離散変量と連続変量が混在する場合の距離とMTシステム

中西　寛子　　大学統計教員育成センター　特任教授

【はじめに】

・MTシステムとは

基準空間と呼ばれる均一な性質を持つデータより構成される母集団を考える．新たなに発生した標本がこの基準空間に含まれるか否かの判断をするのがMTシステムである．たとえば，健康な人の基準空間を考え，ある人がこの基準空間に含まれるか否かを考える．

MTシステムは，新たなに発生した標本が複数母集団のどれに含まれるかを判断する判別分析と考え方は似ている．判別分析と異なる点は，MTシステムで扱う母集団は1つであり，この母集団と新たに得られた標本間距離を用いて基準空間に含まれるか否かを判断することである．

・従来の研究と本研究との関連性

MTシステムは，提案された母集団と標本間距離を用いて判断するが，離散変量と連続変量が混在しているデータ（混在型データと呼ぶ）における距離についてはあまり検討されていない．たとえば，離散変量を連続変量と見なしてマハラノビスの距離を導出し，母集団と標本間距離を提案したものがあるが，実例への応用を行っているだけであって，数理的評価はなされていない．本研究では，混在型データに対する新距離を提案する．その応用例と長所と短所など提案距離を総合的に議論する．

・データの記述

基準空間を作る母集団に属している標本 $\boldsymbol{w}' = (\boldsymbol{x}', \boldsymbol{y}')$ を考える．ここで $\boldsymbol{x}$ は a 個の離散変量からなるベクトル，$\boldsymbol{y}$ は b 個の連続変量からなるベクトルである．新たな標本$\boldsymbol{w_0}' = (\boldsymbol{x_0}', \boldsymbol{y}_0')$ が基準空間に含まれるか否かを判断する．

Location model（Olkin and Tate 1961）に基づき，各標本を k 個のセルに振り分ける．ここで，k は a 個の離散変量の各々が取りうるケースの数を掛け合わしたものである．$\boldsymbol{x}$ の記述によって各標本は含まれるセルm $(m = 1, \cdots, k)$ が決まる．ゆえに，$z_m = 1$，$z_j = 0$ $(j \neq m)$ となるベクトル $\boldsymbol{z}' = (z_1, \cdots, z_k)$ を標本ごとに考えることができる．

【2．混在型データにおける従来のMTシステム】

・MTS（従来の提案距離）

離散変量の $\boldsymbol{x}$ を連続変量と見なしマハラノビスの距離を導く．つまり，基準空間に属する $\boldsymbol{w}$ の平均ベクトルを $\boldsymbol{\nu}$ とし，分散共分散行列を $\Sigma_{\boldsymbol{w}}$ とすると，次のマハラノビス距離によって標本 $\boldsymbol{w}_0$ が基準空間に含まれるかを判断する．

$$\mathbf{MTS} = (\boldsymbol{\nu} - \boldsymbol{w}_0)' \Sigma_w^{-1} (\boldsymbol{\nu} - \boldsymbol{w}_0)$$

たとえば，性別のような離散変量が与えられているとする．男性(または女性)の割合が情報として与えられるが，もし，男女比がほぼ半々であるならわざわざ離散変量を導入する必要はなく，推定すべき変量を減らす意味で，連続変量のみで同様の式を導き利用するほうがよい．

【3．新たなMTシステムの提案】

・MTD（新提案距離）

母集団 π_i $(i = 1, 2)$ において，標本がセルmに含まれる確率を $p_{im} > 0$ とし，$\boldsymbol{z}$ が与えられたときの $\boldsymbol{y}$ の条件付き確率密度関数を f_{im} $(i = 1, 2, m = 1, \cdots, k)$ とする．さらに f_{im} に正規性の仮定 $(\boldsymbol{y} \mid z_m, \pi_i) \sim N(\boldsymbol{\mu}_{im}, \Sigma_m)$ をおく．Nakanishi (1996) は λ-divergence（Taneja 1987）を拡張し，上で示した正規性の仮定を考慮することにより，パラメータ λ をもつ2母集団距離 D^λ を提案した（紙面の関係で$\lambda = 0$ については割愛する）．

$$D^\lambda = \frac{1}{\lambda}\left[\sum_{m=1}^{k}\left(p_{1m}^{1+\lambda} p_{2m}^{-\lambda} + p_{1m}^{-\lambda} p_{2m}^{1+\lambda}\right)\exp\left(\frac{\lambda(1+\lambda)}{2}\Delta_m^2\right) - 2\right] \quad \left(-\frac{1}{2} \leq \lambda,\ \lambda \neq 0\right)$$

ここで，セル m に対し，$\Delta_m^2 = (\boldsymbol{\mu}_{1m} - \boldsymbol{\mu}_{2m})' \Sigma_w^{-1} (\boldsymbol{\mu}_{1m} - \boldsymbol{\mu}_{2m})$ である．

