発表９

科学技術・イノベーション政策研究への統計的因果探索の利活用の新たな可能性

― 博士課程進学率に関する政策論を例として ―

2022年2月18日
第14回政策研究レビューセミナー
文部科学省　科学技術・学術政策研究所
第1調査研究グループ　高山　正行

本日ご紹介する内容

研究内容

アンケート調査とは別途、マクロスコピックな統計データから、**博士課程進学率に関する各種要因の因果関係の全体像、定量的効果を把握**することを目指し、統計的因果探索アルゴリズム"LiNGAM"の利活用を通じた新たなアプローチでの因果推論

政策課題:若手研究者支援・博士人材流動性

- ○ 第1調査研究グループにおける博士人材の様々な調査・課題提起
- ○ 政府でも、第6期STI基本計画に基づき、博士課程大学院生・若手研究者支援をはじめとする、様々な議論・政策手段が存在

×

理研AIPセンターとの共同研究による統計的・数理的アプローチの強化

- ○ LiNGAMの政策科学への応用を通じ、政策要素間の因果関係をデータ駆動的に分析
- ○ 社会物理学的なアプローチの採用による、政策科学における新たなモデル構築

↓

- ✓ 政策立案のためのより正しいロジックモデルの確立、予測精度・信頼度の高い政策シミュレーションの実現を期待
- ✓ 科学技術・イノベーション政策研究のデジタル・トランスフォーメーションへの一歩

※ 本日ご紹介する研究の詳細: 近日学会HPにて公開される(と思われる)以下の発表の予稿を参照。

○高山、小柴、前田、三内、清水、星野
「EBPMと統計的因果探索・数理モデルの利活用」、 **研究・イノベーション学会　第36回年次学術大会　2G02**(2021).

○高山、小柴、前田、三内、清水、星野
「統計的因果探索アルゴリズム"LiNGAM"を用いた若手研究者支援政策に関する研究」、
研究・イノベーション学会　第36回年次学術大会　2G03(2021). ⇐ ベスト・ペーパー・アワード受賞!

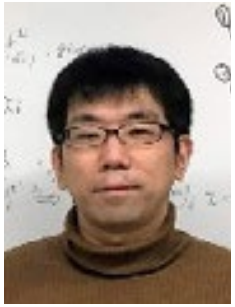
研究体制

高山正行
文部科学省 研究振興局　参事官(情報担当)付
(併)NISTEP　第1調査研究グループ　研究官

小柴等
NISTEP　データ分析政策研究室　主任研究官

清水昌平
滋賀大学　データサイエンス学系　教授
理研AIPセンター　因果推論チーム　チームリーダー
NISTEP　客員研究官

前田高志ニコラス
理研AIPセンター　因果推論チーム　特別研究員
東京大学　空間情報科学研究センター　客員研究員
NISTEP 客員研究官



三内顕義
理研AIPセンター　数理科学チーム　研究員
JST さきがけ　研究員
NISTEP 客員研究官

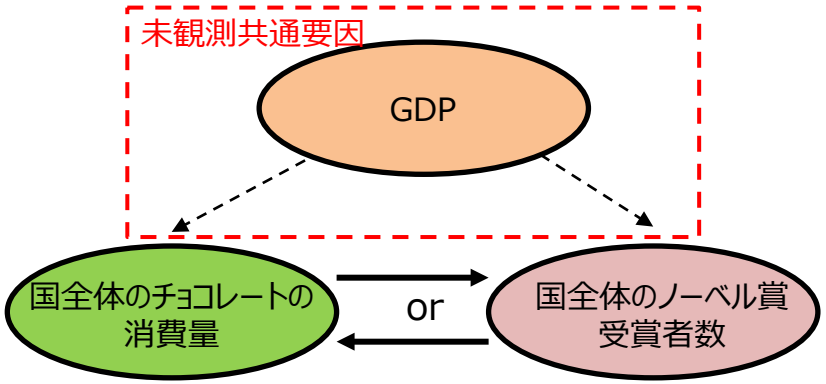
星野利彦
文部科学省　大臣官房付
(併)NISTEP 第1調査研究グループ　総括上席研究官

「因果推論」に関する主要なアプローチ

〇因果推論のポイント:
ざっくり言うと、
関連する変数群を特定し、**定性的に説明可能**かつ**統計的に正しい因果グラフを描写**すること

因果推論に関する有名な例: 各国のチョコレートの消費量とノーベル賞受賞者数には強い相関があるが、
これらの因果関係は?

