計量経済学と機械学習の関係–AI はさだめ, さだめは反事実–

Relationship between Econometrics and Machine Learnings: AI is the Plan, the Plan is Counterfactual

KATAGIRI, Satoshi*

ユリウス暦 2020 年 1 月 6 日

* Twitter Account: @ill_identified

目次

概要

この投稿は, グレゴリオ暦 2019 年 7 月 15 日 (以下, 特に断らない限りグレゴリオ暦法による表記) に開催された第 80 回 Tokyo.R での応用セッションの発表内容を加筆修正したバージョンである. ただし **R の話はほとんどなかった**.

- 『計量経済学と機械学習の関係 –AI はさだめ, さだめは反事実的 – - Speaker Deck』

近年注目を集めている機械学習に対して, 経済学の伝統的な計量経済学 (≃ 統計学) がどう影響を受けているか, また**逆に機械学習がどのように従来の統計学的なアイディアを取り入れているか**について語る. カバーするトピックはかなり広範囲のため, ある程度の知識がないと難しいだろう. こちらもなるべく簡易に書くよう努力するが, とはいえ高度な話題に対してはそれなりの前提知識を要するのも事実である. 想定読者は, 機械学習か計量経済学の基本的なトピックを知っている (例えばいずれかの標準的な教科書を読んで内容をある程度理解している) 人間である. ただし参考文献リストを多く挙げているので, それらを地道にたどればわからないこともない.

また, 発表から時間があいたため, いくらか加筆・変更をしている.

サブタイトルは AI と愛を掛けただけで (Tiptree, 1973), **本文の本質を捉えてうまいこと言ってるということは特にない**[*1].

参考: 過去の記事の中でも特に長いものとして, causal impact , 三国志はいずれも pdf 変換すると 30 ページに少し届かないくらいだが, **今回は 50 ページ以上ある**. そして前 2 者はいずれも画像がかなり多いが, 今回はあまりない. 参考文献リストが全体の 1 割くらいなので体感ではさほど長くないと思う.

1　イントロダクション

以前, 第 79 回 Tokyo.R で, 学生から,「計量経済学と機械学習は補完しあえるのか, 競合するものなのか」という遠大な質問をもらった. 一言で答えるのは難しいというか, 私はおろか最先端の研究者でも答えるのは難しい問いだと思う. 確かに以前, 私は両者を対比して違いを表す内容の記事を書いた.

- 2015 年: 『ディープラーニング VS ディープパラメータ』
- 2017 年: 『計量経済学と機械学習の違い』

上記 2 つの記事で私が主張したかったことは要約すると以下である.

- 機械学習は未来においてもデータの生成過程が斉一であることを暗に仮定している. よって, 構造の変化までは考慮されていない, 重回帰であろうがディープラーニングであろうが, 構造の変化を捉えるパラメータまで特定できなければ, 無意味である.
- 機械学習は計量経済学に部分的に影響を与えつつあるが, 逆に機械学習は経済学のようにデータ構造の特定を考慮しないために当てはまりを悪くする場合がある

ただし, 今見返してみると, 当時は考えが浅く, いまいち伝えたいことの本質を伝えきれていない表現が目立つ.

[*1] たしかこのタイトルは発表前日の深夜のノリで付けた.

私が厳密に正確に何かを書こうとすればするほど学術論文めいた体裁になってしまう. つまり前置きがとても長くなってしまう. しかしこのブログは一般公開されており, また学術論文でもない (私自身も職業研究者ではない) ので誰でも読めるようにしなければならない. 例えば簡単に日本語で「計量経済学 機械学習」でググると私の書いたもの他には以下のようなものがヒットする

1.『機械学習と計量経済学 - SlideShare』(2017)
2.『経済学徒が知らない機械学習の世界 | AI tech studio』(2018)
3.『Academic 機械学習と計量経済学の違いとは？- Block Genius』(2018)

これらを確認してみると, おおまかに「計量経済学と機械学習の互いに欠けているツールを輸入しあえる」(1)「機械学習は『予測』をすることで計量経済学は『説明』することだ」(2, 3) などと書かれており, 大筋の主張としては私の考えと対立していない.

しかしながらこれらの資料では具体的にどう違いが生じるのかには深く立ち入っていない. また, ここ**数年の機械学習分野の流行と発展に伴い, 私の書いたものも含めてここで提示していた事実そのものが時代遅れになっている**面もある. ここ数年で, Varian (2014), Mullainathan and Spiess (2017), Athey (2017, 2018), Athey and Imbens (2019) といった経済学者によるすぐれたエッセイ・レビューがいくつも公開されている. そこでこれを機にもう一度真面目にこの問題を論じてみたい.

しかし, そもそも英文を読むことに何の抵抗もない人間や勤勉な研究者には多くは既知の話であり, そして日本語でも 2019 年 11 月の『経済セミナー』711 号に掲載された 依田 (2019) の評論で簡潔に特徴をまとめられている.

私が付け入る隙があるとすれば, これらは経済学を学ぶ人間に向けて書かれたものだということだ. 私も経済学寄りの視点で書いているが, 加えて**機械学習側の最近の応用研究について掘り下げることで差別化をはかったつもり**だ[*2].

カバーする範囲は多岐にわたり, 無節操に計量経済学や機械学習のトピックをかいつまんで両者の視点から見比べているが, 1 つ 1 つのトピックを基本的な用語の定義から解説していると, 膨大な量になってしまう. よって, 厳密な解説は省き, どうしても細部をごまかした, 直感的な解説になってしまう. よってある程以上勉強した人間ならば私の言わんとしていることを理解し同意なり批判なりしてくれるだろうが, 専門知識のない人にとっては誤解を招く可能性もある[*3]. そこで, **どこまでが先人たちの積み重ねた研究の成果でどこまでが私の勝手な妄想なのかをできる限り明確に区別できるよう**に, より厳密に議論している論文や教科書の引用と紹介をなるべく詳細に書いている. そしてこれらの多くは英文なので, それが読めない/読みたくない人のためにも, 日本語で書かれたもの, あるいは無料公開されている文献のうち有意義なものものできる限り紹介するようにした.

この発表自体は 2019 年 7 月のもので, 当時のスライドは冒頭に挙げたリンク先で公開されている. しかし, 発表からこの原稿の公開まで時間があいてしまったため, かなりの部分を加筆・変更している. 主な箇所は以下の通り.

[*2] 万人に向けた内容であるので, 予防線を貼っておく. 私は一貫して, 統計学, 計量経済学, 機械学習のどれが偉いかなどという主張はしていない. 本文の主張を見れば分かるように, これらの形式的な分類・差別はナンセンスである. 言いたいのはそれぞれに使うのが適切な場面とそうでない場面があり, その使い分けを知る必要がある, ということだけだ.

[*3] 最初は基本的な概念の共有も丁寧にやろうとおもっていたが, テーマが広がりすぎたので諦めた. 最低限, 統計学か機械学習の基礎的なことを知っている必要がある. もしかすると個別のトピックで印象に残ったものは別の機会に再度細かく説明するかもしれない.

- 全体的な構成を見直し, 論題をわかりやすくした (ただし抽象さが増した気がする)
- スライドではいい加減な説明になっていた内的・外的妥当性と機械学習関係, そして最近の機械学習の新たな取り組みの再評価について私の考えをより明確に書いた
- 7 月以降に発表された研究・文献についても改めていくつか取り上げた. 例えば:
  - 2019 年 10 月の Athey and Imbens (2019)[*4]
  - 2019 年 12 月の NeurIPS で行われた CausalML workshop で発表された研究
  - Deaton and Cartwright (2018) の RCT 利用への批判[*5]と Imbens (2018a) による反論
  - 赤池 (2008), Akaike (2010) の「納度」概念[*6]
  - 2019 年 11 月の『経済セミナー』711 号の機械学習特集
  - Pearl (2009), Pearl and Mackenzie (2019) の文脈での, いわゆる「Pearl 流因果推論」と, Imbens (2019) の解説
- ところどころに「脱線した話題」というセクションを設けて関係ありそうで関係のないポエムや雑学を書いた
- その他神経質な人間しか気にしない内容を脚注に書いた

今回の原稿の残りの部分の構成は以下の通り: 第2節では, 計量経済学と統計学の共通点の多さを挙げた上で, 計量経済学と統計学が因果推論を重視し, 一方で機械学習はそうではないことを強調する. 第3節では, これまでの計量経済学の限界や, 方法論の批判について紹介し, それと比較する形で機械学習の特徴と, 限界について言及する. 第4節では, 第 3 節を踏まえて, 計量経済学が機械学習をどう利用して研究を行っているのかを紹介する. 第5節では, Judea Pearl の思想をベースに, 逆に機械学習について深く踏み込み, 機械学習の分野でも新たな研究の流れがあることを紹介する. 第6節では, 逆に機械学習内の研究のうち, 経済学者の興味を惹きそうな最近の研究をいくつか挙げていく. 第7節は, 「計量経済学と AI」というテーマで, 前節で紹介した RCT や自然実験を用いた因果推論とは異なるもう 1 つのアプローチ, 動学的構造推定を紹介し, 数年前に話題になった碁の人工知能 Alpha Go と動学的構造推定が同値であると主張するIgami (2018) の主張を取り上げる.

第8節は, まとめである: 従来の統計学・計量経済学と機械学習の差異とそれぞれの限界を浮き彫りにした上で, 両者に対立構造があると考えるのはもはや時代遅れであることを指摘する.

たとえば既に基本的な知識を身に着けていて機械学習を応用した新しい手法に興味があるのなら, 第 3 節と第 4 節を読めば良いだろう. あるいは機械学習の応用研究のうち, 経済学の知見を活かす可能性が見られるトピックを知りたいのなら, 第 6 節から読むのが良いだろう.

2 再論: $\hat{\beta}$ vs. $\hat{y}$ という両者の古典的対立構造

本題の前に, 統計学と計量経済学は同じなのかという問題について補足する必要がある. 理論面では, 計量経済学は数理統計学に立脚した応用理論だと私は考えている. つまり,

- 漸近性: サンプルサイズが十分大きい時, 推定値 or 分布が真のものに収束するか (つまり, 大**数の法則**や**中心極限定理**のこと)
- 十分性: サンプルサイズの大小に関係なく, データから得られる情報を常に無駄にすることなく活用で

[*4] ただし原稿はそれ以前から ArXiv で公開されていた. 私が単に見落としていただけである.
[*5] 正直に言うと, 発表時点ではこの論文に少し目を通しただけだったので正確にまとめられる自信がなく, 適当にごまかした.
[*6] この論文自体は以前からあったが, 最近になって機械学習工学研究会の slack チャンネルで紹介されて知った

きているか (**十分統計量**アプローチ)

- 効率性: 比較的, あるいは絶対的に推定の誤差が小さいか (*BLUE*, *Cramer-Rao* 不等式)

といった観点から, 推定手法が適切なものであるかが評価される点で共通している. よって, 形式的に[*7]は多くの点で計量経済学は統計学と共通していると考えられ, 以降では特に断りがない限り統計学と計量経済学を同一視して扱う.

一方で, 機械学習の, 特に教師あり学習と呼ばれる部分は, 表面的には計量経済学と共通点が多い. ロジスティック回帰も重回帰も計量経済学で昔から使われている. 機械学習の有名な教科書 "PRML," 邦題『パターン認識と機械学習』(Bishop, 2006) や『統計的学習の基礎』(Hastie et al., 2009)[*8]でも紹介されている**マルコフ連鎖モンテカルロ法** (MCMC) や**ノンパラメトリック回帰** (カーネル平滑化) も計量経済学で使われている. そもそも後者のタイトルは**統計的学習** (*statistical learning*) であり, 機械学習と統計学もまた卑近であることを示唆している[*9]. **ロジスティック回帰**, **線形回帰**, あるいは**決定木**やその発展的アルゴリズム (**ランダムフォレスト**や**勾配ブースト木**), さらには**ニューラルネットワーク**もすべて, **数学の問題としては損失関数の最小化問題を解くこと**である. 他方統計学では決定木系のアルゴリズムはあまり使われないが, よく使われる最尤法は, 尤度関数の最大化問題を解くことであり, ロジスティック回帰や線形回帰は最尤法としても定式化することもできるから, やはり**最適化数学の問題**としてとらえられる.

また統計学では**統計的決定理論**と呼ばれる分野がある. ここでは, 統計学をリスクを最小化する問題という観点で方法論を研究しており, **最尤法や最小二乗法や仮説検定の性質を知ることができる**[*10].

クラスタリングや最近傍法などの**教師なし学習** (unsupervised learning) もまた最適化問題として定式化されるが, 問題のカテゴリとしてはやや異なる. 機械学習が流行したことによる経済学研究への影響について述べた Athey (2018) でも, 教師なし学習の意義について多少述べているが, あまり大きな扱いではない. Mullainathan and Spiess (2017), Merler (2018) では最近の経済学研究における機械学習の応用例がいくつか紹介されている. 最小二乗法だから計量経済学 (統計学) 的, ランダムフォレストだから機械学習的, というようにアルゴリズムの形式的な同一性や, 研究の歴史的経緯で分類することはもはやナンセンスだろう.

よって以降は, **(教師あり) 機械学習と計量経済学が, 最適化ルールを決定するという形式的な共通点以外で何が違うのか**を考えていく.

2.1 計量経済学と機械学習の違いはどこに

では, 違いをどう見つければよいのだろうか. セクションタイトルにも取り入れた "$\hat{\beta}$ vs. $\hat{y}$" という象徴的な見出しは, Mullainathan and Spiess (2017) の引用で, これは計量経済学と機械学習の違いがそれぞれパラメ

[*7] 「形式的に同じ」とは数学的に同型関係があるということで, 類似する数学の問題の解き方を知っている人に問題の構造を連想しやすくするという程度の意図である. ドーナツとマグカップは位相幾何学的に同じである, と言っているようなものである.

[*8] ここでは厳密さと個人的な文献データベース管理の都合から英語の原著を引用したが, 教科書としての完成度は 2014 年に出版された日本語訳のほうが上だと私は考える. 日本語訳は定価 1 万円を超える高額な書籍である一方で, 原著は現在無料でダウンロードできるが, それでも日本語版をおすすめする. 原著が分厚く表紙が茶色と黄色ため一部の人間から「カステラ」と呼ばれているが, 邦訳は別にカステラっぽくない. そもそも原著を出しているシュプリンガー社の教科書はみんなこんな色である. 一体誰が言い出したのか. 英語ではこの表現は見られないので日本独自の呼び方であると思う.

[*9] Hastie et al. (2009, 序章) によると, 著者らは自らを「学習」の研究もしてきた統計学者であり, 数年の研究の結果, 統計学やデータマイニング, 人工知能とも横断的な分野であると考えるようになったことが分かる. さらに言えば, 『パターン認識と機械学習』も全編を通してベイズ統計理論の文脈で説明されている. そういう意味では「純粋な機械学習」なるものを見つけるのは難しい.

[*10] 適切な参考文献を挙げるのが難しい. おおまかな理論を知るには久保川他 (1993) や竹村 (1991) が良いだろうか. 絶版になった本だが.

ータと予測値の推定という目的の違いにあることを浮き彫りにしている. 統計学と機械学習の差異を特徴づけたHarrel (2018) によるエッセイがある.

これは西田氏が翻訳を公開している[*11].

- 統計のモデルと機械学習のモデル、どう使い分ければよいのか

もう 1 つは, Hernán et al. (2019) による, 従来の統計モデルや機械学習・ディープラーニングモデルと**因果推論**の違いを解説するエッセイがある. これもやはり, 西田氏が日本語での要約を公開している.

- 予測と因果関係は何が違うのか - Part 1
- 予測と因果関係 - Part 2: 予測は自動化できても因果推論は自動化できない

Hernán et al.は分析アプローチを**記述** (*description*) **予測** (*prediction*) **反事実予測** (*counterfactual prediction*) という 3 つに分類している. これらの評論は, 従来的な統計モデルと機械学習の差異をうまく特徴づけられていると思う. **反事実予測**とは, **因果推論**に他ならない. これらは簡単な論説にとどまっているので, 今回はこの問題をもっと深く解説していく.

まずは因果推論について具体的な話をする. 経済学者で, 現在は Google の Chief Economist を務めている**ハル・ヴァリアン** (Hal Varian, 2014) が例に挙げる古典的な相関と因果の混同例を紹介する.

- 「犯罪発生率の多い地域には警官が多く巡回する. よって**警官の多い地域は犯罪発生率が高い**という因果関係がある.」

地域ごとの犯罪発生率と警官配置数のデータを, **機械学習の手法** (機械学習を多少知っている人ならば, SVM (SVR), LASSO, Random Forest, あるいは Gradient Boost Decision Tree; XGBoost, Light GBM, CatBoost, HotPot など, 好きなものを思い浮かべれば良い) **で学習させて得られるのは, 両者の相関関係である**. 追加でもう 1 人警官を配備した時, 犯罪発生率がどう変わるかという問いには答えてくれない[*12].

さらに, 計量経済学者**スーザン・エイシー** (Susan Athey, 2017) の挙げる例を要約すると以下のようになる.

- 「*eBay* 社は広告のクリックを説明変数に, 販売額を目的変数とした予測モデルを作成して, 広告投資に対するリターンを知ろうとした. 予測モデルによれば**リターンは *1400%* と見積もられたが, 因果効果を考慮して, 著者が独自に見積もった結果では*-63%*** となった. 商品の広告をクリックする消費者は広告を見ても見なくても, 元から購入を決めていることが多いからこの違いが発生したと考えられる.」

さらにもう 1 つ, グレゴリオ暦 2019 年 1 月より Netflix で配信しているアニメ『空挺ドラゴンズ』の原作 1 巻での例を図 1に示す.

[*11] この内容に関しては, 個人的にはツッコミを入れたい箇所がいくつかある. 翻訳に問題があるという意味ではなく, むしろ原文の記述についてである. 上記では翻訳されていないが, Editorial Comment でもいくらかの批判意見が書かれている. それらを読むのも面白い. が, それは今回の話題に関して本質的に関係ない.

[*12] この記事に興味を持った人間の多くは, 回帰分析という言葉をご存知だろう. 回帰分析という言葉を考案したのはフランシス・ゴルトンである. これは観察データだけで判断したという点から見れば, 相関分析を因果効果と混同してしまった典型例と言えるだろう. また, Box (1966) は回帰分析 (相関分析) の結果を因果関係とみなすことの問題を古くから指摘した好例である.

図1　相関と因果の混同; 桑原太矩『空挺ドラゴンズ』講談社, 1 巻 20 ページ (右) および 21 ページ (左) から

2.2　計量経済学での応用: Rubin 流の因果推論

「追加でもう 1 人警官を配備した時, 犯罪発生率がどう変わるか」「予測モデルと因果効果とでは投資のリターンの見積もりが全く変わる」経済学者は古くからこういった因果推論的な考え方に注意を払ってきた. 先ほど紹介した Hermán らは「既存の統計モデルや機械学習 vs 因果推論」という切り口で比較されていたが, 因果推論的な方法を駆使することは, 計量経済学の重要な特徴である. もちろん全てが因果推論というわけではないのだが, 計量経済学の特徴を端的に表現しろと言われれば, 私は「**計量経済学は因果推論を考えるための統計学である**」と答えるだろう. そこで, 以降では一旦, 機械学習の「予測」という機能が相関関係を見ており, 計量経済学は因果関係を見ているという前提で, 計量経済学でその証拠となる例を挙げていく.

まずは経済学の領域で研究されてきた典型的な因果推論について少し紹介する. 計量経済学者**グイド・インベンス** (Guido Imbens) へのインタビュー (Imbens, 2018b) によれば, 経済学者の間で因果推論に基づいて実証研究するのが加速したのは, 90 年代にプリンストン大学に在籍していた労働経済学者たちのグループ (図 2) によってである. 彼らは後述する**自然実験** (*natural experiments*) というフレームワークを開拓した. 彼が挙げた経済学者は, アングリスト (Joshua D. Angrist), アッシェンフェルター (Orley Ashenfelter), カード (David Card), クルーガー (Alan B. Krueger), そしてラロンド (Robert Lalonde) である. 彼らの研究は, **差の差推定法** (*DID; difference in differences*) での最低賃金の政策効果の分析 (Card and Krueger, 1994), 兵役の抽選を利用した兵役賃金プレミアムの分析 (Angrist, 1990), 以前紹介した **非連続回帰デザイン**[*13] (*RDD; regression discontinuity design*) による少人数教育の自然実験 (Angrist and Lavy, 1999), 双子のデータを取ることで擬似的な RCT とした研究 (Ashenfelter and Rouse, 1998)[*14], そして, そもそも自然実験アプローチ研究の引き金を引いた, 今では当たり前となった因果推論を考慮しない研究の問題を指摘した研究 (Lalonde, 1986) も重要である. このように, 初期の因果推論的な実証研究アプローチは 90 年代に集中している. もちろん, ジェ

[*13] 回帰不連続計画, 分断回帰デザインなど, 未だに訳語が定着していない気がする.

[*14] 世間的には, Ashenfelter の研究で最も有名なのは『その数字が戦略が決める』(Ayres, 2007) で紹介されたワインの価格予測の論文 (Ashenfelter, 2008) かもしれない.

図2　Imbens (2018b) が挙げる, 因果推論を開拓した経済学者たち: 左から Angrist, Ashenfelter, Card, Krueger, Lalonde.

肖像転載元: J. D. Angrist, Orley Ashenfelter, David Card, Alan B. Krueger, Robert LaLonde

図3　Donald Rubin (肖像転載元)

ームス・ヘックマン (James Heckman[*15]) やアンガス・ディートン (Angus Deaton[*16]) のような人物の研究; Heckman and Urzúa (2010), Deaton and Cartwright (2018) も忘れてはいけない.

これらは**平均処置効果** (*ATE; Average Treatment Effect*) を推定する研究と言える. ATE は**ドナルド・ルービン** (Donald Rubin, 1974, 1990b, 図 3) が元になって定式化されたものである. Rubin の提案する因果推理論は, **潜在効果** (*potential outome, potential response*) アプローチ, またはそのまま「**Rubin の因果モデル**」または推測統計学黎明期の統計学者, **イェジ・ネイマン** (Jerzy Neyman) の名を取って「**Rubin-Neyman 因果モデル**」ともよばれるが, 最初の潜在効果アプローチという名称が最近は多く使われる気がする.

Rubin の考え方はいわば, 欠損値の理論である: ある処置 D を行った時, 結果にどう影響するかを表すのに, 結果変数を $Y(D)$ と表す. ここでは, 具体的にイメージしやすいように, 学生が補習を行い, テストを受けた結果得点がどうなるか, という問題を考えてみる. $D=1$ が補習 (**処置群**, *treatment group* という) あり, $D=0$ が補習 (**対照群**, *cotronl group* という) なしとする. このような D を**割り当て変数**または**処置変数**という. データから平均的な D の効果を知りたい場合, 単純に考えれば, それぞれの場合の得点の期待値 $\mathrm{E}Y(1)$, $\mathrm{E}Y(0)$ の差をとればよいのではないか. これが**平均処置効果**である[*17].

$$ATE := \mathrm{E}Y(1) - \mathrm{E}Y(0) \tag{2.1}$$

[*15] 計量経済学の様々な分析方法を開拓したことで, 彼は 2000 年にノーベル経済学賞を受賞している. Imbens が彼を挙げたのは, ここで取り上げた経済学者たちが使った分析手法を一般化された理論として整備したからであろう.

[*16] 2015 年にノーベル経済学賞を受賞した人物である.

[*17] また別の定義として, **処置群の平均処置効果** (*ATT; average treatmen effect on treated*) という概念が登場する. こちらも頻繁に使われるが, ここでは因果推論の最も簡単な形式だけを説明する.

