i^*)\,S(X(i^*),\theta)\,S^T(X(i^*),\theta)$$

となる．$X(i^*)$ は，顕在変量への応答パターンが i^* のときに推定される指標ベクトルである．

目的である潜在確率の最尤推定量の漸近分散・共分散行列の評価は，$I_Y(\theta)^{-1}$ によって求められる．

4.2 数値例

ある潜在クラスモデルを想定し，乱数により顕在変量への応答頻度データを作成し，そのデータから得られた最尤推定値の漸近分散を計算した結果を以下に示す．とくに，作成したデータにランダムな欠測を生じさせて，不完全な応答頻度データを作成し，その場合の推定値の漸近分散も示す．

取り扱ったモデルは，2クラス3項目モデルで，各項目への応答はすべて2値型とした．クラスの構成率及びクラス内正応答の確率を表4.1に示す．

このモデルから乱数により，$N=500$（人）に対応する3項目への応答パターンに関する次の3通りの形態の頻度データを作成した．

（i） 完全応答頻度データ，$N=500$.

（ii） (i)のデータを得る過程において，ランダムに，30%，50%，70% の欠測応答を生じさせた3種の不完全応答頻度データ．

（iii） (ii) のデータから不完全応答パターンに対する頻度を除いた縮小完全応答頻度データ．

上記の各データについて，潜在確率の最尤推定値を各々算出し，それらの漸近分散を求めた結果が表4.2である．

表4.2から，縮小完全データにおいて，標本数が増すに従って漸近分散が減少すること，ま

表4.1 潜在クラスモデル

	構成率	項目1	項目2	項目3
クラス1	0.4	0.3	0.2	0.25
クラス2	0.6	0.8	0.8	0.85

表 4.2　漸近分散の評価：表は実際の数値の 10^2 倍を与えている．また，各枠内での配置は表 4.1 で与えた潜在確率の配置に対応している．

完全データ $N=500$				欠測率 (%)	不完全データ				N	縮小完全データ			
				30	.40	.33 .16	.37 .22	.46 .14	350	.46	.40 .18	.45 .24	.50 .21
.32	.28 .13	.31 .17	.35 .15	50	.49	.38 .19	.44 .27	.59 .13	250	.64	.56 .26	.62 .34	.71 .29
				70	.66	.46 .25	.55 .36	.83 .13	150	1.07	.93 .43	1.04 .56	1.18 .49

た，不完全データでは，欠測率が高くなるほど漸近分散が増大することがわかる．また，データに不完全応答パターンが生じた場合には，それらを除去して便宜的な完全データにすることにより，不完全頻度も推定に利用するほうが推定値の精度を向上させることがわかる．

5．小標本での平均自乗誤差の評価

本節では，シミュレーションにより，漸近理論が適用できない小標本での最尤推定値の性質を調べる．とくに，潜在クラス数 T，項目数 k，標本数 N の変化がどのような影響を与えるか，また，漸近理論が適用できる標本数は実際にどれ程であるかについて言及する．シミュレーション回数は，すべて 1000 回である．

（i）　2 クラス 3 項目モデル（表 5.1）

$N=50$ から 4000 の間の各標本数毎に，最尤推定値の平均自乗誤差を調べた結果が表 5.2 である．とくに，平均自乗誤差を分散と偏りの自乗に分解した数値も添えている．表中の数値は全て実際に得られた数値を 10^2 倍したものである．分散の列における（ ）内の数値は，本実験での平均的な頻度データにもとづく最尤推定値の漸近分散を算出したものである．平均自乗誤差は，標本数が増すと減少する．$N=100$ のとき，本モデルにおいて推定値が偏りをもつため，漸近分散の値と実際の分散値の間に隔たりがみられる．$N=500$ で，偏りがほぼ 0 にちかくなり，シミュレーションによる分散値と漸近分散値とが小数第 3 桁程度まで一致する．

