潜在変数と顕在変数の間の確率関係

潜在変数ベクトル $\mathbf{X}$ と p 個の顕在変数からなる顕在変数ベクトル $\mathbf{Y} = (Y_1, Y_2, ..., Y_p)$ の確率的な関係は次式で表現される。

$$f(y,\gamma) = \int_{\Omega(\mathbf{x})} \pi(y \mid x,\alpha)\varphi(x,\beta)d\mathbf{y}$$

ここで、 $f(y,\gamma)$ は確率変数 $\mathbf{Y}$ の確率密度関数、 $\varphi(x,\beta)$ は潜在変数 $\mathbf{X}$ の確率密度関数、 $\pi(y \mid x,\alpha)$ は $\mathbf{X}$ が与えられた下での $\mathbf{Y}$ の条件付密度関数。 α， β は潜在パラメータ、 γ は顕在パラメータという。

潜在構造分析法とは、顕在変数 $\mathbf{Y}$ に関するデータに基づいて潜在パラメータ α， β の推定を行う分析法の総称である。

局所独立の仮定

潜在構造モデルでは、潜在変数が与えられた下では p 個の顕在変数は互いに統計的に独立であることが仮定される。これを局所独立の仮定という。

$$\pi(\mathbf{y} \mid \mathbf{x},\alpha) = \prod_{j=1}^{p} \pi_j(y_j \mid \mathbf{x},\alpha)$$

ここで、 $\pi_i(y_i \mid \mathbf{x},\alpha)$ は潜在変数 $\mathbf{X}$ が与えられた下での個々の顕在変数 Y_i の条件付確率変数を表す。 $(i = 1,2,...., p)$

潜在クラスモデル

$A, B, C,...$ を p 個の質的な顕在変数とし、各変数の応答カテゴリー数を $I, J, K,...$ とする。今、 $p = 3$ とし、またして各潜在変数 X を仮定し、そのカテゴリー数を T とする。確率を表す記号を以下のように定義する。

w_t　：潜在変数 X に対してカテゴリー t で応答する確率

s_{ijk}^{ABC}　：顕在変数 $(A, B, C).$ に対して応答パターン $(i, j, k).$ を示す同時確率

π_{ijkt}^{ABCX}　：顕在変数 $(A, B, C, X).$ に対して応答パターン $(i, j, k, t).$ を示す同時確率

$\pi_{ijkt}^{\overline{A}\;\overline{B}\;\overline{C}\;X}$　：潜在クラス t が与えられた下で顕在変数 $(A, B, C).$ に対して応答パター

ン(i, j, k). を示す条件付き確率
これらの確率に以下の制約を置く。

$$s_{ijk}^{ABC} = \sum_{t=1}^{T} \pi_{ijkt}^{ABCX} = \sum_{t=1}^{T} w_t \pi_{ijkt}^{\bar{A}\bar{B}\bar{C}X}, \sum_{t=1}^{T} w_t = 1$$

この制約により、各個体の潜在変数 X への応答値はその個体がどの潜在クラスに属しているかを表す値であり、w_t は母集団上の潜在クラス t の構成率であると解釈できる。また、局所独立も仮定される。

$$\pi_{ijkt}^{\bar{A}\bar{B}\bar{C}X} = \pi_{it}^{\bar{A}X} \pi_{jt}^{\bar{B}X} \pi_{kt}^{\bar{C}X}$$

ここで $\pi_{it}^{\bar{A}X}$，$\pi_{jt}^{\bar{B}X}$，$\pi_{kt}^{\bar{C}X}$ は t が与えられた下での A, B, C それぞれに対する条件付応答確率であり、次の制約を持つ。

$$\sum_{i=1}^{I} \pi_{it}^{\bar{A}X} = \sum_{j=1}^{J} \pi_{jt}^{\bar{B}X} = \sum_{k=1}^{K} \pi_{kt}^{\bar{C}X} = 1$$

潜在クラスモデルでは、顕在変数のみに関わる確率 s_{ijk}^{ABC} を潜在確率（潜在パラメータ）、潜在変数にかかわる確率すなわち潜在クラス構成比 w_t と各潜在クラス内での顕在変数への応答比率を合わせて潜在確率（潜在パラメータ）と呼ぶ。

潜在クラスモデル分析の目的は、潜在確率により想定される母集団の構造を顕在変数に関する応答パターン別セル度数 $\{n_{ijk}^{ABC}\} i = 1,2,...I, j = 1,2,..., J, k = 1,2,...K$ から推測することである。

潜在確率の推定

顕在確率 s_{ijk}^{ABC} は通常多項分布の生起確率なので、その最尤推定値は調査集団が示す顕在変数への応答度数 $\{n_{ijk}^{ABC}\}$ を用いて、以下で計算される。

$$s_{ijk}^{ABC} = \frac{n_{ijk}^{ABC}}{N}$$

$$\hat{s}_i^A = \frac{n_i^A}{N}, \hat{s}_{ij}^{AC} = \frac{n_{ij}^{AB}}{N}, ...$$