しかし, データをいくら多く集めても, 実際には同一の生徒に対して補習した場合の結果 $Y(1)$ と, 補習しない場合の $Y(0)$ の両方を観察することは絶対にできない. よって, 実際に計算できるのは以下の τ である.

$$\tau := \mathrm{E}\left[Y(1) \mid D = 1\right] - \mathrm{E}\left[Y(0) \mid D = 0\right] \tag{2.2}$$

ATE をこれらの条件期待値で求めるには, **結果変数 D と割り当て変数 Y が独立であること**が必要である. これならば, 条件期待値は周辺化でき, $\mathrm{E}Y(D) = \mathrm{E}\left[Y(D) \mid D\right]$ となるので, (2.1)と(2.1)を同一視できる.

これが成り立つ状況として, **ランダム化比較試験** (*RCT*), つまり全ての生徒に対してランダムで補習を受けさせるかどうかを決められる場合である. これはネイマンやフィッシャーが考案した**実験計画法**を起源とする, 素朴なアイディアである. データの計測には必ず何かしらのノイズが混じるが, 研究室の中で, 実験動物や培養菌, あるいは薬品や新素材のサンプルを複数用意して, 厳重に環境を管理した状態ならばノイズを抑えられる. **研究室内での実験** (*laboratory experiments*) に対して, 生徒の補習は研究室の外で行われており, 生徒には勉強が得意だったり苦手だったりという個性があるしかし, 個性とは関係なくランダムに補習を行い結果を平均すれば, 個体差やノイズの影響を打ち消し合うことができる. つまり, RCT は研究室内での実験と同じプロトコルとみなせるという発想である.

もっとも単純な ATE の計算は 2 群の Y の標本平均の差を取ることだが, より実用的に, 他にも変数が含まれる実験データが手に入った場合を考える. 以下のような重回帰モデルを当てはめる.

$$Y = \alpha + \tau D + \beta X + \varepsilon$$

X は追加の**説明変数**で, 性別とか, どのクラスに属しているかとかの属性情報を想像してほしい. これらの属性が結果変数 Y に影響を及ぼしている可能性があるのだ. そのためこのような X は**共変量** (*covariate*) とも呼ばれる. 個体ごとに異なる情報を説明変数に追加することで, 結果変数に影響する, D 以外の要因を除去する意味がある. ここから, さらに X は**統制** (*cotrol*) **変数**という呼びかたもされる. ε は期待値がゼロの誤差項である. ここで, $D = 1$ の場合と $D = 0$ の場合で両辺の差を取ると以下のようになる.

$$Y(1) - Y(0) = (\alpha_1 - \alpha_0) + \tau(1 - 0) + \beta(X_1 - X_0) + \varepsilon_1 - \varepsilon_0 \tag{2.3}$$

一方で, それぞれの条件期待値を取れば,

$$\begin{aligned}
\mathrm{E}\left[Y(1) \mid D = 1\right] &= \alpha_1 + \tau\mathrm{E}[1] + \beta\mathrm{E}X_1 + \mathrm{E}\varepsilon_1, \\
\mathrm{E}\left[Y(0) \mid D = 0\right] &= \alpha_0 + \tau\mathrm{E}[0] + \beta\mathrm{E}X_0 + \mathrm{E}\varepsilon_0
\end{aligned}$$

となる. RCT のもとでは $\mathrm{E}\varepsilon_1 = \mathrm{E}\varepsilon_0$, $\alpha_1 = \alpha_0$, $\mathrm{E}X_1 = \mathrm{E}X_0$ と考えられているので, 差を引くと τ が残り, (**??**) が導かれる. よって, D の回帰係数 τ が ATE に等しい. つまり, 簡単な**重回帰モデルでも因果推論ができてしまう**のである. このように, 現実には観測できず, 存在もしない変数をあたかも存在するかのように扱う, もっとくだけた言い方をするなら**処置群と対象群があたかも同一個体が互いに交わらない平行世界でそれぞれ体験した出来事であるかのように扱う**ことから[*18], Rubin の因果推論は**反事実的** (*counterfactual*) とも呼ばれる[*19].

[*18] 発表スライドで言及されている『STEINS; GATE』は, タイムマシンにより平行世界を移動することで問題を解決するという SF アドヴェンチャーゲームである. 私が因果推論に出会ったのは学部生の頃で, その直後くらいにこの『STEINS; GATE』のアニメが放映され, 印象に残っていたからスライドで言及した.

[*19] 実際に起こっていないことをあたかも起こっているかのように仮定するというのは形而上学的であるという批判がDawid (2000) によってなされているが, Deaton and Cartwright (2018) はライヒェンバッハを引用して反論しているが, 私は確認していない. Pearl (2009) は自身の方法論は定式化されたモデルに基づくものであり検証可能であるから科学的であり批判の対象ではないという, ポパー的な反論を述べている.

しかし必ずしも RCT を実施できない場合もある. この例では, 補習の有無で, 生徒間で不公平が発生するからである[*20]. 人間を対象とするため, たいていの場合で費用や倫理的な問題が発生する. そこで補習への参加を任意にすると, RCT にはならない. なぜなら,

- 元から成績の良い人間は, 勉強熱心なのでそうでない人より補習を受けたがる.

あるいは

- 元から成績の悪い人間は, 成績をあげようと, そうでない人より補習を受けたがる.

という 2 つのうちどちらかが起こるかもしれないからだ. これは Athey (2017) が挙げた例とそっくりではないか. そうなると ATE の結果にも偏りが生じる. これは**選択バイアス**の問題と呼ばれる. このように, 経済学をはじめ社会科学では, 観測者の介入操作に実験対象が反応するという問題が常につきまとう. 特定の分野や問題によっては, **プラセボ効果**, **ピグマリオン効果**などと呼ばれる現象である. 知的生物が対象ではない物理学や化学の実験ではこのようなことはほとんどないと思われる[*21]. 経済学ではより広く, 現象全般を指す概念として, **内生性** (*endogeneity*) という用語が使われており, 選択バイアスも内生性の問題の 1 つである.

内生性の問題は既に数理モデルで定式化されており, 既に挙げた経済学者たちによって, RCT を使わずともデータの中から**操作変数** (*IV; instrumental variables*) を見つけ出したり, **傾向スコア** (*propensity score*) というものを計算してバイアスを補正する方法が提案されている. つまり, この文脈で発展してきた Rubin 流因果推論とは, 条件がコントロールされていないデータ (**観察データ**; *observational data* という) から, RCT によって得たデータ (**実験データ**; *experimental data* という) に近い品質の分析結果を引き出そうという取り組みである.

実験をせずに実験に近いことをしようということで, これらのフレームワークは本来の意味での実験と対比的に自然実験または**準実験** (*quasi —*) と呼ばれる.

適切な状況で RCT または自然実験のフレームワークを使えば, **重回帰モデルなどのごく簡単なものでも, ATE をかなりうまく推定できる**.

2.3 予測と因果推論の形式的な特徴づけ

より形式的に言い換えると, 次のようになる. Athey and Imbens (2019) は経済学者向けに機械学習の役割を解説する際に, これまでの計量経済学のアプローチを「**同時分布を推定する**」ものと特徴づけている. これに関して, 杉山 (2013) ではさらに細かく定式化している[*22]. 機械学習モデルは**判別** (*discriminative*) モデルであり, 計量経済学モデルは**生成** (*generative*) モデルであると言われる. 判別モデルは $p(Y \mid X)$ で, つまり目的変数の**条件確率**を表す. 条件確率がわかれば, **条件期待値** $\mathrm{E}[Y \mid X]$ を計算することができる. **機械学習の予測値は, 条件期待値に説明変数 X を入力して計算する**ものだとみなせる (ただし決定木のように条件期待値を厳密に定義できないモデルもある). 一方で, ATE の説明から分かるように, 因果推論ではこの説明変数 X に含まれる, 2 通りの値をとる D に注目している. 因果推論では無条件で $D = 0$, $D = 1$ を入力したときの予測値の差を取っているのではなく, RCT やそれ以外の手法によって条件 X をそろえている. 両者の違いは確率分布 $P(X)$ を考慮するかどうかであるとわかる. 条件確率 $P(Y \mid X)$ に加えて $P(X)$ も特定すれば, 同時分布 $p(X, Y)$ もわ

[*20] 臨床研究でも倫理的な問題がある. 例えば『「操作変数法」の報告事例』では喫煙の健康リスクを評価するために, 喫煙を割り当てる RCT をすることは倫理的に却下されると指摘している.

[*21] と言っても私も詳しくないので, 物理学や化学でもしこういうケースがあるというならば教えてください.

[*22] この定式化はいくつかの教科書で見られるのだが, 最初にこの主張がなされたのがどこなのか私は知らない. 誰か知っていたら教えてください.

かる. ここから, 計量経済学は同時分布を推定することだと特徴づけられる.

例えば「傾向スコア法」は, $D=1$, $D=0$ の 2 グループのそれぞれの**説明変数の分布を調整して, 偏ったデータを較正**するというコンセプトであるから $P(X)$ を調整する方法, と捉えることができる.

また, 適切な RCT では重回帰モデルだけで ATE を推定できると書いたが, 逆に言えば RCT ではなく観察データで重回帰分析をした結果は因果関係を表すとは限らない. つまり, **因果推論はモデルの技術的高度さや自由度の高さ** (どんなデータに対しても当てはまりをよくできるか) **よりもデータの取り方がより重要である**ということで, **説明変数の分布を調整する技術**だとみなせる.

2.4 さまざまな因果推論フレームワーク

既に名前を挙げた「操作変数」「傾向スコア」「RDD」「DID」といった方法は Rubin の因果モデルフレームワークに基づいたものである. これらもわかりやすく紹介したかったのだが, 1 から書きおろそうとすればこの記事のボリュームが際限なく増大するし時間もなくなるので過去の参考文献に投げる.

RDD については, 過去に自分で解説を書いたことがある.

- 非連続回帰デザイン (Regression-Discontinuity Design)
- 非連続回帰デザイン (RDD) 理論編
- 非連続回帰デザイン (RDD) 実践編

操作変数に関する話も書いたが, 因果推論の視点からの説明ではないのでどちらかというと記事内で紹介されている教科書を読んだほうが確実である.

- 非線形モデルでの一般化モーメント法と操作変数
- 非線形モデルと操作変数の応用例

ここではかなり単純化した例で因果効果の推定方法を説明したが, 現実のデータにあてはめる際にはここで触れていない多くの前提条件を確認しなければならず, 名前だけ紹介した操作変数法, 傾向スコア法, RDD, DID などは**それぞれできることに限界がある**. それを無視して濫用してはまったく意味をなさない分析しか生まれない. もしここを見て活用したくなったら, 以降に挙げる教科書を参考に理解を深めてほしい.

Rubin 流の因果推論の, 特に社会科学での応用寄りでより深く知りたい人のためにいくつか教科書・資料を挙げておく. DID 法は単純なのですでにいろいろな教科書で教えられている. 私の記憶の範囲では, 最初に日本語の教科書で言及したものは大森 (2008) の労働経済学の教科書であったと思う. インターネット上で一般公開されている資料としては, 津川 (2015) が分かりやすいと思う. 傾向スコア方は, 専門的だが中村 (2019) も因果推論の暗黙の前提に対する示唆を与えてくれる良い資料である.

RDD や操作変数法など計量経済学の分野で開拓された因果推論を扱った教科書としてはAngrist and Pischke (2009), Angrist and Pischke (2015) が書いたものが有名である[*23]. Varian (2016) のレビュー論文は英語だが[*24], 一般公開されている上に DID や操作変数法, RDD と言った計量経済学で使われる因果推論が全て簡潔に解説されている. 黒澤 (2005) のレビューも一般公開されており, これらの手法について簡潔にまとめている. さらに日本語でかつ無料公開されている田中 (2019) の解説では, ここで触れなかった因果推論の前提条件をいくつも紹介している. 戒能 (2017) も非常に長いが様々なことに言及している. 星野 (2009), 高井他

[*23] なおImbens and Rubin (2015) の教科書もあるが, 私はまだ読んでいない.

[*24] 英語がめんどくさくて仕方がないという人は http://kamonohashiperry.com/archives/2486 を読めばよいだろう.

(2016) は, データの欠測という問題により注目した因果推論の教科書である. 経済学でいう内生性の観点からの説明が知りたければ, Wooldridge (2018), Cameron and Trivedi (2005), Wooldridge (2010) など標準的な計量経済学の教科書でも良い[*25].

数式がどうしても苦手という人にとっては, 理論的に厳密な議論を避け, 簡単な導入にとどめた森田 (2014) もあるし, 岩波データサイエンス刊行委員会 (2016) は理論の説明よりも事例紹介としての側面が強いが, DDD (double difference in differences) 法や Local ATE (LATE) といったより現実的な概念についても言及がある. 実際の分析事例を紹介した, 伊藤 (2017) の新書も良いだろう. 伊藤 (2018) の『データ分析を経営や政策に生かすには?「因果分析」と「予測」の適切な使い分け』はインターネット上で一般公開されており, なおかつ非専門家にもわかりやすい簡易な解説になっている.

2.5 第 2 節のまとめ

- 計量経済学の因果推論の側面を強調すると, パラメータを適切に推定することを重視している.
- 因果推論は必ずしもモデルのデータに対する当てはまりの良さを追求するわけではない.
- 一方で典型的な機械学習は, モデルにデータにどれだけ当てはまるかを追求する
- 最も単純な因果推論フレームワークは, 重回帰でもできるランダム化比較実験 (RCT) である.
- RCT ができない場合の因果推論フレームワークは, 自然実験あるいは準実験と呼ばれ, 既にさまざまなフレームワークが研究されている.

脱線した話題 I: 因果推論は経済学者だけのものだったのか?

DID は現在でこそ経済学の実証研究で頻繁に使われているが, このような実験デザインの起源を遡ると, 麻酔技術を発展させた外科医であり, 疫学・公衆衛生学の開拓者でもあった John Snow によって 1850 年ごろに考案された "controlled before-and-after study" に遡るといわれる[*26]. RDD の基本的なフレームワークは教育心理学の論文誌に掲載された Thistlethwaite and Campbell (1960) による研究が初出とされている (Angrist and Pischke, 2009). そもそも, Rubin 本人も心理学者であり, 彼の理論は 18 世紀の哲学者ヒューム (David Hume) の思想に着想を得ている. 経済学以外の分野でも, かなり以前から因果効果を測定するにはどうすればいいかという問題意識があったことは明らかである.

[*25] 2015-2016 年ごろに日本語の計量経済学の教科書が相次いで出たのだが, どれも手元にないのでオススメを挙げられない. ~~あの頃は貧しかった.~~

[*26] "Difference-in-Difference Estimation | Columbia University Mailman School of Public Health," Angrist and Pischke (2009) にも似たような話が紹介されていた気がするが, 本が手元にないので確認できない.

3　計量経済学と機械学習の限界

3.1　Deaton と Cartwright の批判

RCT や自然実験フレームワークを使った経済学の研究は, 最近 20 年で増加している. では, RCT さえできれば万事問題ないのだろうか. ここで, Deaton and Cartwright (2018)[*27][*28] の主張に目を向けてみる. この論文はかなり大雑把に要約すれば, 因果推論と称して何も考えずに ATE つまり平均の差分を計算するだけで結果が出たとする分析に問題があることを主張しており, 頻出する皮肉めいた言い回しも含めてかなり面白い. タイトルからして RCT だけに対する批判のように見えるが, 実際には RCT よりも広く因果推論を求める分析アプローチ全般に対して刺さる批判になりうる指摘も多い. 今回の話に合うように, 元論文の主張の力点とは少し違うところを強調しているが, 私なりに主張を以下のようにまとめる.

1. RCT の理論フレームワークでは, ほとんど仮定を要求せずに因果効果を知ることができるが, RCT を現実の問題に応用するにあたって, 現実のデータの特徴と理論上の仮定の違いを考慮するとできることとできないことがあるとわかる. これは, 式(2.2)から ATE の式(2.1)へと変形する際に必要な仮定が, 果たして適切なのかという問題である.
2. RCT を否定しているわけではない. **RCT を魔法かなにかのように考えることを否定している**のだ. さらに, RCT や他の手法が, それ以外の手法よりも格上だということもない. RCT や, 経済モデルや操作変数や**ベイジアン因果ネットワーク**や他の因果推論の方法のいずれにもそれぞれの強みと弱みがある.
3. 先験的な知識と, 新しい仮定とデータから得られる新しい含意を統合して新たな知識を産むのが科学的な方法だ. しかし単に毎回 RCT で得られる異なるデータについてその結果を報告するのみではこのサイクルが成り立たない. 「『どうなるか』ではなく『なぜそうなるか』を見つけること[*29]」が重要である.
4. RCT では**外的妥当性** (*external validity*) の担保に向いていいない

彼らは RCT の具体的な技術的限界の問題についてもいくつも挙げており, 実用的な知識が詰まっている. しかしこの論文の主張を全て載せていてはこの記事の本題が不明瞭になるため, 今は私の話に関係のある, 比較的に抽象的な箇所だけを取り上げる. 特に強調したいのは, (3, 4) の 2 つでる.

(3) 科学的な方法とは, それがなぜ起きたのかという問いを繰り返し, 新たな情報に基づいて答えを更新し続けることである. 何も考えず, 先行研究の知見も尊重せずにただ平均の差を求めただけでは何が原因で差が生じたのかもわからずに, 差の大きさを報告することになる. それは実際に何かの政策の成果かもしれないし, そもそも RCT の設計に欠陥があることを見落としていたからかもしれない. RCT の欠陥を見つけることができたら, また別の推定方法に修正したり, サンプルのとり方を変えるなど修正することもできる. しかし RCT の結果を計算して報告するだけでは, なにも進歩がない.

(4) また, 彼らは外的妥当性を「サンプルデータ以外のどこでも同じ結果が成り立つこと」と定義しており, RCT ではしばしば外的妥当性の担保は限定されていると主張する. しかしだからといって RCT を否定することにも批判的である. むしろ RCT 以外のフレームワーク, たとえば経済モデルに基づく分析から得られる先験

[*27] これは 2018 年の論文だが, 議論の大枠は 2016 年のワーキングペーパーの時点であらかた完成している. 機械学習の分野に比べ, 経済学は研究が査読論文として公開されるまでのスパンがかなり長い傾向にある.

[*28] この論文は環境リスク研究者である林 (2019) もブログで取り上げている. 主題は私のこの記事とは違うが, RCT の数値だけをありがたがる風潮に対して批判的なことは一致しているようだ. 実際の口頭発表でどのように言及したのか気になる.

[*29] to discover not "what works," but "why things work"

的な知識と RCT の組み合わせによって新たな知識が生まれる, あるいは既存の知識がより信頼できるものになるとしている. ただしImbens (2018a) はこれに対し内的・外的妥当性の定義が広く普及したものでないためおかしいと指摘している. 彼のいう広く普及した定義は内的・外的妥当性の定義は「観察された共変関係が因果関係を反映しているという推測の妥当性」「原因と結果の関係が, 個人や設定や処置か非処置かの違いに対しても同一であるという推測の妥当性」であり, これらは RCT あるいはその他の実験的方法でも十分に検証できると主張している[*30].

3.2 機械学習では答えられない 2 つの問い

しかし残念ながら, この問題に関しては計量経済学も機械学習も解決方法を提案できていない.

Imbens の挙げた通常の定義であれば, 内的・外的妥当性の問題に対処するべく, いろいろな研究がなされている. 一方で, Deaton and Cartwrightによる定義はむしろ, 機械学習でいうところの**汎化** (*generalization*) の概念に近い[*31]と私は考える. Mullainathan and Spiess (2017), Athey (2018) で指摘されるように, これまで経済学者は, データ全体に対してモデルを当てはめた結果を見ることが多かった. これは予測ではなく, 過去の事例の事後検証を目的にするだけならばさほど問題でないように思えるかもしれないが, それは誤りである. データのノイズに対してモデルが**過剰適合** (*overfitting*, **過学習** *overlearning* と同じ) の問題からは抜け出せない (その理由の説明は『統計的学習の基礎』Hastie et al., 2009, 7 章の説明が分かりやすいと思う). 機械学習の過剰適合対策としては, **ホールドアウト検証**や**クロスバリデーション** (**交差検証**) といった, データを分割して推定用と検証用に分けて使うという方法が典型的である. Athey and Imbens (2019) はこのような機械学習のありかたを, 「(データ外の) **新しい個体の予測値を求めるのが目的である**[*32]」と表現している.

しかし, 経済学においては 1 つ重大な問題が存在する. 機械学習でいう「汎化」とは, 取得したデータと, 取得していないデータ (たいていは未来に発生するデータ) が同じ確率分布から生成されているという, **斉一性の原理** (*principle of the uniformity of nature*) が正しいと暗に仮定している. そしてこれは形而上学的あるいは哲学的な命題のため, 科学的 (実証的) な検証が困難な仮定である. クロスバリデーションをすれば汎化性能を向上できるという説明をときどき見かけるが, これも確率分布が独立かつ同一であるか, それに近い留保条件が必要である[*33].

学習 or 推定という操作が, 取得したデータに対して損失 (尤度) 関数を最小 (最大) 化することであるという基本が変わらない以上, この問題は従来の機械学習であっても, 最近流行しているディープラーニングであって

[*30] 津田もまた自身のブログ記事『妥当性（Validity）と信頼性（Reliability）』で内的・外的妥当性の定義を紹介している. いずれも Shadish et al. (2001) に基づく. やや文言が異なるが, 趣旨は同じだと考える. Imbens はこれ以外にも二人の主張について反論を展開しているので, Deaton らの論文を読む際はこちらも確認することをおすすめする.

[*31] 経済学では**補外**または**外挿** (*extrapolation*) と呼ぶことも多いが, 補外は汎化とは指し示す範囲が異なる. 後述するように反事実的なシミュレーションや, 外的妥当性のさらに細かい特定のケースを指す場合もある.

[*32] "the goal may be to make a prediction for the outcome for new units on the basis of their regressor values."

[*33] この記述もあまり正確ではない. さらに正確に言えば, 交差検証によって改善するのは汎化誤差ではなく汎化損失であり, 確実にできるのは in-sample データへの過剰適合の回避までである. この違いはHastie et al. (2009, 7 章) および渡辺 (2012) あたりを読めばわかるだろうか? あるいはさらに場合によっては「できる」という言葉の定義にまで踏み込む必要があるかもしれない. 手を動かすのが好きな人であれば自分でやってみて動けば「できる」と表現するだろうが, ここでいう「できる」は個別の事例に対して結果的にうまくいったかだけではなく, なぜうまくいったかの理由付けまでできるのか, 可能性があるだけなのか, と様々な意味を含む.

も発生する[*34][*35]. 経済学においては特に, 分析対象である人間の行動のパターンが変わりうる可能性が指摘されている. これは斉一性の原理が成り立たないということである. これが私の以前書いた『ディープラーニング vs ディープパラメータ』で主張したかったことの 1 つである. ただし, この斉一性の原理を持ち出す仮説は私が独自に思い立ったものであり, 経済学の論文でこのような形而上学的な議論は (たぶん) 見られない. 一方で『信頼性の高い推論』(Harman and Kulkarni, 2007, 8 章) によれば, このアイディアはネルソン・グッドマンの「**グルーのパラドクス**」とそこから得られる含意に近いようだ. 合わせて彼らは, 統計的学習理論がこのパラドクスに正面から向き合おうとはしていないことを指摘している[*36].

しかし根本的にパラドクスを解決できていないという点は経済学も同じである. 経済学ではこの分布の変化を, **ディープ・パラメータ** (aka **構造パラメータ**) で記述できるという前提でモデルを仮定して検証している. これが**フォワードルッキング**なモデルと呼ばれるものである. これだけではパラドクスを解決できないが, 経済学が当面の課題としているのはいわゆる「ルーカス批判」(Lucas, 1976) の解決である. この問題に関する一般向け解説は小林によるコラム『「期待」どこまで解明？』が分かりやすいと思う. 小林の表現を借りる.