（ii）　2 クラス 5 項目モデル（表 5.3）

この場合の各標本数における最尤推定値の平均自乗誤差を表 5.4 に示している．$N=100$ で偏りがみられるが，$N=500$ では消失する．2 クラス 3 項目モデルでの結果と比較すると，平均自乗誤差及び偏りの双方で，値は小さくなっている．

（iii）　3 クラス 5 項目モデル（表 5.5）

このモデルでは，とくに標本数が小さいときクラスの構成率の推定値が 0 となることがある．つまり，モデルの推定として，2 クラスモデルが採択される現象が起こる．この現象は，1000 回のシミュレーション中，$N=50$ のとき 189 回，$N=100$ のとき 227

表 5.1　2 クラス 3 項目モデルでの潜在確率

	構成率	項目 1	項目 2	項目 3
クラス 1	0.4	0.6	0.7	0.8
クラス 2	0.6	0.1	0.2	0.3

表5.2 2クラス3項目モデルでの平均自乗誤差表：表中の数値は実際の値の10^2倍である．また，各枠内での数値の配列は表5.1の潜在確率に対応している．

標本数 N	平均自乗誤差				分散（漸近分散）				（偏り）^{2}			
50	7.1	10.6	8.9	4.0	.51 (5.9)	4.9 (5.0)	.29 (4.3)	.00 (3.7)	6.5	5.7	8.6	4.0
		1.9	1.7	2.0		.48 (1.5)	.73 (2.0)	.76 (2.3)		1.4	.94	1.3
100	7.3	2.8	1.6	2.9	.94 (3.0)	1.1 (2.5)	.74 (2.2)	.54 (1.8)	6.4	1.7	.90	2.4
		1.0	3.7	2.4		.00 (.74)	.40 (.98)	1.7 (1.1)		1.0	3.3	.74
500	.64	.55	.43	.38	.63 (.59)	.55 (.50)	.43 (.43)	.37 (.37)	.01	.00	.00	.00
		.16	.23	.25		.15 (.15)	.22 (.20)	.25 (.23)		.00	.01	.00
1000	.31	.28	.22	.19	.31 (.31)	.28 (.25)	.22 (.22)	.19 (.18)	.00	.00	.00	.00
		.08	.11	.11		.08 (.08)	.11 (.10)	.11 (.11)		.00	.00	.00
2000	.15	.13	.11	.09	.15 (.15)	.13 (.13)	.11 (.11)	.09 (.09)				
		.04	.05	.06		.04 (.04)	.05 (.05)	.06 (.06)				
3000	.10	.09	.07	.06	.10	.09	.07	.06				
		.03	.03	.04		.03	.03	.04				
4000	.07	.06	.06	.04	.07	.06	.06	.04				
		.02	.03	.03		.02	.03	.03				

表5.3 2クラス5項目モデルでの潜在確率

	構成率	項目1	項目2	項目3	項目4	項目5
クラス1	0.4	0.6	0.7	0.8	0.85	0.9
クラス2	0.6	0.1	0.2	0.3	0.35	0.4

回起こり，N が500以上では起こらなかった．表中の値は，このクラス縮減が起きなかった場合の平均である．結果を表5.6に示している．ここでは，$N=1000$ 位まで若干の偏りが残る．$N=2000$ で殆どこの影響が消えている．全体的に，2クラスモデルと比較して，大きな平均自乗誤差及び偏りを示している．

（iv） 2クラス3項目モデル（表5.7）——不完全応答データを含む場合——

このモデルに対し，標本数500のデータを先ず作成し，これから第3節の数値例のときと同様な手順で不完全応答データを作り，もとの完全データ，不完全応答を含むデータ，この不完全データから不完全な応答頻度を削除した縮小完全データの3種の頻度データを作った．これらに対して，平均自乗誤差を比較した．表5.8は，欠測率を30%としたときの結果であり，表5.9と表5.10は，それぞれ欠測率を50%と70%にしたときの結果である．30%の欠測の場合，不完全応答の頻度を推定に利用するこ