潜在変数の推定値は陽な式では求められず、一般的には最尤法が用いられる。

すなわち応答パターン別度数n_{ijk}^{ABC}を多項分布からの実現値とみなして、次の対数尤度Lを最大化する潜在確率の値を最尤推定値として求める。

$$L=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}n_{ijk}^{ABC}\log s_{ijk}^{ABC}$$
$$=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}n_{ijk}^{ABC}\log\left[\sum_{t=1}^{T}w_t\pi_{it}^{\bar{A}X}\pi_{jt}^{\bar{B}X}\pi_{kt}^{\bar{C}X}\right]$$

この式を陽な数式で求めることはできないので、EM アルゴリズムが用いられる。

EMアルゴリズム

E-Step（期待値の計算）と M-step（パラメータの更新）の 2 つのステップの反復として定義される。E-Step では実際に観測されたデータと推定すべきパラメータの暫定値が与えられた下で、擬似的な完全データの対数尤度の条件付期待値を計算する。M-step では E-Step で求めた対数尤度の条件付期待値を最大化するパラメータの値を求め、その値でパラメータの推定値を更新する。この 2 つのステップの反復計算により、観測データに基づく対数尤度Lを最大化する最尤推定値を求める。

推定すべき潜在確率

潜在	ｸﾗｽの	A			B			C		
ｸﾗｽ	ｻｲｽﾞ	1	･･･	I	1	･･･	J	1	･･･	K
1	w_1	$\pi_{11}^{\bar{A}X}$	･･･	$\pi_{I1}^{\bar{A}X}$	$\pi_{11}^{\bar{B}X}$	･･･	$\pi_{J1}^{\bar{B}X}$	$\pi_{11}^{\bar{C}X}$	･･･	$\pi_{K1}^{\bar{C}X}$
･･･	･･･									
t	w_t	$\pi_{1t}^{\bar{A}X}$	･･･	$\pi_{It}^{\bar{A}X}$	$\pi_{1t}^{\bar{B}X}$	･･･	$\pi_{Jt}^{\bar{B}X}$	$\pi_{1t}^{\bar{C}X}$	･･･	$\pi_{Kt}^{\bar{C}X}$
･･･	･･･									
T	w_T	$\pi_{1T}^{\bar{A}X}$	･･･	$\pi_{IT}^{\bar{A}X}$	$\pi_{1T}^{\bar{B}X}$	･･･	$\pi_{JT}^{\bar{B}X}$	$\pi_{1T}^{\bar{C}X}$	･･･	$\pi_{KT}^{\bar{C}X}$

潜在クラスモデルにおける擬似的な完全データとは、潜在クラス別の顕在変数に関する応答パターン別セル度数（潜在度数）である。

応答パターン(i,j,k).に対する観測度数n_{ijk}^{ABC}と観測度数$n^{*}{}_{ijk}^{ABC}$の関係は、以下で表せる。

$$n_{ijk}^{ABC} = \sum_{t=1}^{T} n^{*ABCX}_{ijkt}$$

潜在度数 n^{*ABCX}_{ijkt} はすべての潜在確率の十分統計量なので、完全データ（潜在度数）に基づく対数尤度の条件付期待値を評価する＝潜在度数の条件付期待値の評価である。従って

E-step は観測度数 n_{ijk}^{ABC} と潜在確率の暫定値が与えられた下で、潜在度数 n^{*ABCX}_{ijkt} の条件付期待値を計算し、それを潜在度数 n^{*ABCX}_{ijkt} の新しい推定値とする。

E-step：条件付期待値で潜在度数を推定する

$$\hat{n}^{*ABCX}_{ijkt} = E\left[n^{*ABCX}_{ijkt} \mid n_{ijk}^{ABC}; w_t, \pi_{it}^{\bar{A}X}, \pi_{jt}^{\bar{B}X}, \pi_{kt}^{\bar{C}X}\right]$$

$$= n_{ijk}^{ABC} \frac{w_t \pi_{it}^{\bar{A}X} \pi_{jt}^{\bar{B}X} \pi_{kt}^{\bar{C}X}}{\sum_{t=1}^{T} w_t \pi_{it}^{\bar{A}X} \pi_{jt}^{\bar{B}X} \pi_{kt}^{\bar{C}X}}$$

次に M-step では、この潜在度数の推定値を使って潜在確率に関して通常の最尤推定を行い、潜在確率の推定値の更新を行う。

M-Step：潜在確率の推定

$$\hat{w}_t = \frac{\sum_{(i,j,k)} \hat{n}^{*ABCX}_{ijkt}}{N}$$

$$\pi_{it}^{\bar{A}X} = \frac{\sum_{(j,k)} \hat{n}^{*ABCX}_{ijkt}}{N}$$

$$\pi_{jt}^{\bar{B}X} = \frac{\sum_{(i,k)} \hat{n}^{*ABCX}_{ijkt}}{N}$$

$$\pi_{kt}^{\bar{C}X} = \frac{\sum_{(i,j)} \hat{n}^{*ABCX}_{ijkt}}{N}$$

これを収束するまで繰り返す。

（注：ここで N はどうやら n^{*X}_{i} のことのようです。）

出典：マーケティングの数理モデル（朝倉書店）