「ルーカス批判が示す期待の自己言及性とは、『経済システムの中にいる国民と政府が、経済システムの法則（期待）を知っており、それに基づいて行動すると、結果的に法則が変わり得る』という事実である。経済法則は、図 4のような期待形成のループによって生成される。このためマクロ経済政策は効果を失う、とルーカスは述べた。このループは永久に続くので、一般的には経済法則はいつまでたっても定まらない。」

このような経済現象特有の構造がなくならない限り, 機械学習的な汎化性能の評価方法をそのまま経済モデルの評価方法とするのは難しいと私は考える.

そしてこれはよく見られる誤認あるいは過度の単純化だが, **経済学が過剰適合に注意を払っていなかったと言うのは誤解**である. 経済学者も過剰適合という概念を以前から認識していた. たとえば**ブートストラップ法**の一種である**ジャックナイフ推定量**[*37]は *Leave-one-out cross validation* (*LOOCV*) とアイディアを共有している. そしてモデルの**ロバストネスチェック**または**感度分析** (sensitivity analysis), あるいはAthey and Imbens (2017) が "supplementary analyses" と呼ぶプロトコル, つまり自分の仮説に基づくモデルが本当にデータと整合するか (そして手持ちのデータ外でも通用するか) の検証項目は多岐にわたるため, かなり注意を払ってきた[*38]. このプロトコルは多種多様だが, 一般化していえば, 関数の形状を一部変えても同じことが言えるかとか,

[*34] いわゆる「普遍性定理」または「万能近似定理」(Cybenko, 1989) で示されるように, ディープニューラルネットワーク (DNN) は高い表現能力を持ち, 複雑な構造も近似できる可能性のあるモデルなのでしばしば誤解されるが, 複雑な構造を近似できることは, 限られたデータからの学習だけでその近似が達成できることを保証しない (そもそもそこを簡単にできてしまったらここまで研究が盛んになる必要すらなく, すぐにでもシンギュラリティの時代とやらに突入していただろう). この誤解に対する厳密な説明は難しいが, 例えば田中他 (2019, 5 章) は中心極限定理と, **ヴァプニーク＝チェルヴォネンキス** (VC; Vapnik-Chervonenkis) 次元で定義される汎化性能の関係についてのスケッチを与えている. 彼らは触れていないが, 逆に言えば中心極限定理が前提としている確率分布の斉一性 (正確には独立性と同一性) が重要であることが分かる.

[*35] 斉一性の原理が正しいとは限らない, とは絶対になるとは保証できないということだが, 同時に絶対にならないとも限らないということで, 現実の現象は完全に無秩序なわけではなく, 将来においてもあまり変化しない安定した確率分布で表現できることも多いのが事実である. より正確に言えば, 「普段扱っている有限のデータだけで学習した予測モデルに他のデータを補外してもある程度予測があたることが多い」とからそのように推理できると言うべきだが.

[*36] この記事は万人に向けたものなので予防線を貼っておこう. これはあくまで, 機械学習が対象としている範囲を指摘しただけであり, 研究者が怠慢であるなどと難癖をつけたいわけではない. 機械学習に限らず, この問題を解決するのはとても難しい.

[*37] ジャックナイフ推定量はCameron and Trivedi (2005) で言及がある他, 操作変数など計量経済学特有のモデルでも扱う研究がいくつもある. たとえば Angrist et al. (1995), Hahn et al. (2001).

[*38] ロバストネス (頑健性, robustness) という語の指すものはかなり曖昧なように思える. 機械学習周辺の文脈では, 単にバリアンスが小さい, もしくはデータに外れ値が混入していても学習結果が汚染されにくい, という程度の意味で使っていることが多いように見えるが, 経済学の実証研究者たちは, モデルの仮定の一部が崩れた場合にも結果が大きく狂ったりしないことを指すことが多い (外れ値に対してロバスト, という言い回しももちろん使うので, 指し示す範囲がより広いと考えるべきか). 前者は仮説を所与とした上での議論であることが多く, 後者は仮説やその背後にある大前提を根本的に疑っていて, さまざまな項目が検証対象となるので, 誰

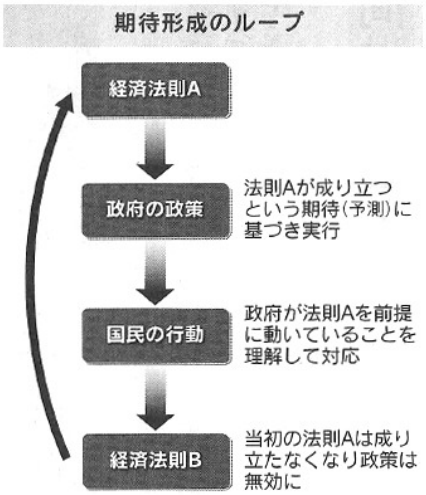

図4　小林による図解

異なる母集団を想定して再試験しても結果が変わらないか (つまりサンプルの分割であり, ホールドアウト検証のアイディアと同じだ), といった評価のことを指す. マクロ経済学で使われる動学的確率的一般均衡 (*DSGE*) モデルでは, 予測精度の評価として**ベクトル自己回帰** (*VAR*) モデルという構造の異なるモデルを用意して予測値と比較するという, **インパルス応答マッチング**とよばれるプロトコルが考案された. 因果推論でも, 傾向スコア法で求められる推定量 (IPW 推定量) に対して一部の係数を調整した場合の結果の変化を確認するプロトコルが提案されている (高井他, 2016, 7 章). そもそも, 経済学者が単なる観察研究から因果推論的アプローチへと興味を向けたのも, 既存の手法が過剰適合することに問題を感じていたからに他ならない[*39].

抽象的な話が続くので, ここまでの汎化に関する話を実際に例を示してみる. 一定のルールと, 小さな乱数を合成して生成したデータがあるとする. それを散布図に描くと図 5のようになった. x1, x2 は説明変数で, 点は目的変数の違いで色分けしている. 機械学習をかじったことのある人間ならば, 分類問題のタスクを連想するだろう. 計量経済学しか知らない人間ならば, 説明変数が 2 つしかない離散選択モデルを想像すればいい. つまり, ロジットモデルやプロビットモデルだ. 目的変数は「商品を買うか買わないか」とか「大学院進学するかしないか」のようなよくある選択問題を想像すればよい. 説明変数が 2 つしかないため, モデルのあてはめ結果を平面的な図で表現できる. 例として, 機械学習の典型的なモデルであるランダムフォレストと, 計量経済学でおなじみのロジットモデル (ロジスティック回帰) をそれぞれ当てはめて, モデルの出力する予測値ごとに下地を色分けした**決定領域**を重ねてみる (図 6). 以前私が『誤った図解から学ぶロジスティック回帰の性質』で書いたように, ロジットモデルは直線的な分類しかできない一方で, ランダムフォレストは複雑な境界線を描くことができるため, うまく 2 つのラベルを分類できている. では, このモデルは補外にも使えるのだろうか. ここで推定したランダムフォレストを使って, 同じ法則で生成した新しいデータの予測をしてみる. その結果が図 7である. これまであなたが見ていたのは右上の部分だけだ. 新たに追加したデータの部分を見ると, 全体として渦を描くような分布であることが分かる. そして実際のデータと予測がまったく一致していないことが分か

がロバストネスと言う言葉を使うかで, 想定している問題が全く異なってくるため注意が必要である. 仮説をメタ的に検証するという点で, 機械学習の汎化性能は経済学ではロバストネスに対応するのかもしれない.

[*39] この研究アプローチの変化は既に述べた 5 人の経済学者らが立役者となっていて, この転換は『信頼性革命』と呼ばれている. このような流れは, 會田『「予想よりも早かった」ノーベル経済学賞』でも触れられている.

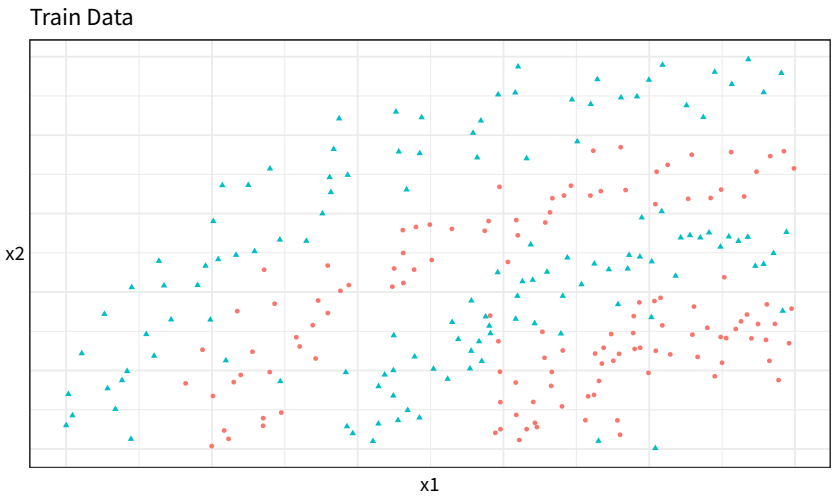

図5　訓練データ

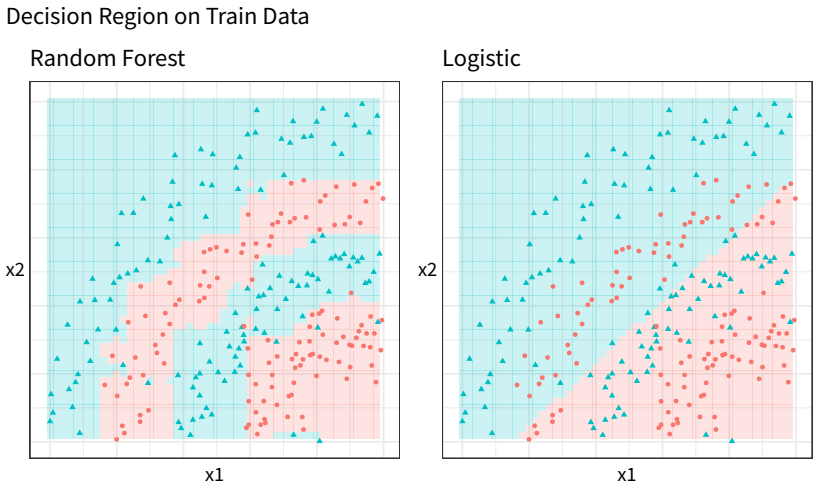

図6　ランダムフォレストとロジットモデルで当てはめ

った. ロジットモデルの結果は言うまでもないだろう.

あなたはこれを**ばかげた引っ掛け問題だ**と思うに違いない. 私もそう思う. わかりやすくするためかなり極端な例を出した. しかし, 現実のデータを分析の対象とするにあたって, これと同型の現象は全くありえないとは言えない (具体的な話は6.2節で少しだけ掘り下げる). 機械学習で予測する際に機械学習では自由度の高いモデルがいろいろと考案されているが, 目前の, 有限のデータへの当てはまりだけでは究極的には何もできない. 単にモデルを当てはめて終わるのではなく, 人間がデータの外にある構造を洞察することが必要になる. これが経済学でいうロバストネスチェックに対応する[*40].

[*40] ところで, 私がかつて『計量経済学と機械学習の違い』で挙げた例は, 多項ロジットモデル (ソフトマックス回帰) と順序ロジットモデルの当てはまりを比較しても後者のほうが優れているため, 機械学習であっても構造を特定するのが重要であると説いた. しかし, 現実の問題ではそのデータの分布の構造に対してどのモデルが一番近いのかは予め知ることができない. 過去の記事のような, 機械学習でよくやっているように単に与えられたデータに対する当てはまりの良さを見るだけならば, 自由度のあるモデルのほう

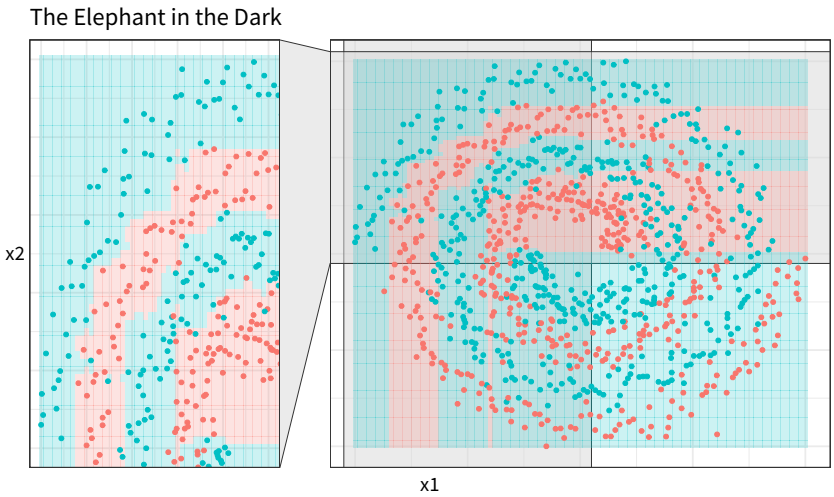

図7　不都合な現実

すると, 次のような疑問が発生する. グルーのパラドクスに答えられないのは機械学習も計量経済学も同じなのだから, 当てはまりの良いモデルを作れる機械学習のほうが優れているのではないか, ということになる. 言い換えるなら, データの外の構造がどうなっているのかという洞察はどちらの方法論でも自動的にやってくれるわけではないのだから, 仮説に基づいて**機械学習の方法で当てはまりの良いモデルを作成すれば良い**ということにならないだろうか?

そこで, **第 2 の問題点**について触れる. 仮説の検証だけでなく反事実的なシミュレーションも経済学者の興味の対象として重要である. 『ディープラーニング vs ディープパラメータ』で書いたように, 経済学の研究ではある説明変数の変化に対して結果変数がどれだけ変化するかという因果関係を求めることが目的意識になっていることが多い. これは**まだ実施されていない政策の結果を予想する**ためであり, 前セクションで言及した RCT や自然実験の手法はこの目的でも活用されている. しかし**機械学習はこのような用途で使えないため, 経済学者はニューラルネットや XGBoost ではなく, それができる重回帰モデルの係数を重視しているのだ**. ここではモデル全体の当てはまり以前に, 係数をバイアスなく推定できていることが前提になる. 機械学習では**バイアス-バリアンス分解定理**から誤差を小さくするモデルを設計するにはバイアスとバリアンスそれぞれに由来する誤差のバランスを取るのが良いとされる (Hastie et al., 2009, 7 章) が, 経済学の研究ではまず**バイアスなく推定するか少なくとも漸近的にバイアスがなくなるように推定し, その上で信頼区間を計算できるようにする**ことが重視される. 逆に機械学習ではこういった漸近的性質について, これまであまり研究がなされていなかった (Athey and Imbens, 2019). ランダムフォレストなど決定木を用いたモデルは, 木構造の条件分岐で記述されるためそもそも**係数が存在しない**ため, ATE を推定しようとして重回帰モデルを決定木やランダムフォレストに置き換えても, 効果の大きさを定量的に提示することが難しい. 同様に, Ribeiro et al. (2016) がニューラルネットを局所的に線形近似する LIME という手法を考案したのが以前話題になったが, 単に係数を提示するだけでは不十分で, 2節冒頭で挙げたような統計的性質をどう担保できるかという知見も必要になる. ここでも, 目的設定の違いから方法論の食い違いが発生している[*41].

が当てはまりやすいという単純な話になってしまい, 例としては不適切だったと反省する.

[*41] 機械学習と計量経済学の差異の一面だけを切り取って端的に表現するとしたら, **赤池の情報量規準** (*AIC*, Akaike, 1974) と**ベイズ情報量規準** (BIC, Schwarz, 1978) に仮託できるかもしれない. 経験的には両者は似たような結果になることが多いが, 理論上は両者

3.3 既製品の機械学習

Athey (2017) は機械学習をそのまま観察データに当てはめることを**既製品の機械学習** (off-the-shell machine learning) と呼び, 有効となる場面もあるものの, **因果推論の問いに答えるには機械学習モデルを使うかどうかだけでなくどうデータを取るか, つまり RCT さもなくば自然実験をどう設計するかも重要である**, としている.

因果関係ではなく, より単純に予測だけが求められるケースでは機械学習をそのまま取り入れることができる. Mullainathan and Spiess (2017) も, 自然言語処理, 画像認識といった技術は従来経済学があまり扱ってこなかったデータから情報を取り出せるとして研究例を紹介している. 国内でも五島他 (2019) が金融分野において時系列予測に機械学習を取り入れている例もあるが, ここでは因果推論に関わる話題のみを話していく.

では, 既製品の機械学習が有効となる場面は何か.

経済学者は仮説を手に入るデータだけでなく補外においても検証することに関心があるので, 手に入るデータの範囲で過剰適合が起こっているかどうかは問題として小さいものだった. とはいえ, できるのにやらない理由はないし, これまでの過剰適合への対処方法は具体的にどの程度効果があるのかといったこともあまり考えられていなかったように私は思う. 例えば**決定係数** R^2 (*coefficient of determination*) に対して, **自由度修正済み決定係数** (*adjusted* R^2) の違いを, 「後者のほうが説明変数の増加に対してペナルティがかかる」以上に正確で詳しい性質を説明できる人間がどれだけいるだろうか[*42]?

Atheyは機械学習で行われるのと同じように, ***LASSO*** (*least absolute shrinkage and selection operator*, Tibshirani, 1996) のような**罰則付き回帰**や, クロスバリデーションといった**システマティックな方法で過剰適合を回避すべき**であることを主張している. Mullainathan and Spiess (2017) も, 検証したいモデルとは別に予測がよくできている機械学習モデルを作成し, 両者の予測誤差を比較することを提案している. さらに, 彼らの引用するAbadie and Kasy (2017) では「実証研究家向けのガイドライン」として通常の最小二乗法とLASSO と**リッジ回帰**の理論的特性を示した上で, 群が複数あることで ATE に異質性が発生するケース, 統制変数や固定効果モデルによって説明変数が多くなるケースなど, いくつかの典型的な経済学の実証問題で想定されるような乱数データに適用して, シミュレーション結果を比較している.

さらにAthey and Imbens (2019) はより具体的に個別の機械学習の手法を特徴づけ, 経済学の実証研究の特定の場面で利用できるヒントを述べている. その指摘は多岐にわたり, 中にはここまでの指摘と重複するものもいくつかある. それ以外では例えば**ディープニューラルネットワーク** (*DNN*) で ATE を推定する方法としてFarrell et al. (2019) を紹介している.

経済学では研究ごとに仮説もモデルも違っているのが当たり前だから, 画一的に機械学習の方法を導入することは難しい. 今泉 (2019) が指摘するように, 機械学習モデルがうまくいく理由を引き出す方法はあまり充実していない. 例えばMullainathan and Spiessの提案する方法で, 仮説モデルより機械学習の方法のほうがはるかに良いスコアを示した場合, 当然ながらその原因を知りたくなるが, 既存の研究の範囲ではその方法は必ずしも充実していない. また別の例として, LASSO を使って, 例えば処置変数の係数つまり因果効果 τ がゼロに推

が一致するという保証はない. AIC はモデルの誤差を最小化するための規準であり, BIC は正しいモデルの選択確率を最大化するための規準だからである (小西・北川, 2004). ただし**最小記述長** (*MDL*) は学習理論の研究の中で生まれたものだが性質は BIC に近いので, これはあくまでたとえのようなものだ.

[*42] Ezekiel (1930) の定義での標準的な条件下での性質は例えば Cramer (1987) が言及している. さらに経済学に特有の状況を想定すれば, Ohtani and Hasegawa (1993), Ohtani and Tanizaki (2004) などの研究がある. 機械学習風に out-of-sample な値を計算する方法としては, Yin and Fan (2001), Shieh (2008) がある.

定されてしまったとして, それはなぜかという含意 (e.g., 仮説検定とどう違うのか) をどう引き出すかという問題がある. 語調としては, 従来のロバストネスチェックの方法以外にも, 理論的性質のはっきりした機械学習の方法を活用する余地が残されているという程度のものではないかと私は考えている.

最後に「既製品の機械学習」活用に関する文献で 1 つ, 技術的な観点から気になった箇所がある. Mullainathan and Spiess (2017) もAthey and Imbens (2019) も LASSO の特徴に触れているが, 一方でスパース推定の**オラクル性** (*oracle property*) を大きく取り上げていない. 川野他 (2018) によればオラクル性とは, 次のような性質である. 線形回帰モデルの係数 (β_0, β_1) について, β_0 の真の値はゼロ, β_1 の真の値は非ゼロであるとすると, オラクル性を持つ推定量 $(\hat{\beta}_0, \hat{\beta}_1)$ とは

1. サンプルサイズが大きくなるにつれ, $\hat{\beta}_0 = \mathbf{0}$ となる確率が 1 に収束する
2. $\hat{\beta}_1$ は漸近正規性を持つ

という 2 つの性質を持つ. つまりオラクル性が保証されれば (1) 意味のない係数が選ばれず, 意味のある係数だけが選ばれるようになる (ただし偽陰性はありえる) こと, (2) 標準誤差を計算したり信頼区間を求めたりできることがわかる. これは推定量の不偏性や信頼区間を重視する計量経済学の視点から大いに強調すべき性質だと私は考えるが, Tibshirani (1996) の提案する**オリジナルの LASSO はオラクル性を持たない** (Zou and Hastie, 2005). オラクル性を得るには, LASSO とは異なる正則化項, 例えば ***SACD*** (*Smoothly Clipped Absolute Deviation*, Fan and Li, 2001, 日本語訳はまだないので適当に書いた) 正則化項, **MC+** 正則化項 (Zhang, 2010) や**適応的 LASSO** (*Adaptive LASSO*, Zhao and Yu, 2006) 正則化項といったものが提案されているが, Mullainathan and Spiessの書き方だとオリジナルの LASSO でもオラクル性を持つかのように誤解されるような気がするし, Athey and Imbensはそもそもほとんど触れていない.

3.4 第 3 節のまとめ

- (教師あり) 機械学習はデータに対するモデルの当てはめに使われてきた. 形式的に言うならば, 条件分布に対する近似の精度をデータで評価することである. データの分布が大きく変わることは通常, 考慮しない.
- 計量経済学は経済学の仮説の検証に使われてきた. 形式的に言うならば, データにない範囲でもどの程度モデルがあてはまるかを検証する (ロバストネスチェック) ことも考慮されていた.
- 機械学習は目的がシンプルなぶん評価方法も定型化されているが, 経済学の仮説の検証方法はあまり定型化されておらず, 一面だけで評価することが難しい.
- 機械学習の方法を計量経済学にそのまま一般化して持ち込むことは難しい. これはそもそも目的が違うことによる相性の悪さが原因である. とはいえ, 一部分を切り取れば, 計量経済学を進化させる場面もある.

ここまでの範囲では私が以前書いた内容とそう差異がない. 以上が, **これから導入する本題に入る前の長いあらすじである**.

脱線した話題 II: 「解釈性」「説明」といった言葉を私がなぜ避けているのか

「計量経済学や統計学はデータを『説明』あるいは『解釈』するもので, 機械学習はデータを『予測』するものだ」という表現がよくある. 英語の文献でも機械学習の役割は prediction, 統計学の役割を reasoning,

interpreting, explaining などと表現されることが多いが, これらの語の持つ意味合いは多義的で一語で表すのは難しく, 受け取る人によって全く異なるものを連想することがある. 言葉の使い方を曖昧にしたせいで認識のずれを発生させたくないため私はこの表現を意図的に避けている.