表5.4　2クラス5項目モデルでの平均自乗誤差表：表中の数値は実際の値の 10^2 倍である．また，各枠内での数値の配列は表5.3の潜在確率に対応している．

標本数 N	平均自乗誤差						分						(偏　り)^2					
50	9.4	15.4	8.9	4.0	2.3	1.0	.23	.68	.21	.04	.00	.00	9.2	14.7	8.7	4.0	2.2	1.0
		2.0	2.5	2.8	2.8	3.2		.44	.57	.57	.59	.63		1.6	1.9	2.2	2.2	2.5
100	3.6	1.8	1.6	3.6	2.2	1.0	.34	1.6	1.3	.41	.03	.02	3.3	.22	.25	3.2	2.2	.98
		1.4	1.4	.82	.95	1.2		.25	.32	.40	.43	.41		1.1	1.1	.44	.52	.78
500	.17	.22	.20	.17	.13	.11	.17	.22	.20	.17	.13	.11	.00	.00	.00	.00	.00	.00
		.06	.09	.11	.12	.13		.06	.09	.11	.12	.13		.00	.00	.00	.00	.00
1000	.07	.10	.10	.08	.06	.05	.07	.10	.10	.08	.06	.05						
		.03	.04	.05	.06	.06		.03	.04	.05	.06	.06						
2000	.04	.05	.05	.04	.03	.03	.04	.05	.05	.04	.03	.03						
		.02	.02	.03	.03	.03		.02	.02	.03	.03	.03						

表5.5　3クラス5項目モデルでの潜在確率

	構成率	項目1	項目2	項目3	項目4	項目5
クラス1	0.4	0.5	0.6	0.65	0.7	0.8
クラス2	0.3	0.9	0.8	0.85	0.3	0.2
クラス3	0.3	0.1	0.2	0.25	0.3	0.4

表5.6　3クラス5項目モデルでの平均自乗誤差表：表中の数値は実際の値の 10^2 倍である．また，各枠内での数値の配列は表5.5の潜在確率に対応している．

標本数 N	MSE						分						(偏　り)^2					
50	15.0	25.0	16.0	12.3	9.00	4.10	.74	.00	.08	.29	.00	.11	14.3	25.0	15.9	12.0	9.00	3.90
	4.89	1.00	3.99	2.25	8.79	4.12	.30	.00	.00	.01	.69	.72	4.59	1.00	3.99	2.24	8.10	3.41
	35.5	13.5	8.99	9.00	4.29	2.52	.40	.60	.63	.60	.65	.64	35.1	12.9	8.36	8.40	3.64	1.89
100	15.1	24.9	35.8	12.2	8.99	4.00	.05	.08	.13	.13	.02	.00	15.1	24.7	35.7	12.1	8.97	4.00
	3.71	1.00	3.93	2.24	6.49	3.94	.45	.01	.09	.01	4.0	2.3	3.26	0.99	3.84	2.23	2.25	1.61
	33.0	11.1	8.37	8.01	3.95	2.37	.59	.55	.44	.42	.35	.34	32.4	10.5	7.93	7.59	3.60	2.03
500	8.10	1.70	0.21	0.21	2.10	5.10	.64	.50	.16	.15	.25	.39	7.46	1.21	0.05	0.06	1.81	4.68
	4.72	0.98	3.43	2.12	8.19	3.76	.36	.06	.53	.20	.70	.33	4.37	0.92	2.89	1.91	7.49	3.43
	0.70	0.99	0.93	1.04	0.53	0.69	.29	.02	.60	.68	.51	.52	0.41	0.97	0.33	0.36	0.02	0.17
1000	1.34	0.30	0.21	0.20	0.51	3.77	.33	.30	.21	.19	.49	.30	1.00	.001	.000	.000	.022	3.48
	0.96	0.48	0.25	0.25	0.56	0.62	.29	.17	.15	.14	.14	.62	0.67	0.31	0.09	0.11	0.42	.000
	0.27	0.40	0.19	0.21	0.33	0.23	.24	.40	.19	.21	.33	.23	0.03	0.04	.008	0.01	0.16	.002
2000	0.28	0.13	0.09	0.08	0.18	0.30	.28	.13	.09	.08	.18	.30	.000	.001	.001	.000	.001	.002
	0.16	0.12	0.07	0.07	0.13	0.26	.16	.12	.07	.07	.13	.26	.000	.000	.000	.000	.001	.004
	0.10	0.10	0.11	0.10	0.12	0.12	.10	.10	.10	.10	.12	.12	.000	.003	.003	.001	.000	.000