因果グラフの例:



・複数の要因の間にある、直接の 原因 → 結果　の関係をまとめて描写

・直接の因果効果だけでなく、**間接的な効果**もありうる

・未観測共通要因がないかどうかの検証も重要

・矢印の向き・効果の正負が**説明可能かどうか**考察・検証が必要

因果推論の主要なアプローチ:

〇実験研究 ⇒ ランダム化比較実験(RCT: randomized controlled trial)
無作為に介入群と対照群に分けて効果検証を行う。ただし特に政策研究の文脈では、規模や実験期間、倫理的問題から、このアプローチをとることが困難なことがほとんど。

〇観察研究　⇒ 能動的な介入は行わず、naturalにデータを取得する。
その分、因果関係の説明を強く裏付けるためには統計学的なアプローチが重要。
～**「統計的因果推論」**という形で、数学的枠組みの構築がなされ、取り組まれてきた。

「統計的因果推論」に関する主要な方法論と共通する課題

主な手法	特徴
(通常の)回帰分析	ある変数が他の変数と線形・非線形の関係にあると仮定し、係数や誤差を評価。
差分の差分法	ある時点での介入群と対照群の差について、さらに異なる時点と比べて差をとることで介入効果を評価。
操作変数法 **Cf. 2021年のノーベル経済学賞は この手法の応用研究!**	ある特定の(直接操作ができない)説明変数による目的変数の因果効果を説明するにあたり、他の説明変数を一切変動させないような別の操作変数を導入し、コントロールすることで、因果効果を抽出推定する方法。
共分散構造分析	構造的因果モデル(構造方程式モデル)に基づいて回帰分析を行う。また、観測変数の因果関係とは別途、複数の観測変数を構成要素とする潜在変数(構成概念)を導入することで、因子分析や因果グラフの解釈の簡略化が可能になる。

○共通する課題:

- **因果グラフを一定程度以上仮定しないと、そもそもこれらの分析ができない**。
  妥当な仮定のもとで進めるにあたり、**質的研究による定性的な予想以外に、人力では思いつかないような可能性も効率よく検証**するには、**統計学的な裏付けも加味した上で因果グラフの自動仮説生成**が有効。**⇒統計的因果探索**

- ナイーブな解釈は線形からであるとしても、現実には線形結合の関係でなかったり、非線形性が働きうるため、より正確な現象記述・予測を行うにあたっては工夫が必要
  ⇒ 数理モデル利活用とのサイクル

「統計的因果探索」とは

○ 因果推論にあたり、定性分析等に基づいた事前知識なしでも、統計データのみからデータ駆動的に仮説となる因果グラフを探索したい ⇒ **「統計的因果探索」**

現在、有向非巡回グラフ構造という仮定での議論を中心に、以下のような手法が確立している。

統計的因果探索の主な手法	特徴	対象となる変数
Bayesian Network (BN)	ベイズの定理に基づいて条件付き独立性を評価し、因果グラフを確率により記述。 Judea Pearl, *Proceedings, Cognitive Science Society*: 329-334(1985).	離散変数値や不等式による場合分け等 (一つ一つの場合に対して確率評価をするので、連続変数値には向かない)
LiNGAM (Linear Non-Gaussian Acyclic Model)	(線形の)構造方程式モデルに基づいて、回帰分析と誤差変数の独立性評価に基づき、因果グラフを係数とともに記述。 S. Shimizu, et al. Journal of Machine Learning Research, 7:2003-2030(2006).	連続変数値

○ 取得しているデータの種類によって、アプローチの選択は変わる

○ **連続変数により数理モデルによるアプローチで第一原理的に社会現象を紐解いていく**という意味では、LiNGAMがより有効な手法と考えられる

因果推論・因果探索の位置づけとデータ駆動型の因果推論への期待

統計的因果推論: データに基づいた定量的な因果推論全般を指す。仮説となる因果グラフの描き方は任意。

統計的因果探索: 統計的因果推論の中でも、因果推論を行うにあたっての仮説となる因果グラフ自体を統計的に自動生成。

統計的因果推論

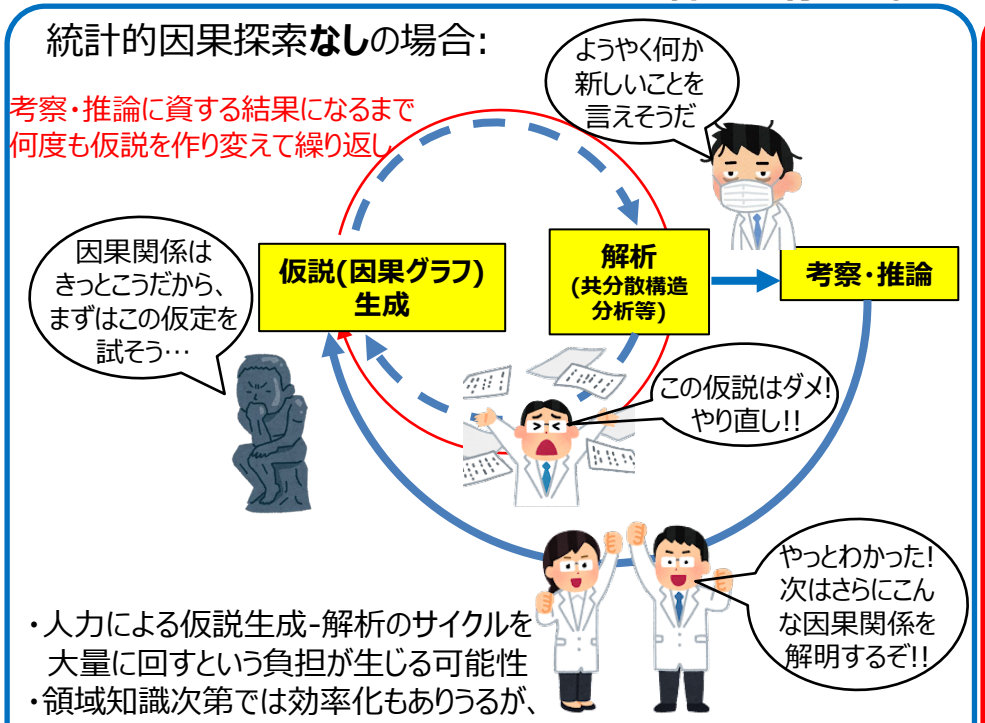
従来の主なアプローチ:
- 回帰分析
- 差分の差分法
- 操作変数法
- 共分散構造解析
- …