「解釈」については, たとえばモデルの結果からどんな含意を引き出すかという問題で言えば, 計量経済学でも機械学習でも, 既存の理論研究の蓄積に基づけば可能である. 今泉 (2019) は特徴量の可視化を最もストレートな機械学習における解釈の手法として, 最近の研究例を紹介している. そして可視化による「解釈」であれば, すでに普及している方法も数多くある. 例えば線形回帰モデルであれば, モデルの残差のヒストグラムを描画してデータの分布が正規分布から程遠いから一般化線形モデル (GLM) のほうが当てはまりが良い可能性があるとか, コレログラムを計算して系列相関があるから分散が推定されるよりも過剰だといったことを[*43]推理できる. グラフを使わないものであれば, ロジスティック分類と分類木の結果を比較して, 分類木のほうがずっと当てはまりが良いから, 分類面が複雑な形状をしておりロジスティック分類に向いていない可能性が高いとか, あるいはランダムフォレストのハイパーパラメータのどれか 1 つを変更して, どういう変化が予想できるか, といった単純な含意を理論さえ知っていればいろいろと推理することができる. 「解釈」が指すのがこのような意味であれば, それは計量経済学でも機械学習でも利用する際に必ずついて回る推理であろう.

さらに今泉は, 経済学での利用を念頭に置いて, 「解釈」を「データを通じた仮説の検証」という意味に限定して議論を深めている. その上で, 仮説として与えられたモデルを検証するという経済学の方法論と, 高い精度 (当てはまり) のために柔軟に特徴量を変換することを許容する機械学習の方法論とでは相性が悪いと指摘している. これは, 経済学で仮定されるモデルが, 上記の単純な仮説検証よりも複雑な制約条件として与えられることが多く, 機械学習モデルの自由度の高さと相反するからである.

この文脈では「説明」も「解釈」も同義語になる. 仮説から構築したモデルをデータに当てはめ, 当てはまりが良いなら「仮説でデータ (観測された事実) を『説明』できる」という意味で説明という言葉が使われる. 単に当てはまりの良さという意味で「説明力」というジャーゴンがたまに使われるのは, このあたりに由来していると思う.

Ribeiro et al. (2016) はモデルの信頼について次のように述べている.

「ユーザがモデルまたは予測を信頼しないのなら, 彼らは使おうとしない. 信頼の定義を 2 つの異なる (しかし互いに関連する) ものに分けることは重要である. (1) 予測を信頼すること, つまりユーザからその予測に基づいて行動するに足る信頼を得られるかどうか, (2) モデルを信頼すること, つまりモデルが配備されたときに, ユーザにとって理由づけできる方法で行動を託せるほどの信頼を得られるかどうか[*44]」

つまり「解釈」「説明」という語にはさらに, 「信頼」という意味を暗示している. この観点で言うと, 私がここまでで挙げたものは結果を受け入れるかどうかではなく, モデルとモデルの出力がどう対応しているかの関係についてなんらかの合意があるということで, 後者の「モデルの信頼」に分類されそうだ.

さらに以下のツイートが興味深い.

- https://twitter.com/tmaehara/status/1148899315722543104

企業での機械学習活用の現場では, 「ニューラルネットその他の複雑なモデルは, 専門知識のない重役に理解し

[*43] 以前, 時系列モデルに関してもう少し具体的な例を書いた: 『あまりに暑いので、ごく簡単に Prophet の分析の質を向上させる方法を書いた』

[*44] *if the users do not trust a model or a prediction, they will not use it*. It is important to differentiate between two different (but related) definitions of trust: (1) *trusting a prediction*, i.e. whether a user trusts an individual prediction sufficiently to take some action based on it, and (2) *trusting a model*, i.e. whether the user trusts a model to behave in reasonable ways if deployed

てもらうには難しすぎるから簡単な決定木や重回帰モデルくらいしか使えない」というのはよく聞く話である. この場合, 非専門家を「納得」「説得」あるいは「信頼」させるために「説明」しているという意味ととれる[*45]. しかし, これらから示唆されるように, 納得させる方法は必ずしも理性や霊性に訴えかけるものでなくても良いことになる.

このように多義的であることから, 私は「解釈」「説明」といった用語を安易に使うことを避けている[*46]. 少なくとも, 日常言語と混同されるような場面や, 知識のバックグラウンドが異なる人と話す際には.

話がさらに逸れるが, 納得という概念はより深く考える余地がありそうだ. これは知識のない人間を説得するという話だが, では計量経済学や機械学習を使う専門家は「納得」して使っているのだろうか? 赤池 (2008), Akaike (2010) は面白い考察をしている. 彼自身の有名な研究である AIC の「なるべく単純なモデルほど評価される」という特徴を, **納得** (plausibility) という観点から再評価するというものである. これは厳密に定式化できない問題を具体例に挙げて言及した論文で, このようなテーマはむしろ分析哲学や科学哲学で議論されそうな問題のように思える. しかし機械学習にしろ計量経済学のフレームワークにしろ, 使う人間 (上記のように非専門家に説明する場面の話ではない) が「納得して」使っているのかというのと, ハイパーパラメータの調整や特徴量の変換を完全に自動化して, 特定のデータに対する対数損失や平均二乗損失の最小値を総当りで, あるいは他のアルゴリズムで探索するような作業 (これもかなり意地悪くカリカチュアされた表現だ) と決定的な違いがあるように思う. 後者はもはや, 試行錯誤というより全自動のくじ引き装置だ. 計量経済学でも機械学習の応用場面でも, 人間の「納得」によるヒューリスティックな探索が要求される余地がまだまだ残されていると思う.

ところで最近は, 「予測」という用語も誤解を招きかねないのではないかと考えるようになった. それは既にほのめかされているように, 機械学習の技術を総動員してできるのはあくまで経験損失の最小化までであり, 将来に対しての汎化は担保せず厳密な意味での (未来) 予測ではない.

とは言っても, 回帰モデルの右辺の変数は「説明変数」と呼ばれ, 目的変数の「予測値」のように用語として定着しているものもあるので, これも言い換えるというのは難しい. 一方で機械学習では学習したモデルをもとに予測値を計算することを「**推論** (inference)」と呼ぶことがある. はじめ私はこの用法に戸惑ったが, しかし「予測」に比べて「推論」は未来に対してコミットしているというニュアンスが薄まるので, 意外と良い表現にも思えてくる. しかし, この文中でも私は「推理」という単語を, 人間の理性によって仮定から結論を導き出す行為という意味で使っており, さらには機械学習の**学習**に相当する操作を統計学では**推定** (estimate) または**推測** (inference) というため紛らわしい[*47].

[*45] やや曖昧な雑感: たとえば, 予測が目的のはずで, ニューラルネットワークを使ったモデルに対しても「解釈できない」とのたまうケースがある, という不満を時に耳にする. しかしこれは, 本人も気づいてるだろうが, 「解釈できない」という言葉は, 文字通り解釈を求めているのではなく, アカウンタビリティを求めているのではないか. あるいはニューラルネットを駆使するエンジニアなり研究者なりは, 納得してニューラルネットを使っているのか, 理解して使っているのか. そういう観点から言えば, ここでの「解釈性」というのは決してバカにするものではない. 以前, 自動運転技術に関して製造技術者の責任範囲に関するニュースが話題になったが, 製造者の責任が全く問われないというのは従来の法律からしても不自然だろう. ただし, むやみに単純なモデルで見やすければアカウンタビリティも容易に得られる, という命題が正しいとは私は思わない. なのでやはり, 字義通り「解釈したい」のかどうかとは切り分けるべきだろう.

[*46] こういう問題に敏感になったのは, 最近になって機械学習についていろいろ教える立場になったことで, 分野ごとの用語の使われ方の違いに気づくことが増えたり, 教えた相手が単語のニュアンスの違いから意図せず誤解に至ったりすることがあったのが大きな原因だと考えている. そのため, 2014-2015 年あたりの私のブログの記事ではこういった単語をもっとためらいなく使っている.

[*47] この点は『パターン認識と機械学習』にも訳注で指摘されている. 厳密には統計学において推定と推測にも使い分けがあるらしいが, 詳細は忘れた.

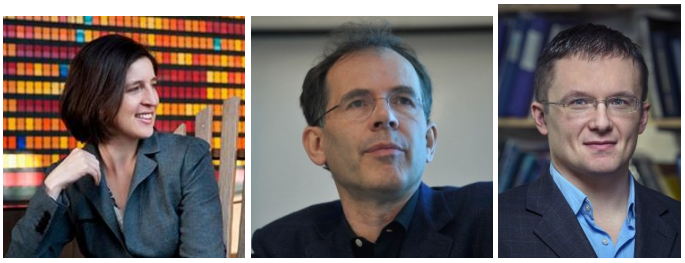

図8　(著者主観で) 近年注目される計量経済学者. 左から Athey, Imbens, Chernozhukov.

肖像転載元: Susan Athey, Guido Imbens, Viktor V. Chernozhukov

4　機械学習を呑み込む計量経済学

機械学習は経済学の問いに必ずしも答えられない. しかし部分的にはそのテクニックを活用できる場面がある. ここまでを踏まえて, 最近は機械学習を取り入れて計量経済学でどのような研究がなされているかをいくつか紹介していく.

特に近年は, 既に紹介した Susan Athey と Guido Imbens, そしてヴィクトル・チェルナジューコフ (Viktor Viktrovich Chernozhukov[*48]) の研究が目立つ (図 8, ただし選定基準は個人の感想. 他にも機械学習に関心を持って研究している経済学者は多く存在する.). 3 人のうち Imbens は因果推論と機械学習の利用を強調した計量経済学の講義を行ったり (Imbens, 2018b), Athey と共著で機械学習の利用を啓発する文章を書いているが, 現時点では機械学習を大きく取り入れた手法の研究発表は Athey らとの共著 (Athey et al., 2019b) のみである. その代わり, Deaton and Cartwright (2018) の批判したような因果推論の外的妥当性の問題に関する研究に熱心である.

そこで, ここでは Athey と Chernozhukov それぞれの特徴的な研究を紹介する.

4.1　異質性のある処置効果

機械学習を融合させることで進展した問題の 1 つに, **異質性処置効果** (*HTE; heterogeneous treatment effect*) の推定がある. これは Athey らの研究の大きな成果である. 前セクションで説明したように, ATE は名前の通り多くの個体の平均的な効果を表すものである. ATE は, 処置変数以外に結果に影響を及ぼす共変量の効果を除外することで, 全ての人間の, 処置による平均的な因果効果を求めるものである. しかし, むしろ個人ごとの処置効果を知りたい場合もある. つまり, ATE では除外してしまった個人の属性 (共変量) の影響をむしろ考慮したいのである.

Wager and Athey (2018), Athey et al. (2019a) ではそれぞれ, **因果木** (*causal tree/forest*) と**一般化ランダムフォレスト** (*GRF*) という, 決定木を利用して, HTE を求める手法を提案している. エッセンスだけを言うと, 2 つの手法いずれも決定木のリーフノードを使用して層別化することで, これまでの課題であった計算量の爆発を抑制している. また, 後者の研究の学術的な貢献としては, これまであまり研究されてこなかったランダムフォレストの予測の漸近分布の性質について証明したということも評価される. これらに日本語で言

[*48] 英語表記では彼は patronym を書かないが, NES によれば, Chernozhukov の patronym は Викторович である. https://www.nes.ru/academic-council

及しているものは今の所あまりない[*49]. 因果木については『木と電話と選挙 (causalTree)』, GRF については 中村 (2019), 『読書日記: 読了：Athley, et al. (2017) 一般化ランダム・フォレスト』, そして私の書いた『Generalized Random Forest (GRF) について』がある.

4.2 アンサンブル学習

Athey and Imbens (2019) はアンサンブル学習・モデル平均化は異なる表現能力を持つモデルを組み合わせることで単なる線形モデルよりも非線形関係や相互作用あるデータへの当てはまりが向上し, ひいてはサンプル外での当てはまりが向上する可能性がある[*50]ため, シンセティックコントロール (SC; synthetic control, Abadie et al., 2010) 法のような使い方ができるのではないかと述べている. 具体的な理論は Athey et al. (2019b) らが提案している. 既に指摘されている SC 法のいくつかの問題点に対してそれぞれ提案された, 3 つの修正モデルをそれぞれ推定した上で, 統合してメタ学習を実践している.

また, Brodersen et al. (2015) で提案された causal impact というフレームワークがある. これは**ベイズ構造時系列モデル** (BSTS, Scott and Varian, 2014[*51]) と DID や SC 法的なアイディアを応用したもので, R のパッケージとしての実装が著者らによって公開されている. ここでもモデルにアンサンブル学習を利用しているため, ある意味 Athey らに先駆けた提案と言える[*52].

4.3 Double Machine Learning

Mullainathan and Spiess (2017), Athey and Imbens (2019) は罰則付き回帰は説明変数の多いモデルの推定で活用の機会があると述べた. 変数の多いモデルに機械学習を活用するというアイディアに Chernozhukov が主に関わった研究として, Belloni et al. (2017), Chernozhukov et al. (2017), Chernozhukov et al. (2018) の**二重バイアス除去機械学習** (*DML; double/debiased machine learnig*) がある. しかしどれも証明が長大である. おおまかなアイディアを知るためには以下が参考になるだろう.

- 『機械学習 × 計量経済学：Double/Debiased Machine Learning』
- 『TokyoR #71 で LT 発表させていただきました. 』

モチベーションとしては, これもやはり ATE を推定するのが目的であるが, データが非常に高次元, つまり処置変数 D 以外の特徴量の種類が膨大であるケースを想定している.

$$Y = D\theta + g(X) + \varepsilon$$

さらに, 特徴量間にはなんらかの相互作用があるため, $g(X)$ のように, 形状の不明な関数として表している. このようなモデル自体は**部分線形モデル** (partially linear —) と呼ばれるセミパラメトリックモデルの一種として, かなり前から研究されていた[*53]. しかし X の次元が巨大な場合は推定が難しい. そこで関数形にこだわらず, かつ次元の呪いの影響を受けにくい手法として, 機械学習を導入したというのが DML である.

[*49] 後に 依田 (2019) においてもアイディアの大枠が言及されている.

[*50] アンサンブル学習の大まかなアイディアについては上記 GRF での私の記事を参照. また, アンサンブル学習に関する基本的な事項の説明や充実した関連研究のリファレンスを持つ教科書として『アンサンブル法による機械学習』(Zhou, 2012) がある.

[*51] bsts の解説も私は以前書いたことがある: 『bsts (ベイズ構造時系列モデル) パッケージの使い方』

[*52] なお, 私は causal impact に関する記事『CausalImpact でできること, できないこと』を以前書いた. かなり単純なシミュレーションではあるが, causal impact ひいては現代的な因果推論の限界について視覚的に説明したつもりである.

[*53] 私もこれまでノンパラメトリック・セミパラメトリックモデルの話をあまり追っていなかったのでこれに際していくらか調べたが, 末石 (2009) によれば研究の嚆矢は 90 年代に遡るようだ.

図9　Judea Pearl (肖像転載元)

脱線した話題 III: AI と経済理論の関係

計量経済学の方法論だけでなく, 経済理論の観点からも AI が産業にどう影響を及ぼすかの研究は盛んである. より正確には最近の AI ブームの前からの風潮で, IT やロボットの導入は労働市場にどう影響を与えるかという研究が以前からなされていた. 典型的なものが, 80 年台以降のアメリカで見られる労働者の賃金の二極化問題の原因を説明するもの, そして以前『「AI の正体は最小二乗法」記事を読み解く』で言及した, **方向付けられた技術変化** (DTC) **理論**に関係する, AI やロボットの導入が新たな雇用を創出するという研究である.

こういった研究テーマは最近出版された 山本 (2019)『人工知能と経済』によくまとめられている. この本は最近のブームからすると内容がやや保守的で, どちらかというと地に足の付いた内生的経済成長理論の応用研究の紹介が多く[*54], 機械学習や計量経済学のテクニカルな面にはそこまで深入りしておらず, また以前『「AI の正体は最小二乗法」記事を読み解く』で言及したような壮大な未来想像図を期待すると面食らうと思うが, これはこれで面白い内容である.

また, NBER[*55] においてThe Economics of Artificial Intelligence というカンファレンスが行われた, ということだけ示しておく. ここまで何度も引用した Athey (2018) のレビューもこのカンファレンスでの発表である.

この分野も調べれば AI とマーケットデザインとか面白い話題が出てくるのだが, そこまでサーベイしていては話が際限なく広がってしまうのでこれ以上は書かない.

5　Pearl の機械学習評論

5.1　3 つのステージ: Pearl 流の因果推論

もし数年以内に計量経済学の新しい教科書が出版されるとしたら, ここまで説明したような最近の研究動向が反映される可能性がある. しかしここで終わっては, 本文中で紹介したすぐれた研究者たちの評論の劣化コピペレポートでしかない. そこで, さらに踏み込んだ話をする.

[*54] 冒頭には総務省の AI ネットワーク化検討会議での編著者の議論と報告が執筆のきっかけとある.

[*55] 非経済学関係者に対して補足説明すると, National Bureau of Economic Research (NBER) はアメリカの民間研究組織である. この組織は規模が大きく, 研究助成金プログラムなども実施しており, 日本でいう日本学術振興会 (JSPS) や科学技術振興機構 (JST) に相当する役割を経済学分野に限定して果たしている.

Quora に『計量経済学と統計学と機械学習の違いは何か』というまさに今回のテーマと同じ問いが投稿されている.

回答者の中には, 既に紹介した経済学者の Angrist がいる. 彼の回答はオーソドックスで, 計量経済学は特に内生性の問題について考えてきた, ということに焦点を当てており, この記事で既に紹介したような自然実験のトピックをいくつか挙げている.

ここで大きく取り上げたいのは Judea Pearl (図 9) の回答である. Pearl の主張はより野心的である. 彼は以下のように述べている.

「計量経済学, 統計学, 機械学習の方法論を比較すると, 標準的な機械学習と, 発展的な機械学習の間には隔たりがあると言わざるを得ない. 標準的な機械学習とは, ディープラーニングやニューラルネットに代表されるような, 分布関数の性質からサンプルを引き出すというこれまでの統計分析が果たしてきた役割と全く変わらず, データの流れに関数をあてはめているだけのものである. それに対して, 発展的な機械学習とはデータを生成する分布を超えて, 政策の介入や反事実的な理由付け (例えば, 『もしこれとは異なることをしていたとしたら?』) を扱うことを可能にするものである[*56]」

ディープラーニングもニューラルネットも昔なじみの最小二乗法, つまり統計学や計量経済学と本質的に変わらない, というのは最近も似たような話[*57]を見たような気がするが, それはさておき, Pearl はなかなか挑戦的な主張をしている.

近年はディープラーニングもニューラルネットも研究が盛んである. (私はほとんどキャッチアップしていないが) 画像認識での応用を中心に次々と新しいテクニックが生まれているようだし, 統計学の極限理論の観点から理論的性質の研究も進んでいるらしい (Imaizumi and Fukumizu, 2018, Nakada and Imaizumi, 2019). では, Pearl がディープラーニングのことを従来の統計学と何ら変わっていないと一蹴したのはなぜか.

彼の回答の以降の記述は, Pearl (2019) の主張とほぼ同じである. これもなかなか野心的な主張をしており, 自身の因果推論フレームワークの要件を達成することが「**人間並みか, 人間を超える人工知能** (AI)」を開発するにあたっての課題になると述べている. つまり Pearl は, よくいわれる「ディープラーニングは特徴量の抽出も自動でやってくれる」というのが幻想であることを批判しているのか? そうではない.

Pearlは強い AI の要件として, 因果関係に対して答えられることを挙げている. 彼はその中で因果推論の強度を表1のような 3 段階で表現した. レベルの数字が大きいほどより難しい問いに答えられ, さらには下位レベルの問いにも答えることができるとしている. 彼はレベル 1 から順に, **関連** (*association*) , **介入** (*intervention*), **反事実** (*counterfactual*) と名付けた. レベル 3 の反事実的な問いに答えられるようになるのが, 強い AI の要件だという. 強い AI とは何か, というのはおそらく今も答えの出ていない哲学的な論争で, 単に Pearl が強い AI をそう規定しているだけだろうから, 今回は AI の定義の正しさには触れない. むしろ, 彼の構築した理論の観点では最近の機械学習と統計学の研究がどう位置づけられているのかが今回のテーマでは重要である.

[*56] "In comparing econometrics, statistics, and machine learning methodologies, one must distinguish between standard and advanced machine learning. The former, exemplified by deep learning and neural networks, fits a function to a stream of data and plays the same role as statistical analysis, taking us from samples to properties of distribution functions. Advanced machine learning, on the other hand, goes beyond distributions onto the process that generates the data, and so, allows us to manage policy interventions and counterfactual reasoning (e.g., 'what if we have done things differently')"

[*57] http://ill-identified.hatenablog.com/entry/2019/03/01/135627

レベル	名称	モデル	行為の例	問いの例
1	関連分析	$p(y \mid x)$	観察	観察された症状からどんな病気を読み取れるか? 選挙の出口調査から何を読み取れるか?
2	介入分析	$p(y \mid do(x), z)$	動作, 介入	もしアスピリンを飲んだら, 私の頭痛は治まるか? タバコを禁止したらどうなるか?
3	反事実分析	$p(y_x \mid x', y')$	予見, 回顧	アスピリンは私の頭痛を止めたか? もしオズワルドが狙撃しなければ, ケネディは生き延びられたか? 私が過去 2 年間禁煙していたらどうなっていたか?

表1　古典的な意味での違い. Pearl (2019) を一部改変.

5.2　Judea Pearl (2019) の因果推論観

さらに彼は既存のフレームワークを強い AI たらしめるために必要なトピックを 7 つ挙げて,「七つ道具」と称した.

1. **透明性** (*transparancy*) と **検証可能性** (*testability*)
2. **do 計算** (*do-calculus*)
3. **反事実のアルゴリズム化** (algorithmization of counterfacutals)
4. **直接・間接効果の評価** (assesment of direct and indirect effect)
5. **適応可能性** (*adaptibility*), **外的妥当性**と**サンプルセレクションバイアス**
6. **欠損データの復元** (*recovering from missing data*)
7. **探索** (*discocverity*)

これらは何を意味するのだろうか. このうち do 計算や直接効果・間接効果は彼の著作で頻出する概念である. 彼が自身の理論のすごさをアピールするために我田引水している感は否めないが, 彼の理論は事実含蓄がある.

先に Pearl の理論の基本的な背景を紹介する必要がある. Pearl (2009), Pearl and Mackenzie (2019) に集約された彼 (ら) の理論は興味深いが, 彼の理論はグラフ理論と, 彼の導入した独特の概念を用いて記述されているため, Rubin 的な理論と雰囲気が全く異なる. Pearl の提案する**因果 (構造) ダイアグラム** (causal structural diagram) あるいは**有向非巡回グラフ** (*DAG; directed acyclic graph*[*58]) のフレームワークは統計学的な数式で書かれたモデルの記述から従来の理論との対応関係を考えると難解だが, グラフで記述できることから視覚的には何を意味しているのかが分かりやすい. DAG 自体はそれ以前から存在したグラフ理論で記述されるが, Pearl は**バックドア・フロントドア基準**や **do 演算子**といった概念を導入することで因果関係を導き出す理論を構築した.

宮川 (2004) によれば, 1920 年代に研究された**パス解析**あるいは**構造方程式モデリング** (*SEM*) フレームワークが因果ダイアグラムの特殊形とみなせる. よって構造方程式モデリングを知っている人間ならば比較的すんなりと理解できるかもしれない. Pearl の理論は因果推論に関するかなり多くの概念を包括的に捉えていて,

[*58] ところで, 宮川 (2004) では Pearl の理論を因果ダイアグラムとして紹介し, DAG とは形式的に異なるものとして扱っているが, 最近の英語の文献では DAG と呼ぶケースが多い. これはなぜだろうか?

