Multivariate Familial Data の統計解析

統計数理研究所 小 西 貞 則

1. はじめに

例えば，ライフサイクルの短い実験動物を使って，遺伝の生物統計学的研究を行うとする．このとき観測されるデータは，第一世代である親と第二世代に相当するその不特定多数の子（同胞）からなる，家族を一つの単位としたものからである．このように，遺伝学を初めとして，医学，疫学，心理学などの分野においては，種々の家族内特性を統計的に明らかにするため，家族を単位として観測されたデータ（familial data）の分析を必要とする．

特に，遺伝的な要因を探るという観点から考えると，(i) 親とその同胞間の関連性の程度（級間相関），(ii) 同胞内の類似性の度合（級内相関）を定量的に計るための尺度を与えることが重要な問題となる．このような問題に対して従来，ある一つの特性（例えば，血圧，身長，体重，I.Q., 肺活量など）に関する評価尺度の統計的な研究が主に行われてきた（Rosner et al.(1977), Konishi (1982, 1985), Srivastava (1984) 他）．

しかし，各個体がいくつかの特性に関して特徴づけられた多次元データとして観測されたときには，これら複数の特性に関して(i), (ii)に上げた関連性の程度を計る評価尺度が必要となってくる．ここでは，各個体が多次元データとして観測された N 組の家族データ（multivariate familial data）の分析を考察し，複数の特性に関する級間および級内相関の評価尺度の提唱と，関連する統計的推測の問題を検討した．

2. モデル

いま，N 組の家族データを観測し，そのうち α 番目の家族のデータを

$$z_\alpha=(y_\alpha', x_{1\alpha}', x_{2\alpha}', \ldots, x_{k_\alpha,\alpha}')' \qquad \alpha=1, 2, \ldots, N \tag{2.1}$$

とおく．ここに，$y_\alpha=(y_{1\alpha}, y_{2\alpha}, \ldots, y_{p\alpha})'$ は，p 個の特性に関する母親のデータ，$x_{j\alpha}=(x_{1j,\alpha}, \ldots, x_{qj,\alpha})'$ は，q 個の特性に関する j 番目の子のデータとする．実際上 $p=q$ とし，親，同胞とも同じ p 個の特性に関する分析を行う場合が多いと思われる．

(2.1) の z_α は，平均ベクトル $\mu_\alpha=(\mu_m', \mu_s', \ldots, \mu_s')'$，分散共分散行列

$$\Sigma_\alpha=\begin{pmatrix} \Sigma_m & e_{k_\alpha}'\otimes\Sigma_{ms} \\ e_{k_\alpha}\otimes\Sigma_{ms}' & I_{k_\alpha}\otimes\Sigma_s+(e_{k_\alpha}e_{k_\alpha}'-I_{k_\alpha})\otimes\Sigma_{ss} \end{pmatrix} \tag{2.2}$$

の $(p+qk_\alpha)$-次元分布に従うとする．ただし，$e_{k_\alpha}=(1, \ldots, 1)'$, I_{k_α} は単位行列，$A\otimes B$ は，行列 A, B のクロネッカー積とする．このモデルでは，α 番目の親は k_α 匹の子を同時に生み，生まれた子の間には順序を考慮する必要がないという設定を考えている．

同胞数（出生児数）$\{k_1, k_2, \ldots, k_N\}$ は，本来ある確率分布からの大きさ N の標本であり，従って出生児数分布の考察が必要となる．しかし，出生児数分布を同時に考慮に入れた多変量家族データの分析は，推測理論上極めて難しい問題となる．実際問題への適用に当たっては，平均出生児数という一次元の尺度を考慮に入れれば，特に問題は生じない．

2.1 複数の特性に関する級間（世代間）相関

N 家族の多次元データ (2.1) の統計的分析において，まず複数の特性に関して，第一世代と

第二世代の関連性の程度を評価したい．このような問題に対しては，第一世代に相当する N 個の親のデータは，平均ベクトル $\mu_m(p\times 1)$，分散共分散行列 $\Sigma_m(p\times p)$ をもつ確率分布から，また第二世代の $\sum_{\alpha=1}^{N} k_\alpha$ 個の同胞のデータは，平均ベクトル $\mu_s(q\times 1)$，分散共分散行列 $\Sigma_s(q\times q)$ をもつ確率分布から抽出され，$\Sigma_{ms}(p\times q)$ が両世代間の相関の度合を反映すると考える．また，同一家族内の同胞間には，当然何らかの相関があるものとする．