統計的因果探索

主なアプローチ:
- ベイジアンネットワーク
- LiNGAM
- …

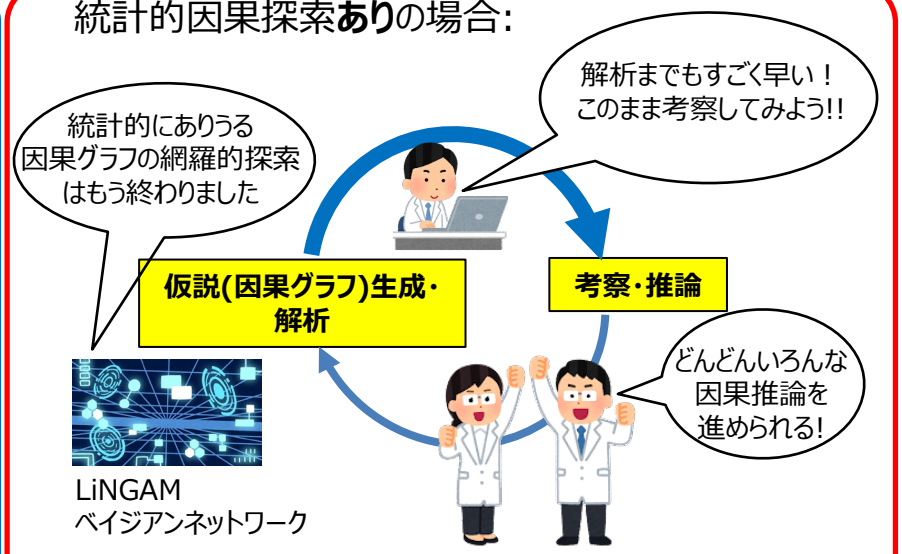
統計的因果探索の有無による因果推論サイクルの比較

統計的因果探索**なし**の場合:



- 人力による仮説生成-解析のサイクルを大量に回すという負担が生じる可能性
- 領域知識次第では効率化もありうるが、どれだけ領域知識の蓄積があるかに強く依存

統計的因果探索**あり**の場合:

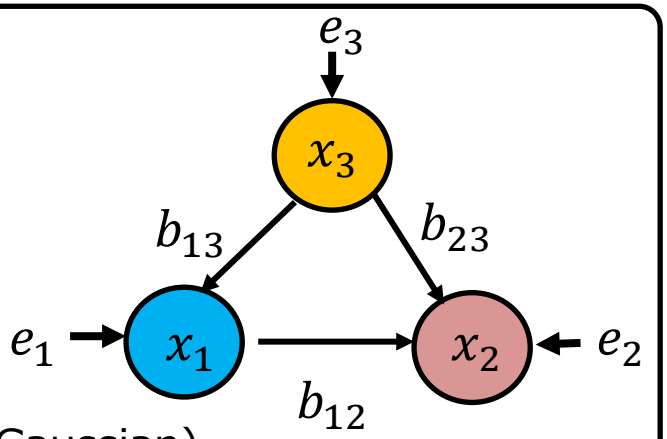


- 統計的に信頼性の高い仮説生成を一度にデータ駆動的・網羅的に行い、場合によっては解析まで自動化
  ⇒ 因果推論の研究サイクルの加速が期待できる
- ただし、このサイクル内で領域知識の利活用をどう進めるか

LiNGAM (Linear Non-Gaussian Acyclic Model) の性質

構造方程式モデル

$$x_i = \sum_{i \neq j} b_{ij} x_j + e_j$$



仮定:
(0) 線形(Linear)
(1) 非巡回性(Acyclic)
(2) 外生変数(誤差変数) e_jの確率分布が非ガウス(Non−Gaussian)
(3) 異なる外生変数同士が互いに独立(未観測共通要因なし)

近年発展形が次々につくられているDirectLiNGAM アルゴリズム:
「回帰分析 + 回帰残差(理想的には外生変数に該当)
同士の独立性の評価」
を繰り返し、残差同士のdependenceが最小
となるような解を出力

S. Shimizu *et al*. Journal of Machine Learning Research, 12(Apr): 1225–1248(2011).
A. Hyvärinen *et al*. Journal of Machine Learning Research, 14(Jan): 111—152(2013).