D^λ を母集団と標本間距離に対応するように書き直す．基準空間に含まれるか否かが不明である標本 $\boldsymbol{w}_0$ に対し，$\boldsymbol{x}_0$ の記述によってこの標本が属するセル m_0 が決まる．$p_{2m} = 1$ $(m = m_0)$，$p_{2m} = 0$ （その他）とし，$\lambda = -1/2$ とすると，証明は省略するが，次の距離が得られる．MTD は 0 から 2 までの範囲に収まり扱いやすい．その他の$\lambda < 0$についても同じような考察が出来るが，本質はλ = -1/2のときと同じである．

$$\mathbf{MTD} = 2\left[1 - p_{1m_0}^{1/2} \exp(-\Delta_{m_0}^2/8)\right]$$

$$\Delta_{m_0}^2 = (\boldsymbol{\mu}_{m_0} - \boldsymbol{y}_0)' \Sigma_{m_0}^{-1} (\boldsymbol{\mu}_{m_0} - \boldsymbol{y}_0)$$

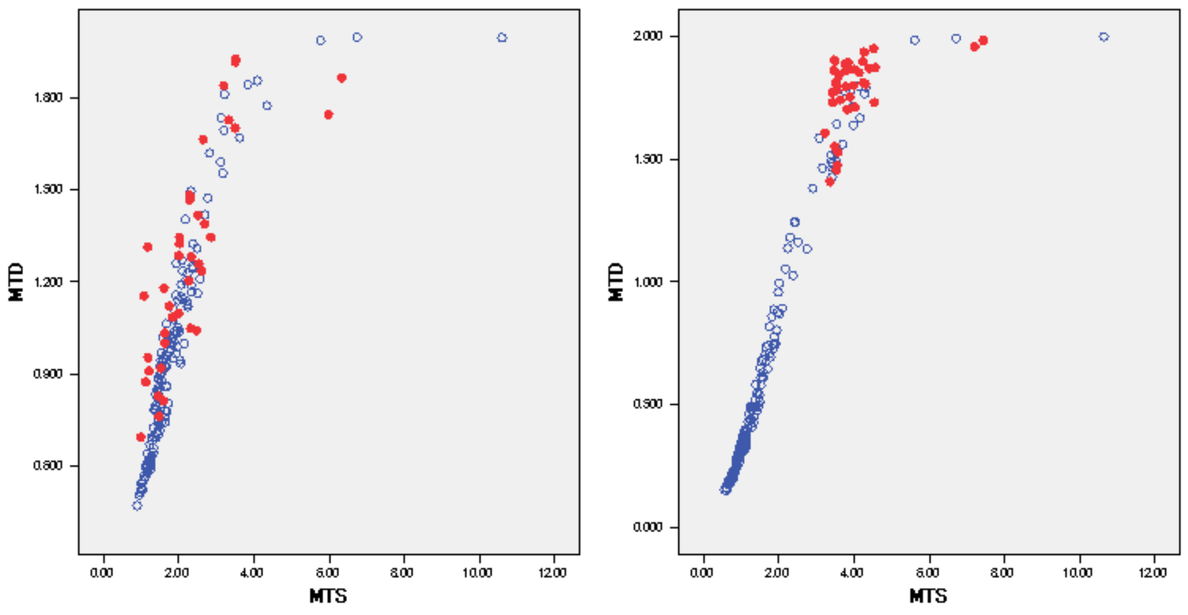
【4．MTシステムの応用】

・虫垂炎（Koepsel他 1981)の例

応用例では，MTD と比較のため MTS の結果を示す．パラメータの推定値は各々，標準的な不偏推定量を用いる．

観測された項目は，性別，壊疽の有無の2つが離散変量であり，年齢，痛みの持続時間，手術までの時間，白血球の4つが連続変量である．179名（内38名が腹膜炎）の虫垂炎の患者のデータを利用した．腹膜炎を起こしていない141名を基準空間とし，腹膜炎をおこしていた38名の MTS と MTD を求めそれぞれを縦軸と横軸にメモリを取り布置した（●印）．また，参考のため腹膜炎をおこしていない患者についても示した（○印）．下図（左）は性別と全ての連続変量を用いた場合の結果，下図（右）は壊疽の有無と全ての連続変量を用いた場合の結果である．

結果より，性別の項目が有効に働いていないことがわかる．従って，Location modelを使うこともない．一方，壊疽の有無は腹膜炎とおおいに関係しているため有効な項目である． MTS と MTD の比較においては，どの程度の値で基準空間より離れていると定義するかによるところが大きく，Location modelを用いることによって腹膜炎の症状のある患者をより明確に見分けることができたとは一概に言えなず，今後の研究が必要である．



Koepsel, T. D., Inui, T. S. and Farewell, V. T. (1981) : *Surgery, Gynecology and Obstetrics*, 153, 508-510.

Nakanishi, H. (1996) : *J. Japan Statist. Soc.* 26, 221-230.

Olkin, I. and Tate, R. F. (1961) : *Ann. Math. Statist*. 32, 448-465.

Taneja, I. J. (1987) : *J. Statist. Planning Inference*, 16, 137-145.