グラフィカルモデルによる欠測のモデリングとその周辺

大阪大学 大学院基礎工学研究科 高井 啓二

データ解析において欠測値の発生は避けられない．欠測があることにより通常のデータ解析が困難になる．そこで，欠測のメカニズムを考慮することが必要になる．ここでは，欠測メカニズムをグラフィカルモデル (GM) で表現することを考える．GM で表現することにより，視覚的に変数間の関係を捉えることができるようになり，解析上有用であると考えられる．しかし，実際には，欠測メカニズムのタイプの中には，GM では表現できないものがあることが分かった．また欠測メカニズムを GM で表現する際に生じる問題についても考えた．

1 準備

グラフィカルモデル (Graphical Model; GM) は，変数間の独立性をグラフに対応させて表現する手法である．GM により変数間の関係を，視覚的に捉えることができる．GM は直観的な視覚と，理論的な確率変数の関係を繋ぐ．ここでは，色々な GM がどの欠測タイプを表現しているのかを調べるとともに，欠測メカニズムのモデリングの道具として，その有用性と限界について考える．

欠測のタイプは，Little and Rubin (2002) などにより三つに分けられている．それは，MCAR(Missing Completely At Random)，MAR(Missing At Random)，NMAR(Not Missing At Random) である．データが MCAR であるとは欠測パターンが完全にランダムであることを言う．データが MAR とは，欠測するか否かが観測されている値に依存していることを言う．MCAR は，MAR の特殊な形である．データが NMAR であるとは，MAR ではないことを言う．MAR かつ本来のパラメタ空間と欠測のパラメタ空間が直積で表わされるときには，欠測メカニズムが無視可能であると言い，そうでないとき欠測メカニズムは無視不可能であると言う．無視不可能な欠測があるときに，偏りのない統計的結論を導くためには，欠測メカニズムをモデリングしなければならない．このモデリングのために，GM は有用な道具となる．

2 GM による MAR と NMAR の表現

MAR・NMAR を GM で表現する．2 変数の場合を考える．変数 (Y_1, Y_2) があり，それぞれの欠測インディケータを (M_1, M_2) とする．また単調な欠測とは，Y_2 が観測されているときは，Y_1 が必ず観測されていることを意味する．非単調な欠測とは，単調な欠測ではないことである．ここでの結果は，3 つ以上の変量の場合にも拡張できる．

2.1 MAR の場合

非単調な欠測で MAR であれば，GM では表現できない．例えば，図 1 のように，M_i と $Y_j (i \neq j)$ の間に辺が存在すると，直感的には MAR である．MAR とは Y_i が欠測するか否

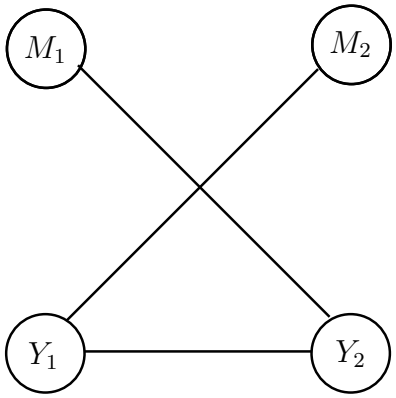

図 1: NMAR:MAR に見えるが NMAR である

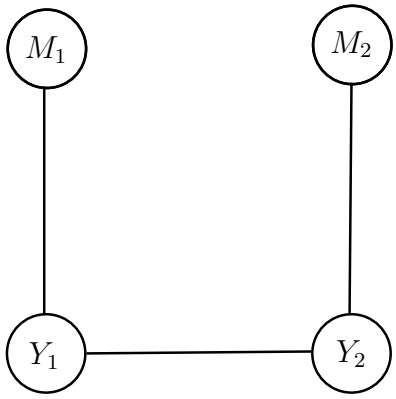

図 2: 欠測構造：Y_1 と Y_2 が独立であると識別性が失われる

か (M_i の値) が，Y_j に依存する欠測だからである．しかし厳密に考えると，図 1 は NMAR を表わしている．MAR のうち GM で表現できるのは，MCAR のみであることが証明できる．

2.2 NMAR の場合

無視不可能な欠測である NMAR の場合には，欠測メカニズムをモデリングしなければならないため，モデリングの道具として GM は重要である．ここでは，図 2 として表される欠測データについて考える．我々は，ここで Y_1 と Y_2 の独立性の尤度比検定を行ないたいとする．

ところが，Ma, Geng and Li (2003) によると，この独立性の仮定の下では識別性が失なわれてしまうことが指摘されている．そこで，識別性を与えるための条件を課した．

この条件の下で，尤度比検定統計量 (LRC;$G_{(2)}$) を構成した．この LRC は，漸近的に自由度 1 のカイ二乗分布に収束することが分かった．また，この統計量の比較対象として，欠測メカニズムを全く考慮しない統計量 ($G_{(1)}$) を構成した．こちらの $G_{(1)}$ も漸近的に自由度 1 のカイ二乗分布に収束する．

上で課した識別性の条件の下では，我々が (推定するのではなく) 事前に与えるパラメタ (a, b) を考える必要がある．この (a, b) を与える定まった方法はない．そこで，シミュレーションによって，適切な (a, b) を探索するとともに，尤度比統計量への影響を調べる必要がある．Takai and Kano (2008) によるアルゴリズムを用いて計算を行なった．その計算結果による概要は，以下の通りである．

(i) (a, b) を正しく特定すると，G の検出力は $G_{(1)}$ の検出力より最大で 10% 以上，上回る．
(ii) (a, b) を正しく特定しないと，G の検出力は $G_{(1)}$ の検出力を最大 5% 程度下回る．

References

Lauritzen, S. L., 1996. Graphical Models. Oxford, Oxford University Press. / Little, R. J. A. & Rubin, D. B., 2002. Statistical Analysis with Missing Data (2nd edition). New York, Wiley. / Ma, W.-Q., Geng, Z. and Li, X.-T., 2003. Identification of nonresponse mechanisms for two-way contingency tables. Behaviormetrika. 30,125-144. / Takai, K. & Kano, Y. 2008. Test of independence in a 2×2 contingency table with nonignorable nonresponse via constrained EM algorithm. Comp. Stat. & Dat. Anal. 52, 5229–5241.