イメージ($x_1 \rightarrow x_2$が正解の時)

① x_1をx_2に回帰した場合
x_1の外生変数に該当するはずの回帰結果の残差に
x_2の外生変数が残り、dependenceがノンゼロとなるため
上記(3)に矛盾。

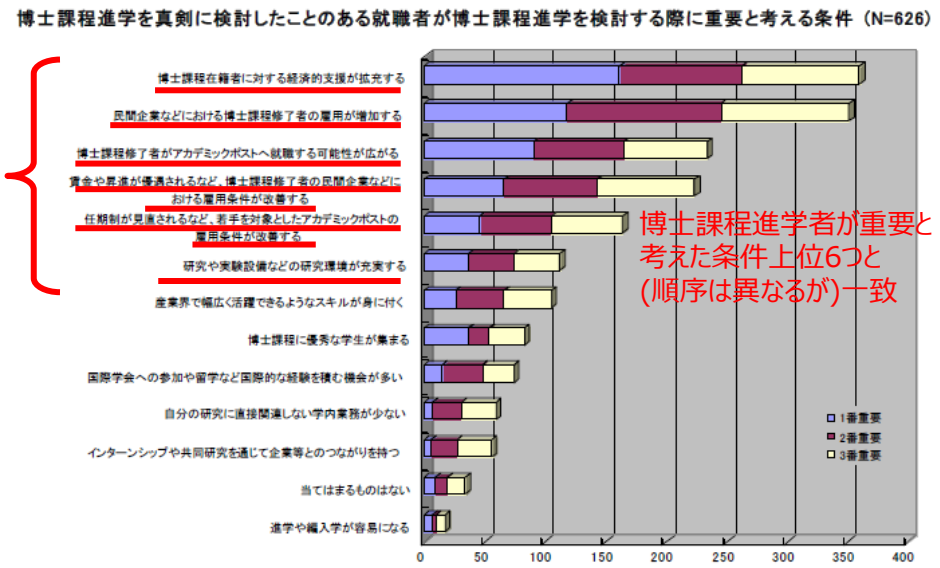
② x_2をx_1に回帰した場合
x_2の外生変数に該当するはずの回帰結果の残差に
x_1の外生変数が残らないので、dependenceもゼロ。

⇒ 因果的順序として$x_1 \rightarrow x_2$が正解であることが計算から分かる。

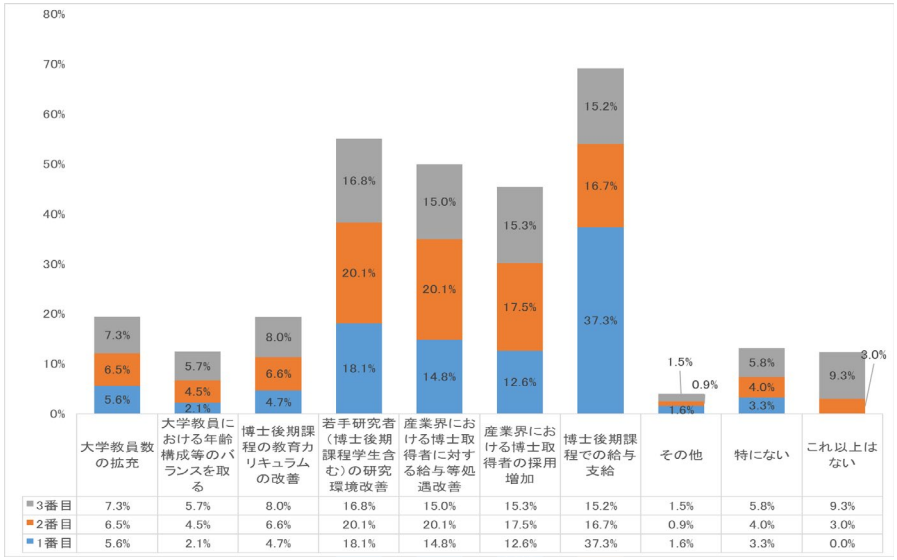
因果の向きを一意に識別可能
(係数行列が一意に定まる)

博士課程進学の重要な条件に関する先行調査・本研究の概要

NISTEP「日本の理工系修士学生の進路決定に関する意識調査」(2009)



NISTEP「修士課程（6年制学科を含む）在籍者を起点とした追跡調査」(2021)



○ いずれの調査でも、**経済的支援・研究環境改善・アカデミア&産業界のポスト拡充、雇用条件改善等の重要性**は、アンケート結果として継続的に示されている

○ 定量化が難しいものも多いが、いくつかの条件についてはある程度定量化も可能

○アンケート調査や政策サイドにおける議論により、とりあえずニーズがあること自体は間違いない
○一方で、「各資源がどれぐらい必要なの?」という問いに答えるには、異なるアプローチが必要。
⇒ 定量的な議論が可能な変数に基づいて、マクロスコピックな統計データから、
博士課程進学率に関する各種要因の因果関係の全体像、定量的効果を把握できないか?

