記述長最小性に基づくモデル選択による統計的因果探索の方法論の再解釈

松島　慎（Shin Matsushima）

東京大学　大学院総合文化研究科

介入情報を含まない観察データから因果構造を推定する統計的因果探索は、近年の機械学習研究においても重要な位置を占め、独立性検定に基づく方法、構造方程式モデルに基づく方法など、多様な手法が提案されてきた。しかし、これらの手法を実証研究の道具として用いるときには、手法が要請する統計的・因果的仮定をどのように正当化し、その仮定のもとで得られた確率的な推定結果を、実際の対象についての因果主張としてどのように解釈するのかという困難が生じる。この困難は手法の「入力側の問題」と「出力側の問題」に分けて整理することができる。

入力側の問題とは、分析者が持つデータが手法の要請する統計的・因果的仮定を満たしていると十分に正当化できなければ得られた出力を信頼することが困難になる、という問題である。例えば観察データから因果グラフを得ようとする場合、忠実性の仮定が多くの手法で要請される。例えばLin (2019)は科学哲学的な立場から忠実性の仮定を正当化しようとしているが、因果方向まで一致性を持つ推定を可能にするためには、因果マルコフ性、因果十分性、構造モデルの仮定なども同様に正当化が必要である。

出力側の問題とは、統計理論が与える結論の様式と実践的な意思決定との間に距離がある、という問題である。因果探索手法の保証は、大標本極限で真の構造を出力する確率が１に収束する、すなわち一致性を持つ、あるいは有限標本で誤推定確率が一定以下になる、といった確率的命題として与えられる。しかし実証研究で必要になるのは、「この」データから得られた構造を採用してよいかという一回的な判断である。出力されたグラフを真の因果構造として扱うなら、統計的保証の意味だけでなくモデルの妥当性、仮定違反の可能性を改めて吟味しなければならない。

本発表は、以上を踏まえ、因果探索手法をモデル選択の枠組みの中で解釈する。具体的には、因果探索手法は分析者があらかじめ用意したいくつかの符号化方式のうち観測データを最も短く記述するものを選ぶ、という解釈を採用する（Grünwald, 2007）。この解釈のもとで統計的・因果的仮定は、世界がその通りであることを分析者が保証すべき命題ではなく、分析者がデータを記述するために暫定的に用意した符号化プロトコルとして理解される。

本発表では、この解釈をシャノンの通信路モデルに基づくデータ圧縮の考え方により説明する。符号化プロトコルをデータの送受信の前に共有しているとき、短い記述は、データに含まれる規則性を効率よく捉えたことを意味する。ここで重要なのは、符号化プロトコルをデータの送受信より前に定めることである。逆に、データを受信した後で都合のよい符号を選べば、どのようなデータも短く記述できてしまい、学習理論的な立

場では過学習と呼ばれるべき問題が起きる。この枠組みの下では、入力側の問題は、仮定を本当に満たすかどうかを外部から証明する問題ではなく、どのようにモデルと符号を紐づけるか、どの符号を比較対象に含めるか、またそれらをどのように事前指定するかという問題として再解釈される。また、複雑なモデルの中に単純なモデルが含まれ、かつ、単純なモデルが真の分布を含む場合、記述長最小性に基づくモデル選択では余分な構造をもつモデルではなく単純なモデルを選ぶ性質を持つ。そのため、忠実性の仮定に関しても、非忠実な分布を恣意的に排除するというより候補間の圧縮性能の比較結果として説明できる。

出力側の問題も同様に再解釈される。記述長最小のモデルは、無条件に真の因果構造に接近するわけではなく、与えられた候補モデル集合とその各要素に紐づいた符号化規則に相対的に、最もよくデータを圧縮する因果的表現である。このとき出力の評価は、選ばれたモデルにおいて、データの記述長がどれほど短くなったか、次善のモデルとの差がどれほどあるかなどによって行うことができる。このような定量的評価は、実応用で必要な判断材料を明示化する点で有用である。

最後に、二変数因果探索のための手法である RESIT（REgression with Subsequent Independence Test; Peters et al., 2017）を再解釈する例を挙げる。RESIT は説明変数と残差の独立性検定に基づき因果方向を決定するための手法であるが、記述長最小性に基づくモデル選択として解釈すれば X から Y への構造方程式モデルと Y から X への構造方程式モデルを比較する手続きとしても理解できる。その場合は手法が持つ有意水準パラメータがモデルの符号長と対応するという理論的な整合性が示せるだけでなく、手法が実践上因果方向を出力できない問題も解消できる。

この再解釈は、統計的保証に基づく従来の解釈と対立するものではないことに注意する。すなわち、RESIT のような特定の手法においては一定の統計的・因果的仮定の下で確率的保証を与えるという理論を保持しつつ、その保証が成り立たない可能性のある実応用においても、出力の意味を相対的・定量的に評価する選択肢を与える。また、再解釈によって因果探索の仮定を不要にすることはできないことにも注意する。むしろ、仮定の位置づけを変えることで、分析者が何を引き受け、何をデータに委ねているのかについて別の見方を与えている。その結果、手法によって得られる構造は、分析者が与えた候補の中でデータを最もよく説明するものとして、相対化された概念として理解される。

Lin, H. (2019). The hard problem of theory choice: A case study on causal inference and its faithfulness assumption. *Philosophy of Science*, *86*(5), 967-980.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT Press.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT press.