結果として操作変数法といった既存の概念と一致する部分が多いが, 独特の発展を遂げており, またGelman (2019) が, Pearl and Mackenzieによる統計学者たちの因果推論への功績を無視するかのような記述に対して反論意見を自身のブログに掲載したことからもわかるように, 本人の主張だけをもとに読み取っていては見誤る可能性がある (Pearl はいつも一言多い性格なのかもしれない).

Pearl (2009) の初版は日本語訳が存在し, いくつかの解説のために教科書や論文もでている. それでも彼の独特の理論はわかりにくいとされている (宮川). 簡単にアクセスできる参考文献としては, 大塚 (2012) による本書の書評があり, 津川 (2014) も簡易な解説をしている. 彼の理論をより詳しく解説する文献としては, 理論の前段階や関連トピックにも言及した宮川 (2004), Pearl と Rubin の理論の差異を明確にしつつ解説しようとしているものとして黒木・小林 (2012), 黒木 (2014) がある[*59]. しかしながらこの著者らは経済学者ではないため, 計量経済学での利用を念頭に置いた解説に乏しい. しかし, 私が去年の発表直前にこの原稿を書いているまさにそのとき, arXiv にImbens (2019) の論文が投稿された. 70 ページを超える内容 (本文は 50 ページ程度) だが, Twitter 上での議論にまで言及して, Pearl 理論と経済学で使われる, 多くの場合で Rubin 的なフレームワークとを比較して整理しようとする彼の解説は非常に有用である.

よって, ここで彼の理論の解説を 1 から書く意義は少ないから, Pearl (2019) の挙げる「七つ道具」の解説に最低限必要な話だけを書いておく. また, 構成の都合上, 彼が挙げた順序とは多少異なる.

まず,「**1: 透明性, 検証可能性**」は, 分析が科学的見地から必要な大前提である. これはあまり解説する必要がないだろう.

5.2.1 「3: do 計算」と「4: 直接・間接効果の評価」

Pearl 独自の概念で, $do(\cdot)$ という演算子で表わされる. 既存の確率概念に置き換えることはできるが, 単なる条件確率とは異なる. 単なる条件確率 $\mathrm{P}(Y \mid X = x)$ は表1では association, 単なる相関関係に対応し, $\mathrm{P}(Y \mid do(X), Z)$ は介入効果に対応する. 宮川 (2004) はこれを明確に, 因果ダイアグラムの構造と $X = x$ を所与とした時の周辺確率密度関数であるとしている[*60].

$$\mathrm{P}(Y \mid do(X = x), Z) := \int \frac{p_V(x, y, z)}{p_{X \cdot pa}(x \mid pa(x))} dz.$$

ここで分母が曲者である. do 計算と密接に関係する**バックドア基準**という概念を表している. 変数の中には, **交絡変数** (confounders) という[*61], 結果変数 Y にも, 他の説明変数 X にも相関する変数が存在しうる. 交絡変数が存在するとき, 単純な差分では因果効果を推定できない. 交絡変数を特定する簡単な判定ルールがバックドア基準である. このバックドア基準に基づいて交絡変数の効果を遮断 (*block*) した上で, という条件付けが分母に表れている. これにより, 因果ダイアグラムでは共変量から結果変数への**直接効果**と, 他の共変量を通じた**間接効果**を識別できるとしている.

ここではこれ以上厳密な説明をする意義がないので, 林 (2011; 2017) や岩波データサイエンス刊行委員会 (2016) の中での林・黒木による, グラフをまじえたバックドア基準と交絡変数の説明を見るのが良いだろう.

この定義から, Pearl のいう介入効果が Rubin の因果効果と非常によく似ていることがわかる. 両者の仮定の違いから生じるより細かい差異は黒木・小林 (2012) も指摘している.

[*59] 黒木はさらに教科書も出している (2017) ようだが, 私はまだ読んでいない.

[*60] ただし宮川は do 演算子に相当するものを set 演算子として紹介している. Pearl はこの概念に以前から言及していたが, 当時は用語や記法が統一されていなかったようだ. Pearl (1995) の定義では介入効果は $\mathrm{P}(Y \mid \hat{x})$ と表記され *do* という記号は見られず, 名称も causal effect としていた.

[*61] Rubin 流の因果推論を解説する際に, 説明変数は共変量とも呼ばれる, と書いた. 正確には共変量は交絡変数と同じ意味で使われることもあれば, 細かく定義を分けていることもあり, 文脈で判断することが多い.

5.2.2 「3: 反事実のアルゴリズム化」

「**3: 反事実のアルゴリズム化**」は, レベル 3 の反事実分析のことを言っている. 表の例はわかりにくいので, Pearl and Mackenzie (2019) の別の例を借りる. 反事実分析は「**アリスは高校を卒業してから 6 年間働き続け, 今の給料は y' である. もし仮に, 彼女が大学を卒業していたとしたら今の給料 y はいくらだろうか**」という問いに答えることである. Pearl (2009), Pearl and Mackenzie (2019) も細かい設定が異なるものの, 同じ例題を提示している. 彼らは自身のフレームワークであればこの問いに答えられるとしているが, Imbens (2019) は反事実分析と RCT を比較して, **RCT ではこの問いに答えることが難しい**と指摘している. RCT が答えられるのは, 大学を卒業した人間の母集団と卒業していない人間の母集団を比較して, 大学を卒業していない人間がもし卒業していた場合に**平均的な**給料がいくらになっていたか, である (平均値が気に入らないなら中央値で答えることもできる). しかし, **個人の**給料がいくらになったかについて, RCT はより強い仮定がなければ答えることができない[*62]. これは厳密な式を書かないと違いが分かりづらいかもしれない.

Imbens は彼らの記述にそって, この考えの問題点を解説していく. 以下の式(5.1)のような線形モデルを使って説明している.

$$\begin{aligned} Y_i =& \alpha_0 + \alpha_1 D_i + \alpha_2 W_i + \varepsilon_i, \\ W_i =& \beta_0 + \beta_1 D_i + \eta_i \end{aligned} \tag{5.1}$$

ここでは交絡変数の関係を明示的にしているため, 前セクションの ATE の説明のために提示した式(2.3)よりも複雑になっている.Y_i は給料, D_i は大学に卒業したかどうかを表す, つまり Rubin 理論でいうところの処置変数で, W_i は観察できない交絡変数である. ε_i, η_i は観察できない要因, 回帰分析の文脈で言うなら誤差項である.

最小二乗法で $\alpha_0, \alpha_1, \beta_0, \beta_1$ を推定すれば, この観察できない要因 ε_i, η_i は残差 $\hat{\varepsilon}_i, \hat{\eta}_i$ として, 個人ごとに異なる値が得られる. よって, アリスという個人に対しても $\hat{\varepsilon}_{Alice}, \hat{\eta}_{Alice}$ を得られる. すると, (5.1)からアリス個人の給料を

$$\hat{Y}_{Alice} = \alpha_0 + \alpha_1 + \alpha_2(\beta_0 + \beta_1 + \hat{\eta}_{Alice}) + \hat{\varepsilon}_{Alice} \tag{5.2}$$

から計算できる. これが Pearl の反事実分析である.

一旦 Rubin の考えを振り返る. (2.2)からわかるように, Rubin は処置変数 D の値が異なると, 残差の値も異なる可能性を考慮している. Pearl の例に沿うならば, $\{college, highschool\}$ の 2 通りである.

$$\begin{aligned} \hat{\varepsilon}_{Alice|highschool} =& W_{Alice|highschool} - \beta_0 - \beta_1, \\ \hat{\varepsilon}_{Alice|college} =& W_{Alice|college} - \beta_0 - \beta_1 \end{aligned}$$

しかし, Pearl の反事実分析は両者が等しいという前提である.

Imbensはあるいは**自己選択バイアス**を引き合いにしてこの仮定に問題があることを指摘している. 仮に, アリスが勉強ができ, 学習能力の高い人間であるとする. 大学進学には学費が必要だが, その問題は無視してアリスの進路が自身の意思以外のなにものからも制約も受けないとすれば, 将来の収入がより増えるであろう大学進学を選ぶのではないか. これは前提条件が逆でも同様に結果を想像できる. この場合,$\hat{\varepsilon}_{Alice|highschool}$, $\hat{\varepsilon}_{Alice|college}$ では後者の方が大きくなる可能性が高い.

この指摘から Pearl の反事実分析フレームワークには, 変数間の関係をはっきり明示できることが前提であるとわかる. Pearl の理論はより強い主張ができるが, そのぶん仮定による制約が強い.

[*62] その答えの 1 つが, 既に紹介したAthey et al. (2019a) の GRF だろう.

5.2.3 Pearl 理論の要約

Rubin 流の因果推論アプローチを採用する人でも, 変数間の複雑な交絡関係を表現するために Pearl のようなグラフネットワークに頼ることはよくある. 厳密には両者の仮定に違いがあるものの, 計量経済学でよく用いられる操作変数法は, むしろPearlのようにグラフの文脈で説明したほうが分かりやすいと思われる. また, Rubin 流のアプローチであっても ATE を推定することを反事実/counterfactual とか因果/causal といった言葉を使わずに, **介入**/*intervention* 効果と表現する例も増えている.

一方で彼の因果ダイアグラムは強い仮定を要求する. do 演算とバックドア基準の説明からわかるように, ネットワーク構造を予め明示しなければならず, (経済学の) 実証研究への応用に対して説得力が乏しいことが指摘されている (Imbens). Gelmanも自身のブログで, Pearl の理論は**因果の方向を記述する定性的な分析には役に立つが, 因果の大きさを記述する定量的な分析は現実的に難しい**, と述べている[*63].

5.2.4 「5: 適応可能性, 外的妥当性, サンプルセレクションバイアス」と「6: 欠損データの復元」

Pearl (2019) は普遍的な「ロバストネス」の問題を軽減する取り組みについて取り上げている. この意味を考えるにあたって, 上記の Quora での彼の回答で「データを生成する分布を超えて」というのがヒントになる. 第3節でも述べたように, ディープラーニングを含む標準的な機械学習および統計学は, あらゆるデータが同一の確率分布から生成されていると仮定している. よって, 決定論的な部分と確率的なノイズとを切り分けさえすれば, 過学習を防ぎ適切に推定できる, ということになる. しかし, 現実にはそうである保証などない. Pearl にとっては機械学習が本来想定していない状況でもある程度適切に機能する, 言い換えるならデータを生成する分布が変化しても当てはまりの良さを損なわない性質を意味すると思われる. つまり彼のいう「ロバストネス」は経済学での用法とよく似ている. 単純な機械学習の方法では, 所与のデータに当てはまるかだけを見ているので, まったく分布の異なるデータを与えられると一気に当てはまりを損ない, 有意義な予測ができなくなる可能性がある (ただし, 機械学習でなくとも, 一般にこういうモデルを作るのは難しい).

彼は AI 研究者たちがこの問題に対して**ドメイン適応** (*domain adaptation*), **転移学習** (*transfer learning*), **生涯学習** (*life-long learning*), **説明可能 AI** (*XAI; explainable-AI*) といったトピックの形で取り組んでいると指摘しつつも, これらの取り組みは一般的なロバストネスの問題解決に対する「副次的作業」にすぎず, もともと関連分析に過ぎないニューラルネットに入力データの分布を変化させるという表層的な変化を与えただけでは不十分だと主張している.

彼は言及してないが, これらはサンプルとサンプル外とで説明変数の分布 $P(X)$ が変化する**共変量シフト**という概念に一般化できる. 古くはShimodaira (2000) が共変量シフトをサンプルの重み付けで対処する方法について定式化しており (ただしジャーナルは統計学寄り), 杉山他 (2014) ではさらに目的変数のラベル比率が変わる「クラスバランス変化」問題への対処にも言及しているが, これらは共変量またはクラスバランスの変化の大きさが既知の場合に適用できる方法である. さらに言えば, 共変量シフトは Rubin の因果推論の文脈での傾向スコアに対応するとも言える. よって, Rubin の文脈ならばこれらのフレームワークは因果推論と形式的には同じではないかと私は思う.

「ドメイン適応」は, 私のサーベイ不足で研究の系譜を把握していないため, とりあえず 東工大原田研究室のサイトを参考にする. ここではドメイン適応を画像認識の問題として扱っており, 同じくPearlの挙げる「転移

[*63] 更に言うと, Pearl ではなく Rubin の流儀に乗っ取るなら, そもそも点推定を捨て, 上下界さえ特定できれば良いという発想である**部分識別** (*partial identification*) を応用すれば, かなり緩い仮定であっても因果効果の推定が可能になる. 部分識別は昔から研究されていたが, あまり知られていない. しかし奥村 (2015), 奥村 (2018) という日本語のすぐれた文献が存在する.

学習」の 1 カテゴリであることがわかる. 当初の動機としてはここに書かれているように学習に必要な膨大な教師データを確保しなければならない問題の解決策として考案されたのだろうが, source domain, target domain という 2 つのデータのグループの分布を調整して近づけるというアプローチは, 共変量シフトの問題と同じである. まだ技術的な課題は多く残されているようだが, 傾向スコア法などの既存の手法と性質に共通点が認められば, 因果推論の新たな方法として利用されるかもしれない.

とうとう引用元がウィキペディアになってしまうが, "Transfer learning" の記事では,「転移学習」とは「ある問題を解くことで得られた知識を別の, 関連する問題に転用する機械学習の問題」とされている[*64]. 転移学習は一般に分布の異なるデータへ適応させるテクニックを言うようだ. ドメイン適応の定義との厳密な違いがわからない (あるいはコンセンサスがないのかも) が, この特徴どおりならば, 因果推論と形式的には同じテクニックかもしれない (私が知っている他の転移学習の手法は fine-tuning くらいで, これが傾向スコア法とどう等価になるのかはすぐに考えつかない).

「生涯学習」は全くノータッチなのでよくわからないが, 彼の引く書籍のサイトによれば, 従来の与えられたデータからのみ学習する機械学習パラダイムとは対象的に, 人間のように開いた環境で自律的な学習をする学習パラダイムであり, **汎用人工知能** (AGI) の必要条件であるとされている (具体的なスキームを調べる余力がない).

「説明可能 AI」は例えば原が『【記事更新】私のブックマーク「機械学習における解釈性（Interpretability in Machine Learning）」』で多くの研究を挙げている.「説明可能 AI」はいわゆる「解釈性」と同義で, 第3節で私が言った「説明」と「納得」どちらの意味でも使われるようで, 指し示す範囲がかなり広い. テクニックとしても多岐にわたり, 非線形かつ複雑なニューラルネットワークをなんらかの方法で縮約, または局所的に線形近似することで「解釈」を得る方法, たとえば LIME(Ribeiro et al., 2016), 各レイヤーでのユニットの重要度を計算する LRP(Bach et al., 2015), 画像認識分野であれば入力画像に対応するユニットの発火をヒートマップで重ねた Grad-CAM (Selvaraju et al., 2017) が提案されていることは以前から知っている[*65]. しかし Pearlがこの文脈でこれらを持ち出す意図は私にはよくわからない.

一方で, 再び話を計量経済学的なテーマに戻すと,「欠損データの復元」も Rubin 的な因果推論アプローチを表す. 彼は例によって自分の枠組みなら欠損データの分布も説明できるという論調だが, 実証面での応用事例は星野 (2009), 高井他 (2016) などで紹介されるように Rubin 的なアプローチに基づくものが多いように思う. Pearlは触れなかったが, 画像認識の分野では入力画像を反転, 剪断, あるいはノイズを重ねることでデータを多用にする data augmentation[*66]という方法が考案されている. この項目で言及されているトピックの多くは, サンプルの偏りによる問題にまとめられそうだ. これをリサンプリングによって解決するというアイディアも, Rubin 流因果推論と似ている. 機械学習においても, 因果推論と同じ問題意識を持った研究があることがわかる.

5.2.5 「7: 探索」

彼はさらに, 自身のフレームワークに基づけば因果関係を探索できることを主張している. その中で,Shimizu et al. (2006) を挙げている. この研究に関しては, 著者の 1 人である清水 (2012, 2017) が日本語で解説している. 因果ダイアグラムや, その前身といえる構造方程式モデルは基本的に線形かつガウシアンという, 単純な回帰分

[*64] 朱鷺の杜でも似たような定義が挙げられている.

[*65] 今泉 (2019) も LIME に言及しているが, 総評として経済学研究にこれらの「解釈」に関する研究を応用するのは現時点では難しいとしている.

[*66] data augmentation の具体的な話に触れた論文は見つけられなかった. みんななんで引用しないの……

析で計算できる「都合のよい」設定である. 複雑高度な理論が優れていると言いたいわけではない. Imbensが指摘するように現状では応用の乏しいフレームワークである. 一方で清水らの提案は探索にとどまらず, よりノンパラメトリックなモデルへの拡張を理論的に整備することで, 具体的な応用の場面を明確にしているように見える.

6　機械学習の中の経済学

ここでは, 後の7節で紹介するIgami (2018) に触発されて, 経済学者が興味を持ちそうな機械学習の研究例を紹介していく.

6.1　AI の差別/公平性

米アマゾン社が社員の採用審査を AI にやらせるという取り組みがなされていたが, AI が女性差別をするということで使用を中止したという話が話題になった[*67]. 神嶌 (2017), Barocas et al. (2018) では, sensitive な特徴量, つまり性別・出自など先天的な属性を予測に使ってしまう問題について言及し, AI の公平性に関する研究をまとめている. この問題の対策として, 従来のように誤差を最小化するだけでなく, 公平性を担保するような指標を新たに開発し, モデルの評価に使用するというものがある.

近年のオンラインサービスと機械学習の組み合わせが生み出す新しい問題を小宮山 (2019) は挙げている.

「Sweeny 本人には犯罪歴がないにもかかわらず, Sweeny の名前で検索を行うと「Latanya Sweeny, Arrested?」というタイトルの広告が掲載されることを報告した. オンライン広告はユーザーのネット上での行動をもとに最適化を行う. (中略) つまり「アフリカ系の名前, Arrested?」という広告が「ヨーロッパ系の名前, Arrested?」という広告がより高い割合でクリックされたとすると, 前者はより多くのユーザに長期的に表示され, 後者は早い段階で消えるようにオンライン広告が最適化された可能性がある.」

このように, 機械学習の最適化がユーザーの行動に対するフィードバックの結果, 自己成就的な行動に誘導してしまうというのは, 全てのデータが独立であるという機械学習 (そして統計学の) ナイーブな仮定では想定していなかったのではないだろうか. また, 差別の問題につながるものではないが, **推薦エンジン**分野の研究では, Chaney et al. (2018) は**機械学習による推薦でユーザーの行動が誘導された結果, ユーザー行動の多様性が損なわれ学習やユーザーの効用にかえって悪影響が発生する**ことを**フィードバックループ**または**エコーチェンバー問題**と呼んで言及している. 推薦エンジン分野ではこの問題が興味を持たれているようで, 最近の例ではJiang et al. (2019) が理論モデルによる分析を進めている. この論文では, **フィルターバブル効果**という名称でも呼ばれている. 日本語のいくつかのニュースサイトでも海外記事の翻訳でこの問題に触れられているようなので, もう少し詳しい話を取り上げる[*68].

従来は推薦エンジンでは暗黙の前提として, t 時点で推薦したアイテム a_t に対する各ユーザーの興味が $\mu_{t+1}(a_t)$ で決まり, 興味を反映したフィードバック c_t が観測できるとした. フィードバック c_t に応じて, 推薦エンジン側の興味予測モデル $\theta_{t+1} = f(\theta_t, a_t, c_t)$ を更新し, 新しい θ_{t+1} から次の推薦アイテムを決めるというシステムになっている. フィードバックデータからこの f を推定し, 推薦するものを決めていたいという (つまり状態空間モデルのようなもの?). このモデルではユーザーの興味は推薦アイテムによって変化しないが, 著

[*67] 参考: ロイターの記事『焦点：アマゾンがＡＩ採用打ち切り、「女性差別」の欠陥露呈で』

[*68] 私は推薦エンジンに関わったことがない, しかしこの分析内容は推薦エンジンの具体的なアルゴリズムを意識する必要がないものに見える. むしろ古典的な時系列モデルを知っているほうが理解しやすいと思う.

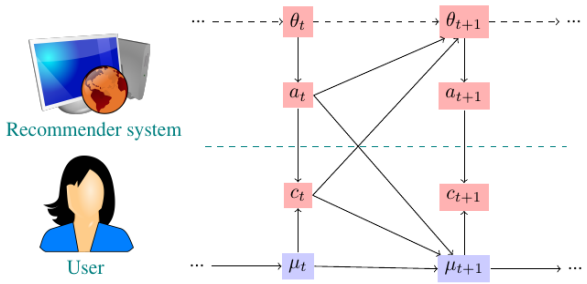

Figure 1: Model of interaction between a recommender system and user over time. Continuous and dashed links indicate existing or possible dependencies, respectively.

図10　エコー・チェンバー効果を分析するモデル (Jiang et al., 2019)

者らはユーザー側の興味も $\mu_{t+1}(a_t, c_t, \mu_t)$, つまり a_t だけでなく過去の推薦アイテムやフィードバックの影響を受けると仮定した場合でフィードバックループ問題を分析した (図 10).

当初のユーザーの興味分布と, 期間が十分過ぎたころのユーザーの興味分布の差が発散する, つまり当初と興味が全く変わってしまうことを「**興味の縮退**」と呼び, 興味の推移を非線形なランダムウォークモデルと仮定すれば, これがかなりゆるい条件下でも発生しうることを示した. つまり, 推薦エンジン次第でユーザーの興味の対象を操作し画一化できてしまうことになる.

次に, ランダムウォークではなくドリフト付き 1 階自己回帰モデルを仮定し, 推薦システムが完全にユーザーの興味を把握できる「オラクルモデル」の場合でも起こりうることを示した. さらに, より現実的な, 誤差のある推薦システムの場合も, 誤差がある程度大きくとも縮退が起こることをシミュレーションで確認した.

最後に, オラクルモデルや強化学習を応用した推薦システムでシミュレーションし, 縮退が起こる速さを比較している[*69].

私は推薦エンジンに関わったことがないので, どういうデータが使用できるかなどよくわからないが, 経済学のモデルでもなにかできそうな気がする. ただ経済モデルでよく使われる動学離散選択モデルは多項ロジットモデルだが, 推薦エンジンであつかうアイテム数は非常に大きいようなので, なんらかの工夫が必要だろう.

ところで経済学では, **統計的差別** (*statistical discrimination*) という研究トピックが存在する. ゲーリー・ベッカー (Gary Becker) やケネス・アロウ (Kenneth Arrow), エドマンド・フェルプス (Edmund Phelps) の理論が元になっており, 例えばCahuc and Zylberberg (2004, Ch. 5) で言及されている. 日本語の文献ならば大湾秀雄 (2017, 1 章) が以下のような簡潔な例を紹介している.

「統計的差別とは, 男性よりも女性の離職率が平均的には高いがゆえに, 企業側が女性への投資に慎重になることを指します. 女性の方がより離職率が高いという統計的事実をもとに, 合理的な意思決定を行っているのです. しかし, 統計的差別というのは実は自己成就的です.「女性は辞める確率が高いから投資をしない」という企業の意思決定が, 女性にとって継続就業の価値を下げ, 離職を促しているわけです.」

ベッカーの理論ではこの差別の理由が合理的ではないものと仮定され, 「嗜好 (taste) による差別」または「使

[*69] ところでこの論文, ほとんど確実な収束とか可測性とか, やたら細かく定義しているがモデルの大筋とあまり関係ないような……

用者差別仮説」と呼ばれる. しかし経済学では, 情報の非対称性など, 理由がある面で合理的であるがゆえに解決が難しいという店で, 「統計的差別仮説」を取り上げることが多い. さらに, 因果推論の失敗から差別の実態を誤認した実証研究の歴史がやはりCahuc and Zylberbergによって紹介されている. これらは最近関心を集めている AI による差別の問題とよく似ているように思える.