博士課程進学率に関する因果探索にあたってのデータセットのデザイン

公開情報ベースで構築したデータセット

〇目的変数:

博士課程進学率(学校基本調査・科学技術指標より)

〇要因変数

-経済的支援に関わるもの

- DC1採択者数 (JSPS HP等から収集。)
- Global COE +リーディング大学院 + 卓越大学院の予算額(億円) (適宜JSPS HP等から収集。)

-博士課程からのキャリアパスに関わるもの

- 博士修了直後の大学教員就職率 (学校基本調査結果より計算)
- 博士修了直後のポストドクターとしての就職率 (学校基本調査より計算)

-研究環境に関わるもの(NISTEP定点調査の項目を参照)

- 国全体での大学における基盤的経費 (国立大学法人運営費交付金等予算額 + 私立大学等経常費補助金)
- 大学研究本務者数 (総務省 科学技術研究調査)
- 大学研究本務者一人あたりの基盤的経費 (上記2つの値で割り算して算出)
- 研究時間割合 (FTE調査の結果を、適宜線形補完したもの)

〇データ点数:

2006～2019年度の年度ごとで14点。
ただし、ポスドク就職率については2011年度以前は学校基本調査で調査・公開されていない
→欠損値については、平均値で補完

データセットの設定と事前知識(prior knowledge)

変数(単位)	変数名	因果探索する上での事前知識	変数の出典(調査によるものについて)
博士課程進学率	x_0		学校基本調査・科学技術指標
前年度のDC1採択者数(人)	x_1	外生変数	
国全体の大学における基盤的経費(億円) ※1	x_2	外生変数	
大学研究本務者数(人)	x_3		科学技術研究調査
一人当たりの基盤的経費(億円)	x_4	$x_2 \div x_3$	
研究時間割合	x_5		FTE調査を線形補完
博士修了直後の大学教員就職割合	x_6		学校基本調査
博士修了直後のポスドク就職割合	x_7		学校基本調査 (2011年以前は調査されていないため、 2012年以降の平均値で欠損値補完)
DC1以外の経済的支援(億円) ※2	x_8	外生変数	

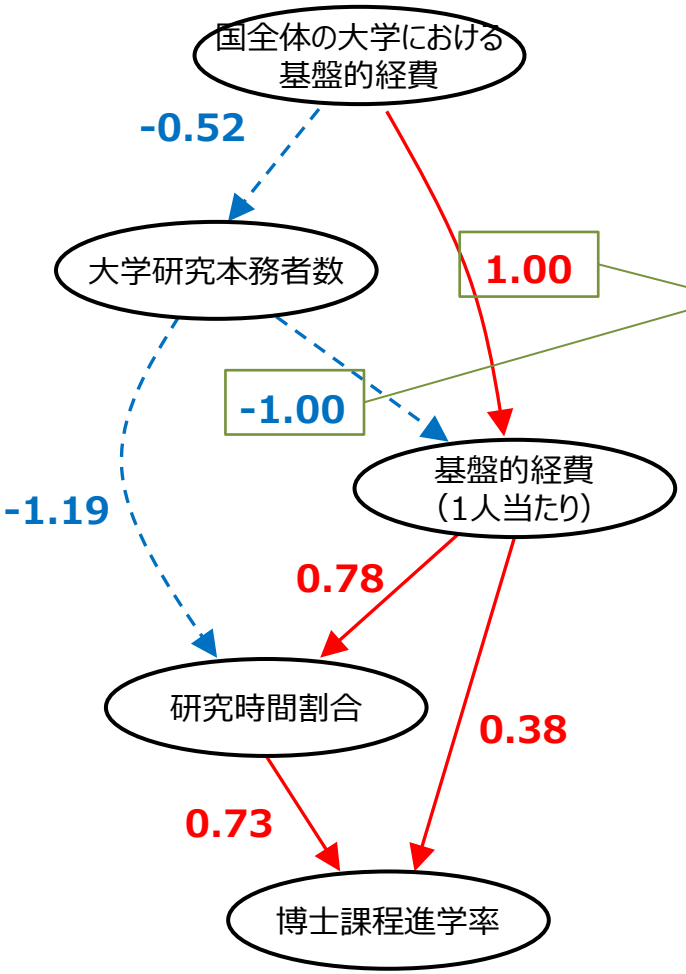
※1: 国立大学法人運営費交付金等予算額 ＋ 私立大学等経常費補助金
※2: グローバルCOE ＋ リーディング大学院 ＋ 卓越大学院 の予算額