このとき，世代間の関連性の強さを計るための自然な尺度は，正準相関係数であり，従って $\Sigma_m^{-1}\Sigma_{ms}\Sigma_s^{-1}\Sigma_{ms}'$ の最大固有値の平方根 ρ_1 を，複数の特性に関する親と同胞間の関連性の程度を計る指標として用いる．

2.2　複数の特性に関する級内相関

同胞内の類似性の度合を定量的に評価したいとき，観測データ $(x_{1\alpha}', x_{2\alpha}', \ldots, x_{k_\alpha\alpha}')'$ は，平均ベクトル $(\mu_s', \mu_s', \ldots, \mu_s')'$，分散共分散行列 $I_{k_\alpha}\otimes\Sigma_s+(e_{k_\alpha}e_{k_\alpha}'-I_{k_\alpha})\otimes\Sigma_{ss}$ の qk_α-次元確率分布に従うとする．このとき，$\Sigma_{ss}\Sigma_s^{-1}$ の最大固有値 λ_1 をもって，複数の特性に関する同胞内の類似性の度合の評価尺度とする．これは，一般にある一つの特性に関して同胞内の類似性の程度を計るときに用いられる級内相関係数の多次元評価尺度への一般化と考えられる．

3.　推定

前章で定義した複数の特性に関する級間および級内相関の評価尺度の推定問題を考察する．いま，N 家族の観測データ (2.1) に対して，

$$Y=[y_1, y_2, \ldots, y_N],\quad \overline{X}=[\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_N],\quad S_\alpha=\sum_{j=1}^{k_\alpha}(x_{j\alpha}-\bar{x}_\alpha)(x_{j\alpha}-\bar{x}_\alpha)'$$

とおく．ただし，$\bar{x}_\alpha=\sum_{j=1}^{k_\alpha}x_{j\alpha}/k_\alpha$ とする．このとき，Konishi and Khatri (1990) は，分散共分散行列 (2.2) の一般化推定量

$$\begin{aligned}&\hat{\Sigma}_m=(\mathrm{tr}\,B_m)^{-1}YB_mY',\quad \hat{\Sigma}_s=(\mathrm{tr}\,B_s)^{-1}\Big(\overline{X}B_s\overline{X}'+\sum_{\alpha=1}^{N}\omega_\alpha S_\alpha\Big),\\ &\hat{\Sigma}_{ms}=(\mathrm{tr}\,B_{ms})^{-1}YB_{ms}\overline{X}',\quad \hat{\Sigma}_{ss}=(\mathrm{tr}\,B_s)^{-1}\Big(\overline{X}B_s\overline{X}'+\sum_{\alpha=1}^{N}\nu_\alpha S_\alpha\Big),\end{aligned}\tag{3.1}$$

を提唱した．ここに，$\omega_\alpha(\geqq 0)$, ν_α は定数，B_m, B_s は $N\times N$ 非負値定符号行列，B_{ms} は $N\times N$ 行列で，さらに $e_N'=(1, 1, \ldots, 1)$ に対して，$B_me_N=0$, $B_se_N=0$, $B_{ms}e_N=0$, $e_N'B_{ms}=0$ を満たすものとする．

(3.1) 式で与えた推定量は，重み付き偏差平方積和行列に基づく推定量で，ウエート $\{B_m, B_s, B_{ms}, \omega_\alpha, \nu_\alpha\}$ を適当な基準に基づいて選ぶことによって，より有効な推定量を構成する目的でつくられた．また，もしウエート間に

$$\sum_{\alpha=1}^{N}\omega_\alpha(k_\alpha-1)-\mathrm{tr}\,B_s(I_N-D_N^{-1})=0,\qquad \sum_{\alpha=1}^{N}\nu_\alpha(k_\alpha-1)+\mathrm{tr}\,B_sD_N^{-1}=0$$

の関係が満たされておれば，(3.1) 式の推定量は，分散共分散行列 (2.2) の不偏推定量となる．ただし，$D_N=\mathrm{diag}[k_1, k_2, \ldots, k_N]$ とする．

