＜修士論文＞

マーケティング・リサーチにおける
統計的因果探索を用いた因果仮説構築に関する研究

(要旨)

滋賀大学大学院
データサイエンス研究科
データサイエンス専攻

修了年度　：　2020年度
学籍番号　：　6019106
氏　　名　：　小西 伶児
指導教員　：　清水 昌平
提出年月日　：　2021年1月17日

1　背景・目的

マーケティング・リサーチの目的は、マーケティング課題の発見や施策の実行などに必要な情報を収集・分析し、企業のマーケティング活動を支援することである。この目的を達成するために、マーケティング・リサーチでは、企業のマーケティング活動と消費者行動の因果関係に関する情報を得ることが重要である。消費者行動の因果関係に関する情報を得るために、マーケティング・リサーチでは一般化線形モデルや構造方程式モデルなどが用いられてきた。しかしこれらの手法は、本来は目的変数の予測を行うものや、因果構造が既知であるという仮定の下で因果関係の大きさを評価するものであり、データから因果構造そのものを推測することは困難な課題であるとされている。

データから因果構造を推測する手法は、統計的因果探索という分野で研究されており、背景理論の不足などにより因果仮説を立てられない場合に活用できることが期待されている。因果構造を識別可能なモデルが複数示されているが、その多くは観測変数が全て連続変数か離散変数のどちらか一方であることが仮定されている。しかし、マーケティング・リサーチでは、商品の購買個数などの計数データ (離散変数) と消費者の感情・意識などの連続変数とが混在しているデータを扱うことが多い。そのため、既存手法ではマーケティング・リサーチで扱う様々なデータにおける因果構造の探索が難しいという課題がある。そこで本論文では、既存の統計的因果探索のモデルを基礎に、マーケティング・リサーチで扱う離散変数と連続変数の両方が混在するモデルを提案し、その識別可能条件と推定法を示す。

2　提案モデル・識別可能性

提案モデルは、加法誤差モデル [1] と 2 次分散関数非巡回有向グラフ (QVF DAG) モデル [2] を用いることによって、以下のように定義される。ここで、それぞれの係数 θ_{jk} は、変数 X_k から変数 X_j への直接的な関係性の強さを表す。

1. p 個の観測変数 $X = \{X_1, \ldots, X_p\}$ は非巡回有向グラフ (DAG) G によって表現されるデータ生成過程から生成されており、各変数の親変数がその変数の直接的な原因である。
2. 連続変数 $X_j (j \in C)$ は、その親変数 $Pa(j)$ と誤差変数 e_j の線形和である。ただし、誤差変数 e_j は、平均 0、分散 σ_j^2 の連続確率変数である。

$$X_j = \theta_j + \sum_{k \in PA(j)} \theta_{jk} X_k + e_j$$

3. 離散変数 $X_j (j \in D)$ は、その親変数 $Pa(j)$ による条件付き確率が 2 次分散関数性を満たす。つまり、以下を満たすような $\beta_{j0}, \beta_{j1} \in \mathbb{R}$ が存在する。

$$\mathrm{Var}(X_j | X_{Pa(j)}) = \beta_{j0} E(X_j | X_{Pa(j)}) + \beta_{j1} E(X_j | X_{Pa(j)})^2$$

また、各変数の条件付き期待値は、その変数の親変数 $Pa(j)$ と任意の単調で微分可能なリンク関数

$g_j : \mathcal{X}_{Pa(j)} \to \mathbb{R}^+$ によって以下のように記述される。

$$E(X_j|X_{Pa(j)}) = g_j(X_{Pa(j)}) = g_j\left(\theta_j + \sum_{k \in Pa(j)} \theta_{jk} X_k\right)$$

2 次分散関数性を満たす確率分布にはポアソン分布や二項分布、負の二項分布などが含まれており、マーケティング・リサーチにおける計数データを表現することができる。また、変数間の関係性については一般化線形モデルと同様であるため、マーケティング・リサーチで用いられてきた従来手法と親和性が高いと言える。

定理 2.1 (提案モデルの識別可能性)**.** 提案モデルは、以下の仮定を満たすとき識別可能である。ここで、π は DAG G における因果順序を表す。

(A) 連続変数が割り当てられた任意の頂点 $j = \pi_m \in C, k \in De(j) \subset C$ のデータ生成過程における誤差変数の分散について、以下が満たされている。

$$\sigma_j^2 < \sigma_k^2 + E(\mathrm{Var}(E(X_k|X_{Pa(k)})|X_{\pi_1}, \ldots, X_{\pi_{m-1}}))$$

(B) 離散変数が割り当てあられた任意の頂点 $j \in D$ について、$\beta_{j1} > -1$ が満たされている。

離散変数の因果順序をは、各変数の 2 次モーメントと条件付き期待値の 2 次式の比が 1 に等しいか 1 より大きいかを検定することで推定できる。一方、連続変数の因果順序は、条件 (A) より、各変数の条件付き分散の大小関係を比較することで推定できる。これらの因果順序の推定法は、離散変数と連続変数が混在していても成立する。そのため本論文による提案モデルは識別可能である。また、提案モデルの識別可能性の証明に基づく推定法は、既存手法と比較して DAG の推定精度が高いことが確認された。

3　今後の課題

マーケティング・リサーチの分野で活用すること念頭に、連続変数と離散変数の両方が混在する構造的因果モデルを提案し、その因果構造の識別可能性と推定法を示した。本論文では理論的な証明に留まっているため、提案モデル・手法を実際のマーケティング・リサーチで得られたデータに適用し、因果仮説構築を行った事例を重ねることで、有用な因果仮説が得られるかどうかを確認する必要がある。また、ゼロ過剰ポアソン分布などのマーケティング・リサーチで頻出する分布を扱うことのできるモデルの開発や、未観測共通原因が存在しないという仮定の緩和などを行うことで、統計的因果探索の手法を用いて効率的に消費者行動の因果仮説を構築できるようになると考えられる。

参考文献

[1] G. Park. Identifiability of additive noise models using conditional variances. *J. Mach. Learn. Res.*, 21(75):1–34, 2020.

[2] G. Park and G. Raskutti. Learning quadratic variance function (QVF) DAG models via overdispersion scoring (ODS). *J. Mach. Learn. Res.*, 18(1):8300–8342, Jan. 2017.