注意点
〇系の遅れは考慮していない
例えば 2016年度の博士課程進学者 のうち，X人 が 大学教員 **…ではない**
2016年度の博士課程進学者 XX人，2016年 の 大学教員 XX人 という別々の考え方

〇DirectLiNGAMでは、事前知識なしでも統計的因果探索を行い結果を出力してくれるが、**各種条件設定や既存の領域知識**により、
-変数Aは他のどの変数の影響も受けず、他の変数に影響を及ぼしうるのみ(外生変数)
-変数Bは他のどの変数にも影響を及ぼさず、他の変数から影響を受けうるのみ
といった**prior knowledge(事前知識)**の導入が可能で、明らかにあり得ない因果グラフが出力されることを防ぎ、因果探索の効率化が可能。

〇 通常のLiNGAMでは「和」の構造方程式モデルであるところ、今回は積の関係があるため、**積の構造的因果モデル**に変形。
データセットの各値を全て対数変換の上、DirectLiNGAMで計算。

本研究におけるDirectLiNGAMでの主要な計算結果



大学教員就職割合

ポスドク就職割合

DC1採択者数

COE/リーディング/卓越 等予算

定義式に従って当然現れるもの

○ **直接博士進学率に寄与するのは、**
<u>一人当たりの基盤的経費、研究時間割合</u>のみ。
間接的な寄与まで含めると、
国全体の基盤的経費（正）、大学研究本務者数(負)の寄与がある。

○ DC1やその他の経済的支援、博士修了直後のPD就職率・大学教員就職率は、この計算の範囲では博士進学率に(間接的にも)影響なし。

○ **<u>研究時間割合が、研究本務者数と一人当たり基盤的経費からそれぞれ負・正の影響を受ける</u>**という結果となった。
(研究時間を増やす文脈では、URAやバイアウト制等が議論されてきたが、若手研究者支援の文脈で独立に議論されてきた他の変数との因果関係の存在可能性が示唆された!)

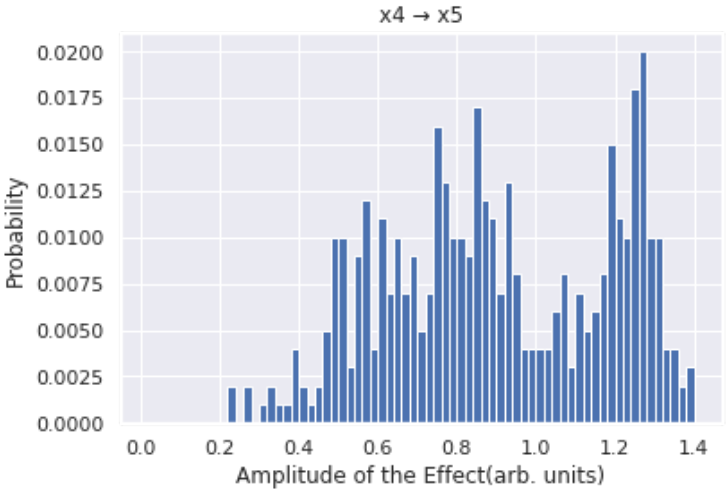
✓統計的因果探索の真骨頂は、**<u>社会科学においてもデータ駆動的に、因果関係に関して人智のみでは不可能な仮説生成が可能になること!</u>** ⇒ **<u>従来の領域知識・手法のみでは到達出来ない斬新な発見</u>**があるかも?

bootstrap法での評価～各有向辺の統計的信頼性

bootstrap法で計算された確率上位15選の因果関係

No.	結果		原因	符号	確率
1	大学研究本務者数	←	研究時間割合	(b<0)	55.6%
2	**基盤的経費（1人当たり）**	**←**	**国全体の基盤的経費**	**(b>0)**	**52.2%**
3	**基盤的経費（1人当たり）**	**←**	**大学研究本務者数**	**(b<0)**	**47.5%**
4	**博士課程進学率**	**←**	**研究時間割合**	**(b>0)**	**43.4%**
5	**研究時間割合**	**←**	**基盤的経費（1人当たり）**	**(b>0)**	**41.9%**
6	**研究時間割合**	**←**	**大学研究本務者数**	**(b<0)**	**39.7%**
7	研究時間割合	←	COE/リーディング/卓越 等予算	(b>0)	21.7%
8	**博士課程進学率**	**←**	**基盤的経費（1人当たり）**	**(b>0)**	**21.6%**
9	大学教員就職割合	←	COE/リーディング/卓越 等予算	(b>0)	19.1%
10	**大学研究本務者数**	**←**	**国全体の基盤的経費**	**(b<0)**	**18.2%**
11	研究時間割合	←	国全体の基盤的経費	(b>0)	16.9%
12	ポスドク就職割合	←	COE/リーディング/卓越 等予算	(b>0)	15.9%
13	博士課程進学率	←	大学研究本務者数	(b<0)	14.5%
14	大学研究本務者数	←	博士課程進学率	(b<0)	13.3%
15	博士課程進学率	←	国全体の基盤的経費	(b>0)	13.3%