機械学習における公平性の取り組みでは, 公平性のスコアリングを提案している. これは経済学の研究では見られなかった取り組みである. 小宮山はこの問題に関して経済学者にとっても重要なのは「長期的なエージェントの振る舞いを分析するとともに, 公平性基準を導入することの厚生 *(welfare)* への影響を検討することである」としている. 第3節でも触れたように, 経済学ではフォワード・ルッキングなモデルの研究を進めてきた. AI による差別をフィードバックループの問題として捉えるならばこの研究蓄積が役に立つかもしれない.

6.2 ナウキャスティング

冒頭にも書いたように, 以前私は『ディープラーニング VS ディープパラメータ』という記事を書いた. 私は現象の構造そのものが変化するケースがあることを問題視して, 機械学習やディープラーニングの欠点を指摘し, また経済学ではその問題の解消を目指していることを主張した. **構造パラメータ** (または**ディープパラメータ**) とは, その構造変化を特徴づけるパラメータのことである.

しかしながら, 過去の私の記述は3節で書いたような目的の違いをはっきりさせていない. 計量経済学のモデルは仮説検証のために作られるのであり, 必ずしも予測誤差の最小化のために作られるのではない.

機械学習を利用する企業では, しばしば過去のデータをだけを当てはめて (未来のデータなんてないのだから当たり前である) 予測モデルを作っている. そして補外された未来の out-of-sample データに対する予測モデルの当てはまりは多くの場合時間の経過とともに悪化するので, 定期的にバッチ学習を繰り返す, あるいはオンライン学習を続けるといった運用がよく見られる. これはごく短期的な未来予測のみを目的としたモデルとも見なせるので**短期予測**あるいは**ナウキャスティング** (*nowcasting*) の一種と言えそうだ.

以前の記事を書いた時点では意識してなかったが, 転職して業務で機械学習を扱うようになって初めて, このような運用形態の是非を意識するようになった. そこで, 実用面をある程度意識してこの問題について考えを整理したい.

まず, 以前の投稿の反省から要件を明確にする. 計量経済学的な観点から言えば, **バイアスなく予測**できることが求められる. 一方でこれを達成するのは難しいので, 例えば常に不偏であることは保証できないが, ある程度**自律的に修正する作用がある**か, というのを第 2 の要件とする.

時系列解析の方法論の類推から, 例えば構造の変化が緩やかなものならば, 平滑化された関数で変化を近似できる. 簡単な例では**移動平均**モデルがある. 学習を繰り返すプロセスそのものを 1 つのモデルとして捉えれば, 学習につかうデータの期間をスライドさせて学習を繰り返すのは移動平均モデルと似ていないだろうか.

典型的な時系列解析の応用で, この問題を掘り下げてみる. データの特性を 2 通りに分けて考えてみる. 1 つは未来の結果は過去の情報に全く依存しない, **マルチンゲール**的なものだとするもの. もう一方は, 過去のデータに未来の予測に繋がる何らかの情報が隠されていて, それが未来の結果 y と関係がある場合. 前者ならば, 理論上特に何かできるということはない. 従来の時系列モデルでも, モデルの特定が正しいとしても捉えきれない誤差項を, マルチンゲール的な性質を持つ成分だと定義している. 例えばランダムウォークならば, 過去の標本平均値だけで期待値を近似できる. 長期的には分散が大きくなることが知られているが, 短期予測ならば比較的問題にならない[*70]. 一方で後者は, 従来の時系列予測モデル, 単純な例で言えば**自己回帰** (AR) モデルでも

[*70] 私の過去の記述では何をもって予測ができたとみなすのかの定義も曖昧だった. ランダムウォークが問題になるのは分散の大きさ

対処できる. では, その関係が単純ではなく, 複雑な非線形関係の場合はどうすればいいのか.

ここで, 『ディープラーニング vs...』で書いたようにディープラーニングの可能性についてもまじめに焦点を当てる. ニューラルネットは理論上は任意の関数を近似できる (Cybenko, 1989). しかし, 具体的にどのようなニューラルネットでならできるかと, それをデータからの学習によってどう達成するかは別の問題だ. どう学習するかという実践的な観点では, 実はすでに多くの研究が存在する. 例えばAlmosova and Andersen (2019) の研究を紹介する[*71].

著者らは, 消費者物価指数 (CPI) インフレ率の時系列データを使い, ランダムウォークモデルや自己回帰モデル, **マルコフ転換 AR** (*MS-AR*) モデル, **季節性 ARIMA** (*SARIMA*) といった従来の時系列予測モデルと比較して, 予測誤差の少ない**再帰ニューラルネット** (*RNN*, 実際には LSTM) を構築した. 彼らの研究で焦点となっている仮説の 1 つは, **ニューラルネットは周期成分を捉えられるために予測精度がよくなるのではないか**, である. AR やランダムウォークモデルは, 平均的な傾向を見ているだけであり, 時系列データにありがちな季節などの周期的な変化を表現できない. よって彼らは, 季節変化の顕著な金融時系列データを例に, MS-AR や SARIMA といった季節変化を捉えることができる従来のモデルとも比較している.

具体的な比較方法は次のようなものだ. FRB セント = ルイスから 1960 年 1 月 ~2018 年 7 月までの, 703 件の月次 CPI インフレ率データのうち, 90% を訓練に, 全体の 10% をテストにした. 金融時系列データは季節調整済みのものも合わせて公開されることが多いが, 上記のように周期成分を捉えられるかが現実のデータ予測でのカギであること, また季節調整処理は年間データで補正するためリアルタイム予測に使えないという理由で, 季節調整のない原系列データを使っている (同様の傾向がほかでもあるか, 個人消費支出データでも確認している).

なお, 著者らの引いた先行研究では, 従来的なモデルとニューラルネットの予測について研究したものもあるが, それらは**どのモデルのパフォーマンスが良いかは個別の問題とデータ次第である**, と結論づけている (当たり前と言えば当たり前だが).

著者らは実験によって, RNN は従来の予測モデルよりもかなり予測誤差が小さいが, 一方で数年以上先の長期予測誤差分散が大きいという発見している[*72]. また, よく言われているように, ニューラルネットはハイパーパラメータの調整が面倒であるとしている. よって「理論上は可能性があるが, どうやったらできるかはまだ分かっていないし, 誰も達成したかどうかを保証できない」ということになる.

最後に再び実用上の課題についてまとめる. 以上から, 理論上は複雑な時系列変化をするデータも予測できる可能性がある, いくつかの例では, ニューラルネットによって既存の予測モデルより予測がうまくいった事例がある. つまり, 単にスライドさせただけでうまくいく保証はない (逆に言えば, 定常的な現象に限定すればこのような方法でもうまくいくことが多い). もし変化を予測できるモデルを用意できれば, 5節で触れたShimodaira (2000), 杉山他 (2014) の方法を応用することができる.

リアルタイム自律的なシステム, 例えば強化学習などであれば, 常にバイアスのない予測ができるとは限らないが, 自律的であるため, 第 2 の要件を達成できる可能性がある[*73]. もし厳密に達成するとしたら, lifelong learning の実用化に等しいのではないか. つまり, 短期であっても, 構造の変化を意識して人の手で予測モデル

であり, ルーカス批判で問題としているのは法則が定まらないことである.

[*71] この論文は原稿であり, 査読を受けていないことに注意する. 大筋では興味深いが, ニューラルネットのモデル選択に BIC を使っているなど, 細かいツッコミどころがいくつかある.

[*72] 長期予測誤差分散が大きくなるということは, ニューラルネットといえど単に当てはめただけでは経済学が求めるタイプのモデルを作ることはできないのではないかという気がするが, 他にも検討すべき可能性はある.

[*73] そうなると, 非定常分布から生成されるデータに対する既存の強化学習アルゴリズムのパフォーマンスが気になるが, 今回はそこまで調べる時間がなかった.

を作ったり, モニタリングしたり, 必要な特徴量を変更したりする取り組みが無意味とは思えない.

6.3 A/B テストと RCT

私の知る範囲では, A/B テストは理論的な骨組みがあるわけではないが, 業界で漠然と使われてきた対照実験フレームワークのようだ. しかし, 第2節の因果推論の説明からわかるように, A/B テストを RCT に沿って行えば理論的根拠が生じる. Katsov (2018)[*74]は web 広告を始めマーケティングに応用できる初歩的な統計学的・機械学習的フレームワークの例を幅広く紹介しており, その中で A/B テストという名称こそ使っていないが, ランダム化実験と観察実験の例についても述べている.

第3節でDeaton and Cartwrightによる科学的な方法としての観点からの批判を紹介したが, 大学の研究者にしか関係ない問題かというとまったくそういうことはない. A/B テストあるいは RCT は「無作為にグループ分けして平均値を比較すれば因果推論ができてしまう」ものと安易に解釈されてしまいがちだが, 現実のデータに応用するに当たって様々な落とし穴が存在する. 一部の企業の研究チームは, A/B テストに独自の改良を施した例を機械学習関係の学会で発表している. 経済学や疫学, 心理学といった社会科学以外でもこのような問題意識が持たれているのは興味深い話である.

LinkedIn 社の研究チームであるSaveski et al. (2017) は, A/B テストが**ネットワーク効果**により RCT の基本的な仮定の 1 つである**処置効果の対象毎の安定性の仮定 (SUTVA)**[*75]を満たしていない可能性に着目した. 例えばある Web サイトについて通知機能のユーザーエクスペリエンス向上のために A/B テストを行っているとする. そしてあなたの友人が処置群に選ばれ, あなたは対照群に選ばれたとする. この場合, 友人は通知機能の変更を直接受けて, 滞在時間が増えるだろう. しかし, **あなたも友人を経由してそのサイトの更新情報を受け取ることで間接的に処置の影響を受けており**, SUTVA に違反している. 経済学では, *peer effect* などミクロなレベルの研究もあるが, **一般均衡効果**のような, マクロな観点で SUTVA が成り立ってないかどうかの話がよく取り上げられる. 参考として, 山名 (2017) は経済学で SUTVA が成り立たない簡単な例を紹介している. Saveski et al.はこの問題について, 自社 SNS サービスのユーザー間のつながり情報に基づいてユーザーをクラスタリングし, 層化を行うことで交絡を排除する方法を提案している. この際, 従来的な単純なランダム割り当てと提案方法とで結果を比較している[*76].

企業によっては, 同時にいくつもの A/B テストを実施していることもある. しかし, これが**サンプルセレクションバイアス**をもたらすこともある. Airbnb の研究チームであるLee and Shen (2018)[*77]はサンプルセレクションバイアスによってもたらされる過大な推定を「**勝者の呪いバイアス**」と呼んでいる (経済学者からすると, この用語はオークション理論のものとは違うので若干違和感があることだろう. ~~というか「生存バイアス」と~~

[*74] この本は pdf 版が無料で公開されている. 紙媒体の書籍は製本の質が悪いため, 紙媒体を好む人間でも pdf で読むことをおすすめする. なお邦訳も出版されているようだが, そちらは内容と製本の質を確認していない.

[*75] Stable-unit treatment value assumption. この語の日本語訳は戒能 (2017) によるものを採用した. Rubin (1990a) によればこの仮定は「全ての群 (処置群, 対照群) はそれぞれ同じ処置を受けその結果が安定的, 言い換えると受けた処置に対して同じ結果で, かつ他の群が受ける処置がなんであれ同じ結果である」ことを意味する. つまりこの仮定の下では, 各個体が互いに他の群に対して影響を与えないことになる. 処置が処置群を通して対照群にも交絡して作用しているなら, 明らかに対照実験の体をなしていないのでこの仮定は考えてみれば当たり前の話のようにも思えるが, 抽象的な言明なので数式に落とし込みにくい. そのためか因果推論の教科書でもあまり詳しく言及せず, 「**処置の条件付き独立の仮定** (CIA)」「**強く無視できる処置割り当て仮定** (SITA)」など数式で定義できるものに置き換えて解説されるものが多い.

[*76] 著者の 1 人であるSaint-Jacques (2019) の LinkedIn の公式ブログでの解説によれば, このアイディアは計量経済学で使われる, Hausman (1978) による**ハウスマン**検定に着想を得たという.

[*77] 内容を日本語で解説したスライドが存在する: https://speakerdeck.com/stakaya/lun-wen-du-nda-winners-curse-bias-estimation-for-total-effects-of-features-in-online-controlled-experiments

~~呼んだほうが分かりやすいのでは~~). 各 A/B テストが完全に独立していて, 全て同時に行われているならバイアスを気にする必要はない. しかし, 実際には A/B テストをして良いパフォーマンスの施策だけが生き残り, またしばらくした後で, また別の施策との A/B テストが行われるということが多い. A/B テストで実際に計算する評価指標は様々で, 一般化しようとするのは難しいが, Lee and Shenは単純な確率モデルを仮定してバイアスが発生することを証明し, その補正方法を提案した.

6.4 強化学習と適応的実験

ここまでは触れなかったが, Athey and Imbens (2019) は**強化学習** (reinforcement learning) にも言及している. 強化学習の特徴ははしばしば「**探索と活用** (exploration and exploitation)」と表現される. RCT のように純粋に結果を比較したいのであれば, 予め処置の割り当てを決め, データの全体を見てから計算する. このような比較実験をする理由の 1 つは, より効果の高い処置を見つけ出したいというものがある. よって, 単に効果を見極めたい (「探索」したい) というよりその効果を「活用」したいということになる. 強化学習は, 観測結果を逐次取得し, 能動的に効果の高いものへの割り当てを増やしていくアルゴリズムの総称である[*78].

強化学習の手法じたいは, かなり昔に発見されている. 例えば Thompson アルゴリズムは 1930 年代に遡る.

そんな中で, 2019 年 12 月のCausalML WorkshopでHadad et al. (2019) が**適応的実験** (*Adaptive Experiments*) でも利用できる介入効果の推定方法を提案している.

適応的実験とは, 2節で紹介したような RCT ではなく, 実験計画を状況に応じて「適応的に」変化させる手法のことらしい. 私は始めて聞いた単語だが, 「適応的実験計画」で検索すると, 日本語でもいくらか材料物性開発や, 医療エビデンスに関する論文や研究者ページがヒットする. 古くは 1940 年代に**逐次検定** (*sequential testing*) または**逐次解析**という, 通常の仮説検定とは違い有意な結果が出た時点での早期打ち切りができるようにする, というものがあるが, その流れだろうか. 具体的に適応的に変化させる方法として, 著者らは**強化学習の手法を挙げている**.

著者らの提案の新しいところは, 強化学習で適応的実験をする際の効果を測る際に単純に平均を計算したり傾向スコア法などを使っても推定結果にバイアスが発生することを指摘し, 因果推論のアイディアを使いバイアスを修正し, さらには**漸近分布の存在を証明したことで信頼区間の計算もできる**ようになった, という点である[*79].

強化学習の名前を出したが, 提案方法は Rubin の因果推論的な文脈に大きく依拠しているので, 強化学習のメカニズムじたいはあまり分かっていなくても良い. 適応的実験では, 時間が経過するごとに新しい観察データ $(W_t, Y_t(W))$ を入手できる. W_t は処置変数 D に対応するが, 今回は 0 か 1 かではなく, もっと多くのパターンがあると想定している. $Y_t(w)$ は結果変数である. では, ある 2 種類の処置 $W = w, w'$ 間での効果の差, $\Delta(w, w') := \mathrm{E}\left[Y(w) - Y(w')\right]$ を知るにはどうすればよいか.

RCT のように, つまり t 時点ごとに独立してランダムに割り当てる前提ならば, 単純に標本平均を計算するだけで良い. しかし著者らが想定するような強化学習を用いた適応的実験では, 過去の結果変数 Y の観測結果に適応して, 効果が高いとみなされたものの割り当てを自動的に多くしてしまう. これは明確に RCT の条件に反する. 因果推論のテクニックですぐ思いつくのは, 傾向スコア法による補正である. 強化学習は過去の結果に基づいて割り当てを決めるため, 過去の結果を共変量と見なせる. よって, 割り当て確率を

[*78] Sutton and Barto (1998) による強化学習の代表的な教科書では, 強化学習を「快楽主義的な学習システム」とも呼んでいる

[*79] 最近の研究でも, Johari et al. (2017) のようにオンライン A/B テストの結果に古典的な逐次検定を改良したものを利用して信頼区間を得ようという, さほどテクニカルない話がある. すると, こういった研究はいままであまりされてこなかったのだろう.

図11　Halbert White Jr. (肖像転載元)

$e_t(w) := \mathrm{P}(W_t = w \mid \mathcal{I}_{t-1})$ と書ける. $\mathcal{I}_t$ は履歴, つまり t 期間までの $\{(W_1, Y_1), \cdots (W_t, Y_t)\}$ の情報を表す. まずこの割り当て確率を推定して, そこから計算した IPW 推定量(6.1)の標本平均(6.2)を使えばいいという考えに至る.

$$\hat{\Gamma}_t(w) := \frac{1\{W_t = w\}}{e(w; \mathcal{I}_{t-1})} Y_t, \tag{6.1}$$

$$\hat{Q}_T(w) := \frac{1}{T} \sum_{t=1}^{T} \hat{\Gamma}_t(w) \tag{6.2}$$

このように傾向スコア補正すればうまくいくと思われがちだが, e が非常に小さい時, 推定量の分散が増大する. バンディットアルゴリズムは効果が小さいとみなした処置 w への割り当てを減らすので, そのような $\hat{Q}_T(w)$ の分散が大きくなる. 適応的実験の目的は推定結果の分散をできる限り早期に縮小することなので, この問題は深刻であると著者らは言う. 強化学習により, 後になって取得したデータほど割り当ての偏りが強くなっているはずなので, 傾向スコアが小さくなる. そこで, 逆に初期のデータにウエイトをかけた重み付け平均を取れば分散を補正できるというアイディアになる.

著者らは IPW 推定量にバイアス修正項を追加した**拡張 IPW** (Augmented IPW) 推定量なるものが存在することに着目し, AIPW 推定量をさらに一般化して, **評価ウエイト**という概念を導入し, **過去の値との重み付け平均**で推定する方法を提案している.

この推定量は, 現実的な条件で漸近的に正規分布となることを筆者らは数学的に証明し, シミュレーション結果を掲載している. ただし評価ウエイトの選び方については厳密には分かっておらず, うまく行きそうなものをいくつか提案するにとどまっている.

脱線した話題 III: 経済学者で最初に人工知能に注目したのは誰か?

多くの人は, 経済学者が機械学習に注目したのはここ最近の話で, 先進的な考えを持った経済学者として Athey や Imbens の名前を挙げるだろう. しかし, Igami (2018) によればディープニューラルネットワーク (DNN) を研究した経済学者がいる. 「DNN が十分複雑ならば, 任意の関数を近似できる」といういわゆる**万能近似定理** (*universal approximation theorem*) は, Cybenko (1989) が最初に研究したのだと私は思っていた. しかし, これとほぼ同時期にHornik et al. (1989) もまたニューラルネットの近似性能を研究している. 著者の一人であるハルバート・ホワイト (Halbert White Jr., 図 11) は計量経済学者としても有名で, White (1980) の提案した**不均一分散に対し一致性を持つ標準誤差** (*HCSE*) は多くの教科書で取り上げられている.

7　計量経済学と AI: 因果推論を超えた反事実分析

7.1　第 3 の反事実モデル

5-6節で紹介したように, 最近は機械学習を因果推論フレームワークに取り込む研究があれば, 機械学習にも因果推論的な考え方が取り入れられていたりもする. しかし, 私はひねくれているので, 以降はあえて違うタイプの研究をメインに紹介していく. 違うタイプといっても, 決してこれまでの研究とまっこうから相反するものではない. むしろ, Imbens にダメ出しされた Pearl の counterfactual とは別の意味で「反事実」分析ができるフレームワークだからだ. それは**動学的構造推定** (*dynamic structural model estimation*, あるいは単に**構造推定**と略される) と呼ばれるフレームワークである. ただし, 構造推定もまた因果関係を分析するフレームワークではあるが, 因果推論のカテゴリで紹介されることはない. これは経済学の問いに答えるために独自に発展した方法であること, また Rubin のような実験ベースの方法とは根本からアプローチが異なるという技術的な理由があると思われる.

Rubin 流因果推論, Pearl 流因果推論に続く第 3 の反事実分析フレームワーク[*80]である構造推定は, 5節で触れた構造方程式モデリングと名前が似ているが, 全く異なる. 更に言うと経済学では「構造方程式」というと経済モデルを表す同時方程式を指し, 因果推論のネットワークという意味付けはされていないことが多い (よって, 以降で**単に「因果推論」と言えば, 2節で説明した Rubin 流の因果推論のことを指すことにする**).

構造推定アプローチは, 経済学的な仮説に基づいて数理モデルを作成し, 実際の観察データを用いてパラメータを推定する. このとき, 因果推論のように RCT によってデータを取得する必要はない. なぜこのようなモデルを作るかというと, 第3節で触れた「ルーカス批判」に対処するためである. 詳しくは「因果推論」や「構造推定」といった方法論の妥当性に関する経済学界内での議論について, 今井他 (2001), 小林によるコラムや山名 (2017) が簡潔にまとめている. 経済セミナー編集部 (2016) にも, これまで紹介した因果推論 (構造推定と対比して誘導形と呼ばれる) と構造推定の比較を簡単に論じたものがある.

構造推定はここ最近 20 年間で研究がさかんになっているようだが, 専門的かつ体系的な教科書はあまりない[*81]. 日本語で言えば現在はおそらく山口 (2017) が一番参考になる[*82]が, そもそも構造推定というジャンルは経済モデルをどう計算するかという大きな括りなので, ミクロ経済学やゲーム理論などに基づく様々な理論モデルに依存しており, 網羅するのは困難だろう.

モデルによって適したアルゴリズムが変わるため, 数値的解法に関するものも定番の教科書がなく, 論文以外ではネット上に一般公開されていた講義ノート・スライドに頼っていた学生が多いのではないだろうか. 日本語なら基本的な事柄を抑えた工藤 (2007), 阿部 (2017) が有名だった. さっき見つけた楠田 (2019) は, そんな中で構造推定の代表的なアルゴリズムの基本的なところを解説している. この分野は数値シミュレーションをする理工系と同じで, 研究者は matlab か Fortran で書いたり, 有償の数理計画法ソルバソフトウェアを使うこと

[*80] ただし, Rubin 的な因果推論は最近は反事実 (counterfactual) ではなく, **介入** (intervention) と表現されることが多くなっている.

[*81] 私が学部生の頃にとある先生から日本語の授業を受けたことがあり, 当時の有用な講義ノートも公開されていたのだが, 今は公開されていない.

[*82] ただし, 山口の最後のスライドの文言は, 自然実験と構造推定が対立しているようなふしがあるが, これには注意が必要である. ただし, 私は口頭発表を聞いていないので, 当時どういう説明したのかは知らない. しかし山口先生がそんなおかしな主張はしないと思うので, 単に難解に思われ敬遠されがちな構造推定の啓発活動だろう. 反事実モデルは自然実験の結果と整合することが**必要条件**である. よって, 必ずしも両者は対立するものではない. これはDeaton and Cartwright (2018) の指摘とも重なってくるところであり, たとえばBlundell (2017) によって構造推定の結果を RCT で追試する重要さが説かれている. ただし, RCT できないときに必要に迫られて自然実験や構造推定をするというそもそもの動機も留意すべきだろう.