表5.7 2クラス3項目モデルでの潜在確率

	構成率	項目1	項目2	項目3
クラス1	0.4	0.3	0.2	0.25
クラス2	0.6	0.8	0.8	0.85

表5.8 30% 欠測の下での平均自乗誤差表：表は実際の数値の 10^2 倍を示している．また，各枠内での配置は表5.7で与えた潜在確率の配置に対応している．

標本数 N	平均自乗誤差				分散				(偏り)^2			
500	.33	.29	.32	.38	.32	.28	.30	.37	.01	.01	.02	.01
完全		.13	.19	.13		.11	.19	.13		.00	.00	.00
500	.42	.34	.39	.58	.41	.33	.37	.55	.01	.01	.02	.03
不完全		.15	.23	.14		.15	.23	.14		.00	.00	.00
350	.45	.37	.39	.62	.43	.36	.36	.60	.02	.01	.03	.02
縮小完全		.14	.24	.14		.14	.24	.14		.00	.00	.00

表5.9 50% 欠測の下での平均自乗誤差表：表は実際の数値の 10^2 倍を示している．また，各枠内での配置は表5.7で与えた潜在確率の配置に対応している．

標本数 N	平均自乗誤差				分散				(偏り)^2			
500	.38	.29	.36	.50	.38	.28	.35	.50	.00	.01	.01	.00
完全		.12	.27	.15		.12	.27	.15		.00	.00	.00
500	.55	.43	.55	.70	.53	.42	.55	.70	.02	.01	.00	.00
不完全		.18	.43	.24		.17	.42	.23		.01	.01	.01
250	.55	.53	.63	.79	.54	.53	.63	.79	.01	.00	.00	.00
縮小完全		.22	.48	.24		.21	.47	.23		.01	.01	.01

表5.10 70% 欠測の下での平均自乗誤差表：表は実際の数値の 10^2 倍を示している．また，各枠内での配置は表5.7で与えた潜在確率の配置に対応している．

標本数 N	平均自乗誤差				分散				(偏り)^2			
500	.33	.37	.40	.33	.33	.37	.38	.33	.00	.00	.02	.00
完全		.12	.24	.19		.12	.23	.19		.00	.01	.00
500	.75	.68	.80	.92	.75	.68	.78	.92	.00	.00	.02	.00
不完全		.31	.53	.43		.30	.51	.43		.01	.02	.00
150	1.2	3.0	5.4	1.6	1.0	1.1	1.1	1.2	.20	1.9	4.3	.40
縮小完全		.38	3.2	1.3		.37	.64	.58		.01	2.6	.68

とで，逆に捨ててしまうより幾分大きな偏りを示すが，分散部分が小さくなるので，平均自乗誤差は不完全応答を利用するほうに若干の優位がみられた．この傾向は，欠測率が50%に増加したとき，顕著になる．さらに，70%まで欠測率が上がると，不完全応答頻度の割合が大きくなり，従ってそれを捨ててしまうことの損失が偏り部分と分散部分の双方で同時にみられた．

6. 結　　言

潜在クラス分析において，不完全な応答頻度も利用できる潜在確率の最尤推定法を与えた．また，その際の漸近分散の評価法も併せて示した．潜在クラス分析において，ある程度信頼のおける潜在確率の推定値を得ようとすれば，かなりの標本数が必要となることが，シミュレーションにより示唆された．従って，不完全な応答データが多いときに，それらを捨てることはかなりの損失である．ここでは，不完全な応答データを推定に利用できる手法を提唱し，不完全な応答データを利用する場合と，それらを捨てる場合の双方における推定値の漸近分散及び平均自乗誤差を比較し，提唱した推定法が有用であることを示した．