「一人当たり基盤的経費 → 研究時間割合」の係数と確率に関する分布



○データセット全体から再標本化を行い統計評価を行うbootstrap法とDirectLiNGAMを組み合わせることで、「原因 → 結果」の各パターンの因果関係が表れる確率や因果係数に関する統計的信頼性の評価も可能

ここまでの結果と課題

ここまでできていること・得られた結果・知見

○ LiNGAMによって、過去14年分程度の統計データから、データ駆動的に因果グラフを探索し、定量的に評価。
○ 因果推論という観点では、研究者一人当たりの基盤的経費や研究時間割合が博士課程進学率に正の影響を与えうる等、先行研究と整合する結果も現れた。
○ 一方で、経済的支援による効果は、因果関係の存在の可能性も含めて先行研究等で期待される結果と異なる部分も現れた。⇒ 未観測共通要因の存在可能性も検討する必要があるが、政策研究の知識に基づいた人力による従来型のアプローチでは現れない示唆

解決できていない課題

○ そもそもの**データ点数**が少ない事と、時間の経過による別のトレンドによる影響が表れうること
○ **分野別の事情の違い**がありうること
○ **遅延効果**がありうること
⇒この結果は直ちに信頼するのではなく、これを基にした派生的な追究が必要。

今後期待される深堀・派生的研究

○ 「修士人材追跡調査」等の個票データを使ったアプローチ
(ただし、データバイアスが存在しうることに注意が必要なのと、変数が離散(2値等)となる可能性が高く、その際はBaysian Networkの利用か、離散変数-連続変数混合の場合のDirectLiNGAMといった技術的拡張に期待。)
○ 分野別のアプローチ(ただし、基盤的経費や経済的支援等には困難な部分も考えられ、慎重な議論が必要)
○ 遅延効果込みでの因果探索(VAR-LiNGAM)
○ 未観測共通要因の因果グラフ上の位置の推定(RCD)
○ 若手研究者支援にあたっての、年齢別の議論への拡張
※詳細は研究・イノベーション学会 第36回年次学術大会 2G03参照。
○ その他、変数の見直し等

EBPMへのLiNGAM利活用に向けて

Q. LiNGAMの利用には特別な環境が必要になるのか?

○ LiNGAMは、S. Shimizuをはじめとする研究者の方々の多大な努力により、
Pythonでの汎用パッケージとして丁寧なtutorialつきで公開されている！
https://github.com/cdt15/lingam

⇒ Python関係の環境構築さえできれば、tutorialに従って非常に簡単な操作で統計的因果探索が可能。

LiNGAMの利用にあたっての注意

○ **丁寧なデータセットの構築(何を変数にとるか・どれだけのデータ点数を集めるか)**
- 強い疑似相関はあるけれども、明らかに因果関係が考えにくいものについては、事前知識(prior knowledge)を入れて縛るか、変数の組み合わせ方を見直す(絞る)か、検討が必要
- (データ点数)/(変数の数) が小さいほど(特に1を下回る場合)因果探索の結果の統計的信頼性は低下していく

○ **背景知識の入れすぎ・変数の選定等の恣意性に注意**
- これを過度にやりすぎると、Policy-Based Evidence Makingでしかなくなる

○**結果の評価**
- 因果探索の結果をそのまま受け入れるのではなく、各有向辺の意味を領域知識に基づいて論理的に議論し、明らかに筋としておかしいものは棄却する等の考察は必要
- 有向辺の確からしさ等、統計的信頼性の評価を怠らない
- 結果の定量的解釈でも、LiNGAMの仮定を満たしているか否か、アルゴリズムの性質も含めて正しい理解が必要

動かすこと自体のハードルは決して高くないが、
誤用することのないよう、リテラシー&各領域知識からの説明性が必須

LiNGAMの派生形・参考文献

特に、DirectLiNGAMについては、以下のように拡張・仮定の緩和が進められており、
様々なニーズに合わせた分析が可能となり、Pythonパッケージにも随時導入されている。
(赤字部分については本発表で紹介した分析においても利用)

○ prior knowledge(事前知識)を加えた上での因果探索
- 明らかに因果関係がある/ない、他の変数から影響を受けない、他の変数に影響を及ぼさない
といった条件付けのもとでの因果探索が可能
(明らかにおかしい因果グラフが出力されることを防げる)

○ bootstrap法(再標本化による統計的推論)に基づく各因果の妥当性の確率評価
- データセットから任意の数だけ再標本化しDirectLiNGAMの実行を行うことで、
特定の因果係数がノンゼロである確率や因果係数の分布を評価する等の分析が可能

K. Thamvitayakul *et al.* In *Proc. 2012 IEEE 12th International Conference on Data Mining Workshops (ICDMW2012)*, pp.659--668, Brussels, Belgium, 2012.