が多く (山口も言及しているように数値計算や計算機科学の知識があったほうがいい), matlab のサンプルコードが掲載されている. **しかし kindle はコピペできないという問題がある**. 経済学の理論モデルと密接に関係するため, 従来的なマクロ経済モデルの計算テクニックとも重なる点があり, 経済動学モデルの教科書, 例えばStokey et al. (1989), Ljungqvist and Sargent (2018), Barro and Sala-i-Martin (2004) のような伝統的な教科書も参考になるかもしれないが, 専門家でもないのにこのようなヘビーなものを買うのは敷居が高いだろう (私もこの辺のはあまり読んでない).

上記に挙げたものと重複するが, 『構造と識別〜構造推定と計量経済学に関するトピックを紹介する』というブログ記事でも参考になりそうな資料がリストアップされている.

7.2 AI と構造推定

この節では, Igami (2018) の解説をする. 構成を変えたり, 細かい文言を変えているが, 大筋で彼の主張と反したものにはなってないと思う[*83]. ただし, チェス・将棋・碁の AI の話は今回始めて知ったので, 読み違えはあるかもしれない.

7.2.1 Deep Blue と動学的最適化

Igami (伊神満) は, チェス用 AI Deep Blue, 将棋用 AI Bonanza, そして数年前に話題になった **Alpha Go といった AI が, 構造推定と数学的に同一である**と主張している. これらの AI の仕組みについて解説する資料はすでにいくつもあるが, 彼は構造推定との関係を強調しているため, 視点が異なる.

Igamiは, これらの AI が経済学の有名で古典的な論文の 1 つである Hotz and Miller (1993) とおなじフレームワークに基づいているとしている. Deep Blue (Campbell et al., 2002), Bonanza (保木・渡辺, 2007), AlphaGo (Silver et al., 2016) について, いずれも次のような抽象的な遷移式で一般化して表現できる.

$$s_{t+1} = f(a_t, s_t) \tag{7.1}$$

s_t とは t 時点でのゲームの状態を表す変数で, 例えば, 勝ちとか, 負けとか, 引き分けとかである. もうすこし細かくいうならば, チェス盤・碁盤・あるいは将棋盤のどこにどの駒があるかの状態を表す (盤面がはっきり分かれば勝ち負けも特定できる). a_t は t 時点でゲームの参加者が選んだ選択を表す. つまり, 次の手番 (ターン) での**状態** (*state*) s_{t+1} は, 直前の手番の状態と選択によって決まる, という想定である. a_t は, その時点の盤面で決まるから, ターン t で打てる手は s_t に依存して決まる集合 $\mathcal{A}(s_t)$ として表せる. 3 種類のボードゲームはいずれも, 対戦者どうしが同じ条件で交互に打つ手を選択するゲームだから, この定式化は一般に成り立つ. このような定式化は, 経済学でもよく使われる**ゲーム理論**的な考え方では**動学ゲーム** (または逐次ゲーム) と呼ばれるカテゴリである[*84].

勝利の状態にある時, $u_t = 1$ に, 負けなら -1 に, どちらでもないなら 0 とする. 次に行うのは $u_{t+1} = 1$ になる確率, つまり次の手で勝つ確率を**価値関数** (*value function*, aka **評価関数** *evaluation* —) $u_t = V(s_t, \boldsymbol{\theta})$ で評価することである. $\boldsymbol{\theta}$ は, この関数の形状を特徴づけるパラメータである. つまり, t 時点で取るべき最適な手 a_t^*

[*83] 著者本人が『経済セミナー』699 号で『ゲーム理論と AI—コンピュータ囲碁・将棋の事例』という題で寄稿している. 私は内容を確認していないが, 彼の論文の初稿は 2017 年に公開されているので, タイトルからしてこの論文をなにかしら解説していると思われる.

[*84] ゲーム理論は学生の時からまじめに勉強していないので適切な教科書を紹介しづらい. 教科書なら岡田 (1996), 一般公開されているもので動学ゲームの特徴を簡単に説明したものとしては横尾他 (2012a,b) がある.

は, この価値関数を最大化するものである. 専門用語では, 最適な手を**政策** (*policy*) といい, パラメータや状態と政策の関係を陽に表す関数 $a_t^* = \sigma(s_t; \boldsymbol{\theta})$ を**政策関数**と呼ぶ. 政策関数を求めるのが, このタイプの問題の最終目標である. (7.1) からわかるように, s_t は a_{t-1}, s_{t-1} に依存して決まる. そこで, Deep Blue ではターン t の手 a_t を決めるのに, L ターン後の状態までを計算して求めた. 形式的に書くなら, Deep Blue が計算すべき最適な手 a_t^* は以下のように定義される.

$$a_t^* := \arg \max_{a \in \mathcal{A}(s_t)} V(s_{t+L}; \boldsymbol{\theta}) \tag{7.2}$$

このような方程式を**ベルマン方程式**といい[*85], このようなタイプの最適化問題は, 応用数学の分野で**動的計画法** (*DP; Dynamic Programming*) または**動学的最適化** (—optimization) という.

チェスのようなルールのはっきりしている対戦ゲームの場合, 可能性のあるあらゆるゲーム展開は, 木構造で列挙できる. よって, 探索木のアルゴリズムを応用することで, この答えを知ることができる. ただし, すべての可能性を総当りで調べるのは現実的に不可能である. そこで, s_{t+L} から分かるように Deep Blue では数手先までだけを計算して決めている. この際の計算方法が, **ミニマックス探索法**と呼ばれるものである.

ただし, 探索の前に関数 V の形状をはっきり特定する必要がある. Deep Blue ではこの価値関数を特徴づけるパラメータ $\boldsymbol{\theta}$ が 8,150 個もあった. そして当時の計算機の性能の限界から, かなり泥臭い方法でパラメータを決定している. 過去の対戦データと, チェスのグランドマスターたちのアドバイスを元に, 部分的には探索アルゴリズムを利用したものの, $\boldsymbol{\theta}$ を**手動で調整して**勝率を高めていたという. つまり, 実態としてはありうる可能性の一部を総当り的に探索していただけで, **動的計画法の問題を厳密に解いていたわけではない**.

7.2.2 Bonanza

次に 2005 年に公開された将棋用 AI, Bonanza の話をする. 将棋はチェスよりルールが複雑である. 単純に盤面が広く, 駒が多いからというだけでなく, 捕った相手の駒を自分の手駒として使えるというルールにより, ゲームの進行に対して状態の組み合わせが単調減少しない. そのためパラメータも膨大となり, Deep Blue のように手動で調整できる数ではなくなる (Deep Blue の時点でも相当な労力だと思うが). チェスの状態数が約 10^{47} 通りなのに対し, 将棋は約 10^{71} 通りとなり, Deep Blue の価値関数のパラメータが 8,150 個だったのに対して, Bonanza のそれは約 50,000,000 個に増加している.

Bonanza では, (7.2)をアルゴリズムに基づいて計算している. この計算は 2 段階に分けられる. まず第 1 段階では, 観察されるデータとして約 50, 000 件の対戦データを使う. 1 局で平均して 100 手程度のため, データは約 5,000,000 件の (a_t, s_t) のペアになる. これに(7.2)を当てはめることで, 最適な選択 a_t^* を機械学習モデルの予測値として得る. 観察データよりパラメータが多いため, 機械学習でおなじみの正則化を用いている. ただし, 評価関数(7.2)の計算には L ターン後までの探索が必要なので, 既製品のライブラリでは計算できず, 最適化ソルバのプログラムの反復計算の中にネストして, なんらかの探索アルゴリズムで $V(s_{t+L}; \boldsymbol{\theta})$ を計算するような実装を自らする必要がある.

Igamiは Bonanza の方法がRust (1987) の提案した, 路線バスのエンジン交換の最適なタイミングを求める手法と似ていると評している[*86]. 彼が似ているとした点は, 以下の 2 つにまとめられる.

[*85] 正確には, ベルマン方程式の定義として一般的ではない. 阿部, 工藤の定義を参照.

[*86] 正確には, 2 人の対戦ゲームか, シングルエージェントか, 有限期間か, 無限期間か, という問題設定の違いはあるが, これは本質的な違いを産まない.

1. データに基づいた経験的 (empirical) な手法を使っていること. つまり, 従来の将棋 AI と違い, (個人の) 勘と経験に頼っていないこと.
2. 最適化を二重にネストしていること.

Rustの流れを書く. 彼の持つデータは, ある時点でエンジンを交換したかどうかと, バスの状態である. よって, Bonanza と同じく行動と状態のペア (a_t, s_t) があり, これを観察データとして $a_t = V(s_t;\theta)$ に当てはめる. V の形状はロジスティック回帰モデルを予め与えてあるので, 最尤法を用いる. 最尤法の計算 (ニュートン法などの反復アルゴリズムが使える) の各イテレーションで, $\boldsymbol{\theta}$ の暫定解 $\boldsymbol{\theta}'$ とそこから得られた予測値 a'_t を所与として, 条件に対応する以下の政策関数を最適化する.

$$a'_t = \sigma(s_t; V(\cdot;\boldsymbol{\theta}'))$$

この計算には, ***Value Function Iteration*** (*VFI*) というアルゴリズムが用いられる. **不動点定理** (縮小写像定理) の応用で, このアルゴリズムは解, つまり最適な政策関数に収束することができることが証明されている. 一連の計算手順は彼によって提案され ***nested fixed-point*** (*NFXP*) 法と名付けられている.

7.2.3 Alpha Go

そして Silver et al. (2016)[*87]による, 囲碁を対象とする Alpha Go においては, 状態数は約 10^{171} 通りとさらに膨大なものとなる. それだけではなく, 碁はゲーム中のある一手が有利になるか, 不利になるかの判断がプロでも難しい. よって, それが良い手か悪手かを判断するにはゲームの終端まで探索して結果を確認する必要が出てくる. すると, Bonanza のような方法で計算するのが難しくなる.

Alpha Go の手順は 4 つに分かれており, 第 1 のパートは, Bonanza 同様教師あり学習を用いている. ただし, ロジスティック回帰ではなく, **畳み込みニューラルネット** (*CNN*) でおこなう. なお, 使用したデータは 256,000,000 件, パラメータ数は 4,600,000 個である. 元の論文ではこの学習で得られた CNN に基づく政策関数を, *SL policy network* と呼んでいる. SL とは, **教師あり学習** (supervised learning) の略である. さらに第 2 のパートでは, **強化学習**[*88]によって SL policy network を改善している. つまり, 当初の CNN の学習で得た $\hat{\boldsymbol{\theta}}_{SL}$ を与えた政策関数 $\sigma(s_t;\hat{\boldsymbol{\theta}}_{SL})$ よりも勝率の高くなるような政策関数 $\sigma(s_t;\tilde{\boldsymbol{\theta}})$ に与える推定パラメータ $\tilde{\boldsymbol{\theta}}$ を求める, つまり, x, y が対戦して x が勝つ確率を $\mathrm{P}_{win}(x, y)$ とすると,

$$\mathrm{P}_{win}(\sigma(s_t;\tilde{\boldsymbol{\theta}}), \sigma(s_t;\hat{\boldsymbol{\theta}}_{SL})) > \mathrm{P}_{win}(\sigma(s_t;\hat{\boldsymbol{\theta}}_{SL}), \sigma(s_t;\hat{\boldsymbol{\theta}}_{SL}))$$

を満たすようなものである. しかし, 現時点では価値関数を定義できず, ゆえに従来の最適化問題の枠組みではパラメータを最適化することはできない.

そこで, 開発者たちは当初は様々な政策関数に基づく AI たちをランダムにマッチングして対戦させたり, さまざまな s_t をランダムに設定して勝ち抜いたものを優秀な候補として選別することにした. ここで選別されたものを $\hat{\boldsymbol{\theta}}_{RL}$ とする. ここで得られた政策関数を, *RL policy network* と呼んでいる. RL は強化学習の略である.

そして第 3 のパートは, 勝率を適切に評価できる価値関数 V の構築である. そこで, $\hat{\boldsymbol{\theta}}_{RL}$ から無数のゲームをシミュレーションで生成し, そこからリサンプリングしたデータをもとに, s_t から適切に勝利確率を予測でき

[*87] この論文を日本語で解説したブログがある: http://7rpn.hatenablog.com/entry/2016/06/10/192357

[*88] 私は強化学習に関して昔Sutton and Barto (1998) を少し読んだだけであまり詳しくないので, これが一般的な強化学習の定義にあてはまるのかよくわからない. そもそも強化学習も動学的最適化を応用しているので, 構造推定とも共通点がいくつかある. なおこの教科書は第 2 版 (2018) の草稿が無料公開されている.

AI	パラメータ推定	解の探索
Deep Blue	文字通り「手探り」	ミニマックス探索
Bonanza	ロジスティック回帰 + 正則化	ミニマックス探索
AlphaGo	ニューラルネット (CNN)+ 強化学習	モンテカルロ木探索 (MCTS)

表2　各 AI の仕組みの違い

るモデルを構築する. この結果, ある状況における $\sigma(s_t;\hat{\boldsymbol{\theta}}_{LR})$ どうしの勝率がはっきりと計算できる. この勝敗結果を教師データとして, 価値関数を推定することができる. 価値関数の推定には, やはり自由度の高いモデルということで CNN を用いている. この価値関数を近似したモデルは *value network* と呼ばれている.

第 4 パートは, AlphaGo が実際に対戦する際に使われる, **モンテカルロ木探索** (MCTS) というアルゴリズムである. value network は勝率を表すものなので, 最適な手がなんであるかを直接教えてはくれない. そこで Bonanza と同じように, 探索アルゴリズムが必要になる. MCTS の詳細はIgamiの引用する文献を参考にしてもらうとして, 彼は MCTS の特徴を「最適手の候補をシミュレーションで比較評価する[*89]」と表現している. これにより, 解があることは分かっているものの複雑すぎて厳密解を導けない問題の答えを近似できる. この近似は, 異なる近似の誤差をそれぞれ打ち消し合う「アンサンブル効果」で正確さを得る.

以上 3 種類の AI の違いを表2に示した.

7.2.4　AlphaGo の経済学的な解釈

Igamiは構造推定と Alpha Go を比較して, 同型関係を見出した. 第 1 のパート, SL policy network の推定はHotz and Miller (1993) で提案された 2 段階のアプローチ (以下, HM 法) の 1 段階目の推定に対応する. Bonanza で触れた NFXP 法は状態数が多い場合に計算量が多くなりすぎ, かつ反復計算中の $\boldsymbol{\theta}$ の値ごとに繰り返し計算する必要があった. 一方で HM 法の 1 段階目の方法とは, 「観察されたデータは与えられた状態の元での最適の選択を表している」という前提ならば, データに直接当てはめるだけで政策関数を推定できるというアイディアから出発している. よって, 厳密に動的計画法の問題を解く必要がない. ただし, モデルの制約を強くしないような方法で推定する必要がある. というのも, 分析の目的は価値関数のパラメータ θ (経済学の場合は, 選好とか, 技術構造といった経済現象の根本的なメカニズム) を求めることであり, それは観察されたデータから求めるものであって, 仮定のみによって得られるものではないからである. そういう意味で, HM の方法は**計算資源上の**次元の呪いを, **データの**次元の呪いに置き換えた手法である, と彼は評している. 実際, AlphaGo の SL policy network は, サンプル外での的中率が 55% であり, 一方で CNN より単純なロジスティック回帰モデルではたったの 27% だったというから, モデルの自由度が重要であることが分かる.

次の RL Policy netowrk をどう解釈するか. SL policy netowork $\sigma(s_t;\hat{\boldsymbol{\theta}}_{SL})$ は, いわば人間のトッププレイヤーによる戦略を近似したものだった. 一方で AI どうしを対戦させて構築した RL policy netowrk は $\sigma(s_t;\hat{\boldsymbol{\theta}}_{LR})$ より強い戦略を近似したものである. つまり, 実際には存在しない, 人間より強いプレイヤーを想定したデータを擬似的に観察したという, 「反事実的な」実験とみなせる.

Igamiはこの工程と, 続く第 3 段階の policy network の推定の工程は, HM 法の改良であるHotz et al. (1994) による手法 (以下, HMSS 法) の第 2 段階と同じであるとしている. HM 法では 2 段階目で, 状態数の多い場合の計算負荷が問題となっていたため, HMSS 法では直接解くことをやめ HM の第 1 段階の推定に

[*89] a similar state-evaluation task by simulating the outcomes of the candidate moves

基づいた前向きのシミュレーションを複数実行することを提案している. この複数回シミュレーションがAlphaGo の成功の原因ではないかと彼は考えている.

第 4 の MCTS は, 最適な手というよりはある程度はランダムな手を選択する粗雑な戦略である. つまり, AlphaGo の実態は, (1) トップレベルの人間の選手の戦略をさらに強化したもの, (2) そこから導かれる価値関数, (3) より単純化したリアルタイムの「拙速な」シミュレーション, という 3 種類の戦略のハイブリッドである.

さらに, AlphaGo の後に発表された AlphaGo Zero は, 現実のプロ棋士の対戦データを使わずに, 完全にシミュレーションだけでオリジナルを上回るパフォーマンスを示したことで, 大きな反響を呼んだ. つまり, 初期バージョンでの第 1 のパート, SL policy network を省いて, 白紙の状態からいきなり RL policy netowrk の構築を開始したのだ.

因果推論の文脈で言えば, これは観察データをそのまま反映するような推論は因果推論になりえないという話に繋がるどころか, 従来の計量経済学の範疇も超えている. しかし, AlphaGo Zero 成功の理由付けは経済学的な観点で理由付けできる. policy network (政策関数) と value network (価値関数) は**双対的**, つまり片方からもう一方を導けるため, 冗長である. MCTS によるアンサンブル効果は追加的なパフォーマンスの向上をもたらしたかもしれないが, これは究極的には各パーツの近似誤差の現れである.

AlphaGo は碁の対局において人間のチャンピオンに勝利したという点で,「**人間を超える AI** (強い AI)」であることを見せつけた. しかし, これは碁という限定されたゲームのルールの中での成果である. Igamiはさらに, 構造推定アプローチあるいは経済理論の観点から, AlphaGo にはいくつかの明示されていない制約が含まれていることを指摘している. 事実, HMSS 法の提案後にも, より完全な答えを出すためさらに全く異なる方法が提案されている. これらが「強い AI」開発のヒントに繋がるのではないかと彼は考えている.

また, 逆に経済学側にとっての発見として, Alpha Go の設計でニューラルネットワークを使ったことは構造推定にも応用できるのではないかと彼は考えている. 構造推定の 1 段階目でも自由度の高いノンパラメトリックモデルを当てはめるのが望ましいためだ.

ところで, 冒頭の私の過去の記事で, ディープラーニングではルーカス批判に対応できるモデルを作れない (あるいは選好などの経済現象の根本的なパラメータを特定できない) という主張をした. 今回明らかにしたその理由付けは, 限られた (そして偏った) データの学習だけではできないというものである. 一方でIgamiの指摘によれば, 同じ命題のために用いられている構造推定はニューラルネットを利用した AI との共通点が見られる. よって, 単なる当てはめだけでできるとは言えないものの, 適切な使い方をすればニューラルネットは経済学で求められるモデルを作るために有用かもしれない, と修正することになるだろう.

8　結論とまとめと反省

今回は私のポエム披露と近年の計量経済学・機械学習の研究サーベイを行った. 機械学習は計量経済学と補完しあえるのか, 相容れないものなのか, という当初の問いに対しては, まず計量経済学を因果推論と仮説検証で特徴づけ, 機械学習を関数の近似と特徴づけた上で, 先行するサーベイ論文とマイポエムを引きつつ, 3節において, 現時点では両者は目的の違いから等価ではないものの, 部分的には技術を流用できると結論づけた.

続いて Judea Pearl の視点を交えて, 近年の機械学習で関心を呼んでいる研究トピックをいくつか取り上げ, それらが経済学者が長年研究してきたテーマと共通点があることを指摘した. 最後に, 碁の AI, Alpha Go が計量経済学特有の反事実分析フレームワークである動学的構造推定と共通しているというIgami (2018) の論文を取り上げた. これまで強調してきた計量経済学と機械学習の目的の違いという話に反して, 動学的最適化と

いうテクニックの共通点だけでなく, 人間の行動の正確な記述という観点から, 計量経済学 (構造推定) と AI 研究が最終的に行き着こうとする場所がかなり近いことを彼は示唆した.

Igamiはまた, こう述べている.

「計量経済学の暗黙の仮定を緩めることは, モデルを現実的なものとするだけでなく, より解釈可能なものにもする. 構造モデルを開発し推定することの利益の *1* つは, より単純な設定の下での原因と相関という基本概念 *(*例えば, 何らかの変数 X, Y の間の統計的な関係性*)* を超えたその先で, 結果を経済学的に解釈可能にすることだ. 『解釈可能』と『説明可能』という言葉は様々な分野で異なる意味で用いられているようだが, 『構造的な解釈可能性』の概念はよりフォーマルな定義への指針として有用に思える. これは『構造的な解釈可能性』と「*DNN* を説明すること」への挑戦とを混同すべきではないという提案である[*90]」

この言明からは重要な示唆がいくつも見られる. 私はこの記事の6節までは, 計量経済学あるいは機械学習で定式化されたモデルの範疇で, 相関関係や因果効果の大きさを論じる方法論の特性に言及してきた. その強調のために「**ニューラルネットや XGBoost ではなく重回帰モデルの係数のほうが有意義**」などというカリカチュアライズされた命題すら唱えた. しかし, 彼の主張からわかるように本質的には線形近似にこだわることもまたナンセンスである. 彼の提案する「解釈可能性」はより拡張された概念であり, LIME や Grad-CAM に代表されるようなニューラルネットの「解釈性」研究とも明らかに異なる指針付けである.

Tokyo.R での発表だから当初はなにかデモンストレーションができたらいいと考えていたが, 膨大な先行研究のサーベイで持ち時間を使い果たし, 今また自分で課した締め切りを 6 日も超えてまで書き続けることになっている. この記事をアップロードする数時間前にも, 新たな参考文献を Amazon で購入するという計画性のなさだった. さらに私は集中力がないため, この記事を書いている途中に他のネタを思いついて脱線してしまう (実際の作業時間は 2 週間相当の余暇時間にも届かないかもしれない). 特に今回はサーベイ論文の体で書いているため, なおさらごった混ぜが許容されやすい. ところどころに「脱線した話題」というセクションがあるのも, この記事を書いている途中に生まれては消えていったネタのうち, 比較的関係のありそうなものの残滓である. 前半の計量経済学に関するサーベイで終わるならまだしも, 独特で難解と評判の Pearl の因果推論を Rubin 因果推論や機械学習と相対化して論じるなどという難物に挑むのは無計画すぎて, 明らかにただ恣意的にピックアップした記述の羅列になっている (実際, Pearl の節は他とのつながりがあまりない. いっそこの箇所をまるごと削除したほうが冗長にならずに済んだかもしれない). さらに後半の話題ほど露骨に引用文献が減り, 伝聞口調になっていることからも息切れしていることがわかる (ナウキャスティングのセクションでナウキャスティングの話をしていない).

私の現在の仕事からすれば, Edelman et al. (2005), Ostrovsky and Schwarz (2011), Ostrovsky and Schwarz (2016) の話にでも言及したほうがよかったのかもしれないが, 後の祭りである.

参考文献

Abadie, Alberto and Maximilian Kasy (2017) "The Risk of Machine Learning," January, retrieved from *here*.

[*90] "Relaxing the implicit econometric assumptions would make the models not only more re- alistic, but also more interpretable. One of the benefits of developing and estimating a structural model is that the results are economically interpretable, above and beyond the basic notions of causation and correlation in simpler settings (e.g., determining a statistical relationship between some variables X and Y). The words 'interpretable' and 'explainable' could mean different things in different fields, but the concept of 'structural interpretability' seems useful as a guide for a more formal definition. Note this proposal about structural interpretation should not be confused with the chal- lenge concerning 'explaining DNNs.'"

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010) "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, Vol. 105, No. 490, pp. 493–505, June, DOI: 10.1198/jasa.2009.ap08746.