また，同じ標本数の下で項目数を増やせば，推定すべき潜在確率も関連した分多くなるが，応答パターン数が増えるので，推定値の精度を向上させることがわかった．逆に，潜在クラス数が増せば，応答パターン数は変らず，推定すべき潜在確率の数のみ多くなるので，推定値の精度は著しく悪くなる．また，潜在クラス数に比べて標本数が小さければ，実際より潜在クラス数を少なく推定する危険性があることがわかった．シミュレーションの結果は想定した潜在確率の真値にも依存するものである．この関係は，今後の研究課題とする．

謝　　辞

本報告は，昭和60年度統計数理研究所共同研究 "数量化の方法論と応用に関する研究会" (60-共研-12)での発表をもとに作成したものである．有意義なご意見をお聞かせ下さった参加者の先生方，発表と執筆の双方の機会を与えて下さった統計数理研究所 駒澤勉教授にここに感謝致します．また，丁寧な査読をして下さいましたレフェリーの先生方に感謝致します．

参 考 文 献

Anderson, T.W. (1954). On estimation of parameters in latent structure analysis, *Psychometrika,* **19**, 1-10.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J.R. Statist. Soc.,* B**39**, 1-38.

Formann, A.K. (1978). A note on parameter estimates for Lazarsfeld's latent class analysis, *Psychometrika,* **43**, 123-126.

Gibson, W.A. (1955). An extension of Anderson's solution for the latent structure equations, *Psychometrika,* **20**, 69-73.

Goodman, L.A. (1979). On the estimates of parameters in latent structure analysis, *Psychometrika,* **44**, 123-128.

Green, B.F., Jr. (1951). A general solution for the latent class model of latent structure analysis, *Psychometrika,* **16**, 151-161.

林知己夫，樋口伊佐夫，駒澤　勉 (1970). 情報処理と統計数理，産業図書.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis, *Measurement and Prediction*, Princeton University Press, Chapter 10.

Lazarsfeld, P.F. and Henry, N.W. (1968). *Latent Structure Analysis,* Houghton Mifflin.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm, *J.R. Statist. Soc,* B**44**, 226-233.

Morita, T. (1980). Stability of the solution in latent class analysis, *J. Japan Statist. Soc.,* **10**, 37-46.

Morita, T. (1984). A procedure to deal with improper solutions in latent class analysis, *J. Japan Statist. Soc.,* **14**, 29-34.

Okamoto, M. and Isogai, T. (1978). Stability of the solution in latent class analysis, *Proc. 11th SDA Symp.,* 15-19.

Sato, Y., Sugawa, K. and Kawaguchi, M. (1980). On an error estimation in the latent class analysis, *Bull. Facul. Engin.,* Hokkaido University, No. **97**.

Sundberg, R. (1976). An iterative method for incomplete data from an exponential family, *Scand. J. Statist.,* **1**, 49-58.

A Maximum Likelihood Estimation Procedure and the Numerical Evaluation of the Resultant Estimates in Latent Class Analysis With or Without Missing Entries

Michiko Watanabe

(Department of Science, Research Institute of Fundamental Information Science, Kyushu University)

In this paper, we introduce an algorithm for deriving the maximum likelihood estimates and their asymptotic variance-covariance matrix in the latent class model, which is applicable to the case of incomplete data as well as the case of complete data. This algorithm is constructed in accordance with the EM algorithm from the viewpoint that the latent class model can be considered as a model based on a mixing multinomial distribution.

The latter part of the paper is concerned with the numerical experiment for evaluating the properties of the maximum likelihood estimates yielded by the above estimation method. Specially, our interest is concentrated on the following two points: the first is whether it is more efficient or not, in estimating the model parameters, to make use of the response patterns with missing entries than to abandon them, and the second is how it reflects on the precision of the estimates to change the number of latent classes assumed, or the number of test items examined. As the result, the superiority of using the incomplete data is confirmed in both respects of the mean square errors and of the asymptotic variances. Concerning the latter question, we have got the result that the larger number of test items and the less latent classes we employ, the more efficient estimates we can get.