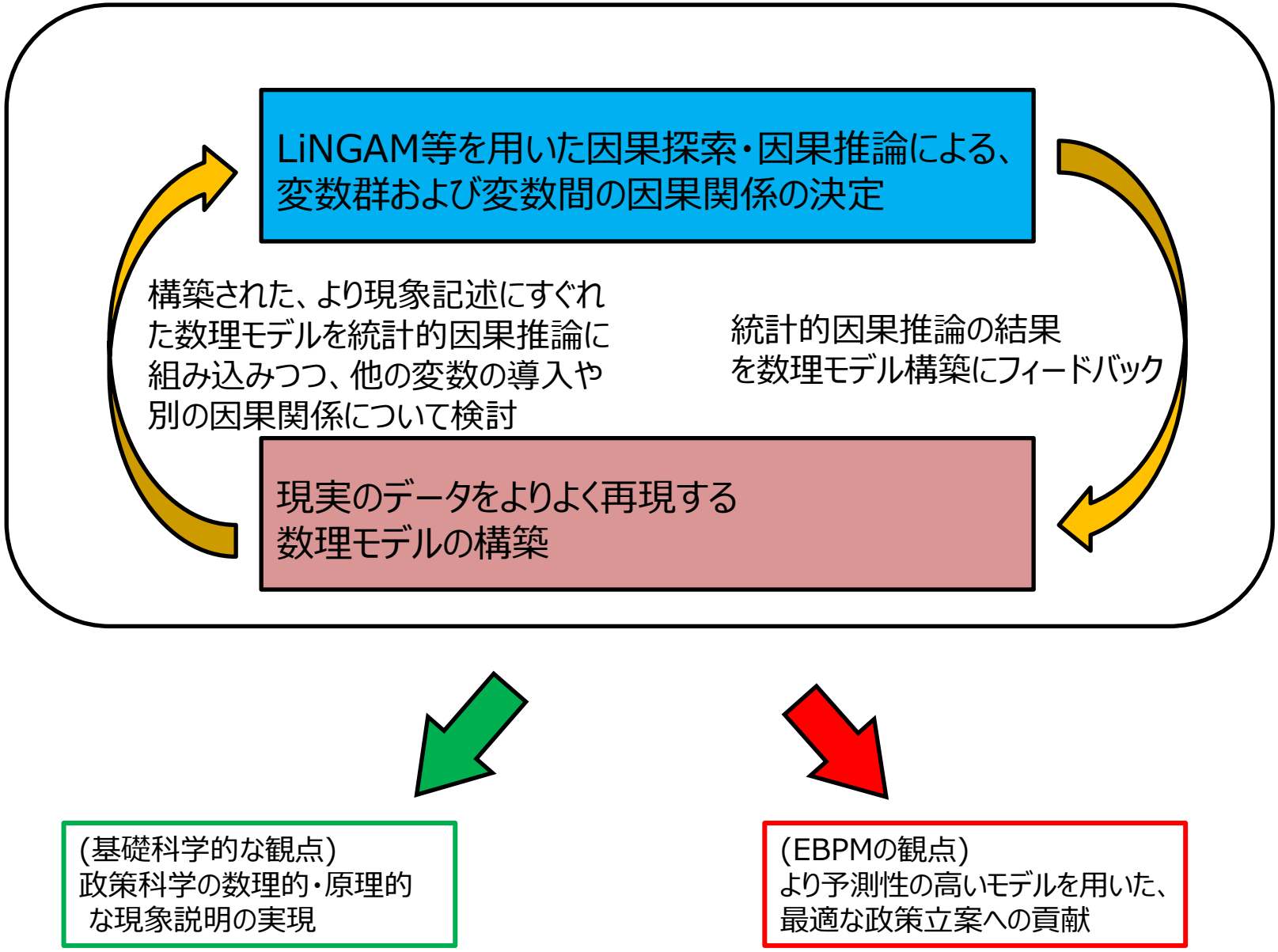
○ 自己回帰モデルに基づいて、遅延時間依存性を考慮したLiNGAM
- 影響の表出に遅延がある場合についても、遅延の長さに応じた影響の度合を評価
⇒ 離散的なシステムダイナミクスの研究にも貢献が期待できる

A. Hyvarinen, K. Zhang, S. Shimizu and P. O. Hoyer. *Journal of Machine Learning Research*, 11:1709–1731, 2010.

○ 未観測共通要因がある場合でも因果グラフ全体の構造を推定する
- RCD(Repetitive Causal Discovery)アルゴリズムにより実行可能

T. N. Maeda and S. Shimizu. In Proc. 23rd International Conference on Artificial Intelligence and Statistics (AISTATS2020), Palermo, Sicily, Italy. PMLR 108:735-745, 2020.

展望: 単純なLiNGAMの政策研究への適用を超えて～
因果探索・因果推論と数理モデル構築の両輪での研究



(基礎科学的な観点)
政策科学の数理的・原理的な現象説明の実現

(EBPMの観点)
より予測性の高いモデルを用いた、最適な政策立案への貢献