一般化推定量 (3.1) を用いると，複数の特性に関する級間相関の評価尺度 ρ_1 は，$\hat{\Sigma}_m^{-1}\hat{\Sigma}_{ms}\hat{\Sigma}_s^{-1}\hat{\Sigma}_{ms}'$ の最大固有値の平方根 r_1 で推定し，級内相関の評価尺度 λ_1 は，$\hat{\Sigma}_{ss}\hat{\Sigma}_s^{-1}$ の最大固有値 l_1 で推定する．これから，r_1, l_1 を各々級間，級内相関の多変量評価尺度として用いる．

また，l_1 は各個体の q 個の特性に関するデータをそれらの線形結合で置き換え，一次元評価尺度の級内相関係数を最大にするように係数を選ぶことと同値である．

推定量の分布は，多変量正規性およびウエートに関してある種の仮定を置くことによって，漸近的な結果が求まる．これらの結果は，ρ_1, λ_1 に対する信頼区間の構成等に用いることができる (Konishi et al. (1991) を参照)．

なお，Srivastava et al. (1988), Konishi and Khatri (1990) は，モデル (2.1) に基づいて，級間，級内相関行列を $P_{ms}=D_m^{-1/2}\Sigma_{ms}D_s^{-1/2}$, $P_{ss}=D_s^{-1/2}\Sigma_{ss}D_s^{-1/2}$ と定義し，関連する統計的推測の研究を行った．ただし，D_m, D_s は各々 Σ_m, Σ_s の第 (i, i) 要素を，第 i 対角要素にもつ p-および q-次元対角行列とする．

4. ウエートについて

一般化推定量 (3.1) を提唱した目的は，推定論，分布論を統一的に扱うことができるということに加えて，例えば提唱した推定量の平均二乗誤差を最小にするようなウエート $\{B_m, B_s, B_{ms}, \omega_\alpha, \nu_\alpha\}$ を見つけ，有効な推定量を見いだすことにあった．以下は，ウエートの取り方の一例であるが，詳細は Konishi et al. (1991) を参照されたい．

(i) 正準相関：(3.1) 式において

$$B_m=B_s=B_{ms}=D_N-k(N)k'(N)\Big/\sum_{\alpha=1}^{N}k_\alpha, \qquad \omega_\alpha=1$$

とおく．ここで，$D_N=\mathrm{diag}[k_1, k_2, \ldots, k_N]$, $k(N)=[k_1, k_2, \ldots, k_N]'$ とする．

(ii) $\hat{\Sigma}_{ss}\hat{\Sigma}_s^{-1}$ の固有値：(3.1) 式において，(i) で与えた B_s に加えて

$$\omega_\alpha=(N_p-N+1)\Big/\sum_{\alpha=1}^{N}(k_\alpha-1), \qquad \nu_\alpha=-(N-1)\Big/\sum_{\alpha=1}^{N}(k_\alpha-1).$$

とおく．ここで，$N_p=\sum_{\alpha=1}^{N}\sum_{\beta\neq\alpha}^{N}k_\alpha k_\beta\Big/\sum_{\alpha=1}^{N}k_\alpha$ とする．

参 考 文 献

Konishi, S. (1982). Asymptotic properties of estimators of interclass correlation from familial data, *Ann. Inst. Statist. Math.*, **34**, 505-515.

Konishi, S. (1985). Testing hypotheses about interclass correlations from familial data, *Biometrics*, **41**, 167-176.

Konishi, S. and Khatri, C.G. (1990). Inferences on interclass and intraclass correlations in multivariate familial data, *Ann. Inst. Statist. Math.*, **42**, 561-580.

Konishi, S., Khatri, C.G. and Rao, C.R. (1991). Inferences on multivariate measures of interclass and intraclass correlations in familial data, *J. Roy. Statist. Soc. Ser. B*, **53**, 649-659.

Rosner, B., Donner, A. and Hennekens, C.H. (1977). Estimation of interclass correlation from familial data, *Applied Statistics*, **26**, 179-187.

Srivastava, M.S. (1984). Estimation of interclass correlations in familial data, *Biometrika*, **71**, 177-185.

Srivastava, M.S., Keen, K.J. and Katapa, R.S. (1988). Estimation of interclass and intraclass correlations in multivariate familial data, *Biometrics*, **44**, 141-150.