Akaike, Hirotugu (1974) "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, pp. 716–723, DOI: 10.1109/TAC.1974.1100705.

——— (2010) "Making Statistical Thinking More Productive," *Annals of the Institute of Statistical Mathematics*, Vol. 62, No. 1, pp. 3–9, February, DOI: 10.1007/s10463-009-0238-0.

Almosova, Anna and Niek Andersen (2019) "Nonlinear Inflation Forecasting with Recurrent Neural Networks," retrieved from *here*.

Angrist, Joshua D. (1990) "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *The American Economic Review*, Vol. 80, No. 3, pp. 313–336, retrieved from *here*.

Angrist, Joshua D. and Victor Lavy (1999) "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, Vol. 114, No. 2, pp. 533–575.

Angrist, Joshua D and Jörn-Steffen Pischke (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*: Princeton University Press, retrieved from *here*, （大森義明・小原美紀・田中隆一・野口晴子訳, 『ほとんど無害な計量経済学 –応用経済学のための実証分析ガイド–』, NTT 出版, 2013 年）.

Angrist, Joshua David and Jörn-Steffen Pischke (2015) *Mastering 'metrics: The Path from Cause to Effect*, Princeton ; Oxford: Princeton University Press.

Angrist, Joshua, Guido Imbens, and Alan Krueger (1995) "Jackknife Instrumental Variables Estimation," Technical Working Paper 172, National Bureau of Economic Research, Cambridge, MA, DOI: 10.3386/t0172, Publication Title: Journal of Applied Econometrics.

Ashenfelter, Orley (2008) "Predicting the Quality and Prices of Bordeaux Wine," *The Economic Journal*, Vol. 118, No. 529, pp. F174–F184, June, DOI: 10.1111/j.1468-0297.2008.02148.x.

Ashenfelter, Orley and Cecilia Rouse (1998) "Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins," *The Quarterly Journal of Economics*, Vol. 113, No. 1, pp. 253–284, February, DOI: 10.1162/003355398555577.

Athey, Susan (2017) "Beyond Prediction: Using Big Data for Policy Problems," *Science*, Vol. 355, No. 6324, pp. 483–485, February, DOI: 10.1126/science.aal4321.

——— (2018) "The Impact of Machine Learning on Economics," in *The Economics of Artificial Intelligence: An Agenda*: University of Chicago Press, pp. 507–547, retrieved from *here*.

Athey, Susan and Guido W. Imbens (2017) "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives*, Vol. 31, No. 2, pp. 3–32, May, DOI: 10.1257/jep.31.2.3.

——— (2019) "Machine Learning Methods That Economists Should Know About," *Annual Review of Economics*, Vol. 11, No. 1, pp. 685–725, August, DOI: 10.1146/annurev-economics-080217-053433, ArXiv version:1903.10075.

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019a) "Generalized Random Forests," *The Annals of Statistics*, Vol. 47, No. 2, pp. 1148–1178, April, DOI: 10.1214/18-AOS1709.

Athey, Susan, Mohsen Bayati, Guido Imbens, and Zhaonan Qu (2019b) "Ensemble Methods for Causal Effects in Panel Data Settings," NBER Working Paper w25675, National Bureau of Economic Research,

Cambridge, MA, p. w25675, DOI: 10.3386/w25675.

Ayres, Ian (2007) *Super Crunchers: Why Thinking-by-Numbers Is the New Way to Be Smart*, New York: Bantam Books, （山形浩生訳,『その数学が戦略を決める』, 講談社, 2007 年）, OCLC: ocn122973719.

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek (2015) "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLOS ONE*, Vol. 10, No. 7, July, DOI: 10.1371/journal.pone.0130140.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2018) *Fairness and Machine Learning*: fairmlbook.org, retrieved from *here*.

Barro, Robert J. and Xavier Sala-i-Martin (2004) *Economic Growth*, Cambridge, Mass: MIT Press, 2nd edition, （大住圭介訳,『内生的経済成長論 (<1>, <2>)』, 九州大学出版会, 2006 年）.

Belloni, Alexandre, Victor Chernozhukov, Iván Fernández-Val, and Christian Hansen (2017) "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, Vol. 85, No. 1, pp. 233–298, DOI: 10.3982/ECTA12723.

Bishop, Christopher M. (2006) *Pattern Recognition and Machine Learning*, Information Science and Statistics, New York: Springer, retrieved from *here*, （村田昇・元田浩・栗田多喜夫・樋口知之・松本裕治訳,『パターン認識と機械学習: ベイズ理論による統計的予測 (上・下)』, 丸善出版, 2007 年）.

Blundell, Richard (2017) "What Have We Learned from Structural Models?" *American Economic Review*, Vol. 107, No. 5, pp. 287–292, May, DOI: 10.1257/aer.p20171116.

Box, George E. P. (1966) "Use and Abuse of Regression," *Technometrics*, Vol. 8, No. 4, p. 625, November, DOI: 10.2307/1266635.

Brodersen, Kay H., Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott (2015) "Inferring Causal Impact Using Bayesian Structural Time-Series Models," *The Annals of Applied Statistics*, Vol. 9, No. 1, pp. 247–274, March, DOI: 10.1214/14-AOAS788.

Cahuc, Pierre and André Zylberberg (2004) *Labor Economics*: MIT Press.

Cameron, AC and PK Trivedi (2005) *Microeconometrics: Methods and Applications*, Cambridge: Cambridge University Press, DOI: 10.1017/CBO9781107415324.004.

Campbell, Murray, A.Joseph Hoane, and Feng-hsiung Hsu (2002) "Deep Blue," *Artificial Intelligence*, Vol. 134, No. 1-2, pp. 57–83, January, DOI: 10.1016/S0004-3702(01)00129-1.

Card, David and Alan B. Krueger (1994) "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania," *American Economic Review*, Vol. 84, No. 4, pp. 772–793, retrieved from *here*, NBER Working Paper Version: 10.3386/w4509.

Chaney, Allison J. B., Brandon M. Stewart, and Barbara E. Engelhardt (2018) "How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility," in *Proceedings of the 12th ACM Conference on Recommender Systems - RecSys '18*, pp. 224–232, Vancouver, British Columbia, Canada: ACM Press, DOI: 10.1145/3240323.3240370.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey (2017) "Double/Debiased/Neyman Machine Learning of Treatment Effects," *American Economic Review*, Vol. 107, No. 5, pp. 261–265, May, DOI: 10.1257/aer.p20171038.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018) "Double/Debiased Machine Learning for Treatment and Structural

Parameters," *The Econometrics Journal*, Vol. 21, No. 1, pp. C1–C68, February, DOI: 10.1111/ectj.12097.

Cramer, J.S. (1987) "Mean and Variance of R2 in Small and Moderate Samples," *Journal of Econometrics*, Vol. 35, No. 2-3, pp. 253–266, July, DOI: 10.1016/0304-4076(87)90027-3.

Cybenko, George (1989) "Approximation by Superpositions of a Sigmoidal Function," *Mathematics of Control, Signals, and Systems*, Vol. 2, No. 4, pp. 303–314, December, DOI: 10.1007/BF02551274.

Dawid, A. P. (2000) "Causal Inference without Counterfactuals," *Journal of the American Statistical Association*, Vol. 95, No. 450, pp. 407–424, June, DOI: 10.1080/01621459.2000.10474210.

Deaton, Angus and Nancy Cartwright (2018) "Understanding and Misunderstanding Randomized Controlled Trials," *Social Science & Medicine*, Vol. 210, pp. 2–21, August, DOI: 10.1016/j.socscimed.2017.12.005, NBER working paper version: 10.3386/w22595.

Edelman, Benjamin, Michael Ostrovsky, and Michael Schwarz (2005) "Internet Advertising and the Generalized Second Price Auction: Selling Billions of Dollars Worth of Keywords,"Technical Report w11765, National Bureau of Economic Research, Cambridge, MA, p. w11765, DOI: 10.3386/w11765.

Ezekiel, Mordecai (1930) "The Sampling Variability of Linear and Curvilinear Regressions: A First Approximation to the Reliability of the Results Secured by the Graphic "Successive Approximation" Method," *The Annals of Mathematical Statistics*, Vol. 1, No. 4, pp. 275–315, November, DOI: 10.1214/aoms/1177733062.

Fan, Jianqing and Runze Li (2001) "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, Vol. 96, No. 456, pp. 1348–1360, December, DOI: 10.1198/016214501753382273.

Farrell, Max H., Tengyuan Liang, and Sanjog Misra (2019) "Deep Neural Networks for Estimation and Inference,"Technical report, arXiv: 1809.09953.

Gelman, Andrew (2019) ""The Book of Why" by Pearl and Mackenzie," January, retrieved from *here*.

Hadad, Vitor, David A. Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey (2019) "Confidence Intervals for Policy Evaluation in Adaptive Experiments," November, arXiv: 1911.02768.

Hahn, Jinyong, Jerry Hausman, and Guido Kuersteiner (2001) "Higher Order MSE of Jackknife 2SLS," retrieved from *here*, Publication Title: Order A Journal On The Theory Of Ordered Sets And Its Applications.

Harman, Gilbert and Sanjeev Kulkarni (2007) *Reliable Reasoning: Induction and Statistical Learning Theory* in , The Jean Nicod Lectures, No. 2007, Cambridge, Mass: MIT Press, （蟹池陽一訳,『信頼性の高い推論\, 帰納と統計的学習理論』, 勁草書房, 2009 年）, OCLC: ocm73926873.

Harrel, Frank (2018) "Road Map for Choosing Between Statistical Modeling and Machine Learning," September, retrieved from *here*, 和訳: 西田勘一郎『統計のモデルと機械学習のモデル、どう使い分ければよいのか』.

Hastie, Trevor, Robert Tibshriani, and Jerome Friedman (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*: Springer, 2nd edition, retrieved from *here*, （杉山将・井出剛・神嶌敏弘・栗田多喜夫・前田英作・井尻善久・岩田具治・金森敬文・兼村厚範・鳥山昌幸・河原吉伸・木村昭悟・小西嘉典・酒井智弥・鈴木大慈・竹内一郎・玉木徹・出口大輔・冨岡亮太・波部斉・前田新一・持橋大地・山田誠訳,『統計的学習の基礎 ―データマイニング・推論・予測―』, 共立出版, 2014 年）.

Hausman, Jerry A. (1978) "Specification Tests in Econometrics," *Econometrica*, Vol. 46, No. 6, p. 1251,

November, DOI: 10.2307/1913827.

Heckman, James J. and Sergio Urzúa (2010) "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify," *Journal of Econometrics*, Vol. 156, No. 1, pp. 27–37, May, DOI: 10.1016/j.jeconom.2009.09.006.

Hernán, Miguel A., John Hsu, and Brian Healy (2019) "A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks," *CHANCE*, Vol. 32, No. 1, pp. 42–49, January, DOI: 10.1080/09332480.2019.1579578, 西田勘一郎による要約: 『予測と因果関係は何が違うのか - Part 1』『予測と因果関係 - Part 2: 予測は自動化できても因果推論は自動化できない』.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989) "Multilayer Feedforward Networks Are Universal Approximators," *Neural Networks*, Vol. 2, No. 5, pp. 359–366, January, DOI: 10.1016/0893-6080(89)90020-8.

Hotz, V. J., R. A. Miller, S. Sanders, and J. Smith (1994) "A Simulation Estimator for Dynamic Models of Discrete Choice," *The Review of Economic Studies*, Vol. 61, No. 2, pp. 265–289, April, DOI: 10.2307/2297981.

Hotz, V Joseph and Robert A. Miller (1993) "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies*, Vol. 60, No. 3, pp. 497–529, DOI: 10.2307/2298122.

Igami, Mitsuru (2018) "Artificial Intelligence as Structural Estimation: Economic Interpretations of Deep Blue, Bonanza, and AlphaGo," March, arXiv: 1710.10967.

Imaizumi, Masaaki and Kenji Fukumizu (2018) "Deep Neural Networks Learn Non-Smooth Functions Effectively," in *AISTATS*, retrieved from *here*.

Imbens, Guido (2018a) "Understanding and Misunderstanding Randomized Controlled Trials: A Commentary on Deaton and Cartwright," *Social Science & Medicine*, Vol. 210, pp. 50–52, August, DOI: 10.1016/j.socscimed.2018.04.028, working paper version: here.

——— (2018b) "Causal Inference and Machine Learning," June, retrieved from *here*.

Imbens, Guido W. (2019) "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics," *arXiv:1907.07271 [stat]*, July, arXiv: 1907.07271.

Imbens, Guido W. and Donald B. Rubin (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*: Cambridge University Press, DOI: 10.1017/CBO9781139025751.

Jiang, Ray, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli (2019) "Degenerate Feedback Loops in Recommender Systems," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society - AIES '19*, pp. 383–390, Honolulu, HI, USA: ACM Press, DOI: 10.1145/3306618.3314288.

Johari, Ramesh, Pete Koomen, Leonid Pekelis, and David Walsh (2017) "Peeking at A/B Tests: Why It Matters, and What to Do about It," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, pp. 1517–1525, Halifax, NS, Canada: ACM Press, DOI: 10.1145/3097983.3097992.

Katsov, Ilya (2018) *Introduction to Algorithmic Marketing: Artficial Intelligence for Marketing Operations*: Grid Dynamics, retrieved from *here*, （株式会社クイーヴ訳,『AI アルゴリズムマーケティング自動化のための機械学習/経済モデル、ベストプラクティス、アーキテクチャ』，インプレス，2018 年）, OCLC: 1047855023.

Lalonde, Robert J. (1986) "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, Vol. 75, No. 4, pp. 604–620, retrieved from *here*.

Lee, Minyong R. and Milan Shen (2018) "Winner's Curse: Bias Estimation for Total Effects of Features in Online Controlled Experiments," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, pp. 491–499, London, United Kingdom: ACM Press, DOI: 10.1145/3219819.3219905.

Ljungqvist, Lars and Thomas J. Sargent (2018) *Recursive Macroeconomic Theory*, Cambridge, Massachusetts: MIT Press, 4th edition.

Lucas, Robert E. (1976) "Econometric Policy Evaluation: A Critique," *Carnegie-Rochester Confer. Series on Public Policy*, Vol. 1, No. C, pp. 19–46, DOI: 10.1016/S0167-2231(76)80003-6.

Merler, Silvia (2018) "Machine Learning and Economics," November, retrieved from *here*.

Mullainathan, Sendhil and Jann Spiess (2017) "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, Vol. 31, No. 2, pp. 87–106, May, DOI: 10.1257/jep.31.2.87.

Nakada, Ryumei and Masaaki Imaizumi (2019) "Adaptive Approximation and Estimation of Deep Neural Network with Intrinsic Dimensionality," November, arXiv: 1907.02177.

Ohtani, Kazuhiro and Hikaru Hasegawa (1993) "On Small Sample Properties of R^2 in a Linear Regression Model with Multivariate t Errors and Proxy Variables," *Econometric Theory*, Vol. 9, No. 3, pp. 504–515, June, DOI: 10.1017/S0266466600007805.

Ohtani, Kazuhiro and Hisashi Tanizaki (2004) "Exact Distributions of R2 and Adjusted R2 in a Linear Regression Model with Multivariate t Error Terms," *JOURNAL OF THE JAPAN STATISTICAL SOCIETY*, Vol. 34, No. 1, pp. 101–109, DOI: 10.14490/jjss.34.101.

Ostrovsky, Michael and Michael Schwarz (2011) "Reserve Prices in Internet Advertising Auctions: A Field Experiment," in *Proceedings of the 12th ACM Conference on Electronic Commerce - EC '11*, p. 59, San Jose, California, USA: ACM Press, DOI: 10.1145/1993574.1993585.

——— (2016) "Reserve Prices in Internet Advertising Auctions: A Field Experiment," Working Paper 2054, Stanford Graduate School of Business, Stanford, CA, USA, p. 23, retrieved from *here*.

Pearl, Judea (1995) "Causal Diagrams for Empirical Research," *Biometrika*, Vol. 82, No. 4, p. 669, December, DOI: 10.2307/2337329.

——— (2009) *Causality: Models, Reasoning, and Inference*, Cambridge, U.K. ; New York: Cambridge University Press, 2nd edition, retrieved from *here*, （黒木学訳,『統計的因果推論–モデル・推論・推測–』, 共立出版，2009 年）, 初版訳.

——— (2019) "The Seven Tools of Causal Inference, with Reflections on Machine Learning," *Communications of the ACM*, Vol. 62, No. 3, pp. 54–60, February, DOI: 10.1145/3241036.

Pearl, Judea and Dana Mackenzie (2019) *Book of Why - the New Science of Cause and Effect.*, OCLC: 1107553348.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016) ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pp. 1135–1144, San Francisco, California, USA: ACM Press, DOI: 10.1145/2939672.2939778.

Rubin, Donald B. (1974) "Estimating Causal Effects of Treatments in Randomized and Nonrandomized

Studies.," *Journal of Educational Psychology*, Vol. 66, No. 5, pp. 688–701, DOI: 10.1037/h0037350.

——— (1990a) "Formal Mode of Statistical Inference for Causal Effects," *Journal of Statistical Planning and Inference*, Vol. 25, No. 3, pp. 279–292, July, DOI: 10.1016/0378-3758(90)90077-8.

——— (1990b) "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies," *Statistical Science*, Vol. 5, No. 4, pp. 472–480, November, DOI: 10.1214/ss/1177012032, On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.

Rust, John (1987) "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica*, Vol. 55, No. 5, p. 999, September, DOI: 10.2307/1911259.

Saint-Jacques, Guillaume (2019) "Detecting Interference: An A/B Test of A/B Tests," June, retrieved from *here*.

Saveski, Martin, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M. Airoldi (2017) "Detecting Network Effects: Randomizing Over Randomized Experiments," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, pp. 1027–1035, Halifax, NS, Canada: ACM Press, DOI: 10.1145/3097983.3098192.

Schwarz, Gideon (1978) "Estimating the Dimension of a Model," *The Annals of Statistics*, Vol. 6, No. 2, pp. 461–464, March, DOI: 10.1214/aos/1176344136.

Scott, Steven L. and Hal R. Varian (2014) "Predicting the Present with Bayesian Structural Time Series," *International Journal of Mathematical Modelling and Numerical Optimisation*, Vol. 5, No. 1/2, p. 4, DOI: 10.1504/IJMMNO.2014.059942.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra (2017) "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, Venice: IEEE, October, DOI: 10.1109/ICCV.2017.74.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell (2001) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston: Houghton Mifflin.

Shieh, Gwowen (2008) "Improved Shrinkage Estimation of Squared Multiple Correlation Coefficient and Squared Cross-Validity Coefficient," *Organizational Research Methods*, Vol. 11, No. 2, pp. 387–407, April, DOI: 10.1177/1094428106292901.

Shimizu, Shohei, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen (2006) "A Linear Non-Gaussian Acyclic Model for Causal Discovery," *Journal of Machine Learning Research*, Vol. 7, pp. 2003–2030, retrieved from *here*.

Shimodaira, Hidetoshi (2000) "Improving Predictive Inference under Covariate Shift by Weighting the Log-Likelihood Function," *Journal of Statistical Planning and Inference*, Vol. 90, No. 2, pp. 227–244, October, DOI: 10.1016/S0378-3758(00)00115-4.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis (2016) "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature*, Vol. 529, No. 7587, pp. 484–489, January, DOI:

10.1038/nature16961.

Stokey, Nancy L., Robert E. Lucas, and Edward C. Prescott (1989) *Recursive Methods in Economic Dynamics*, Cambridge, Mass: Harvard University Press.

Sutton, Richard S. and Andrew G. Barto (1998) *Reinforcement Learning: An Introduction*, Adaptive Computation and Machine Learning, Cambridge, Mass: MIT Press, 1st edition, （三上貞芳・皆川雅章訳, 『強化学習』, 森北出版, 2000 年）.

——— (2018) *Reinforcement Learning: An Introduction*, Adaptive Computation and Machine Learning Series, Cambridge, Massachusetts: The MIT Press, 2nd edition, the draft is here.

Thistlethwaite, Donald L and Donald T Campbell (1960) "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment.," *Journal of Educational Psychology*, Vol. 51, No. 6, pp. 309–317, DOI: 10.1037/h0044319.

Tibshirani, Robert (1996) "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58, No. 1, pp. 267–288, January, DOI: 10.1111/j.2517-6161.1996.tb02080.x.

Tiptree, James, Jr (1973) *Love Is the Plan The Plan Is Death*: Ballantine Books, retrieved from *here*, 邦題『愛はさだめ、さだめは死』.

Varian, Hal R. (2014) "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, Vol. 28, No. 2, pp. 3–28, May, DOI: 10.1257/jep.28.2.3.

——— (2016) "Causal Inference in Economics and Marketing," *Proceedings of the National Academy of Sciences*, Vol. 113, No. 27, pp. 7310–7315, July, DOI: 10.1073/pnas.1510479113.

Wager, Stefan and Susan Athey (2018) "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, Vol. 113, No. 523, pp. 1228–1242, July, DOI: 10.1080/01621459.2017.1319839.

White, Halbert (1980) "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, Vol. 48, No. 4, pp. 817–838, DOI: 10.2307/1912934.

Wooldridge, Jeffrey M (2010) *Econometric Analysis of Cross Section and Panel Data*: The MIT Press, 2nd edition, retrieved from *here*.

Wooldridge, Jeffrey M. (2018) *Introductory Econometrics: A Modern Approach*, Mason, OH: Cengage Learning, 7th edition.

Yin, Ping and Xitao Fan (2001) "Estimating R^2 Shrinkage in Multiple Regression: A Comparison of Different Analytical Methods," *The Journal of Experimental Education*, Vol. 69, No. 2, pp. 203–224, January, DOI: 10.1080/00220970109600656.

Zhang, Cun-Hui (2010) "Nearly Unbiased Variable Selection under Minimax Concave Penalty," *The Annals of Statistics*, Vol. 38, No. 2, pp. 894–942, April, DOI: 10.1214/09-AOS729.

Zhao, Peng and Bin Yu (2006) "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, No. 7, pp. 2541–2563, retrieved from *here*.

Zhou, Zhi-Hua (2012) *Ensemble Methods: Foundations and Algorithms*: CRC Press, （宮岡悦良・下川朝有訳, 『アンサンブル法による機械学習 — 基礎とアルゴリズム —』, 近代科学社, 2017 年）.

Zou, Hui and Trevor Hastie (2005) "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 67, No. 2, pp. 301–320, April, DOI:

10.1111/j.1467-9868.2005.00503.x.
赤池弘次 (2008)「「納度」概念の利用について」,『統計数理』, 第 56 巻, 第 2 号, 253–258 頁, retrieved from *here*.
阿部修人 (2017)「上級マクロ経済学講義ノート動的計画法」, Technical report, retrieved from *here*.
依田高典 (2019)「経済分析のツールとしての機械学習」,『経済セミナー E-Book 機械学習は経済学を変えるか』, 日本評論社,『経済セミナー』711 号より.
伊藤公一朗 (2017)『データ分析の力: 因果関係に迫る思考法』, 光文社, 東京, OCLC: 990298721.
——— (2018)「データ分析を経営や政策に生かすには?「因果分析」と「予測」の適切な使い分け」, 8 月, retrieved from *here*.
今井晋・有村俊秀・片山東 (2001)「労働政策の評価: 「構造推定アプローチ」と「実験的アプローチ」」,『日本労働研究雑誌』, 第 497 号, retrieved from *here*.
今泉允聡 (2019)「機械学習はデータを解釈できるか」,『経済セミナー E-Book 機械学習は経済学を変えるか』, 日本評論社,『経済セミナー』711 号より.
岩波データサイエンス刊行委員会 (2016)『岩波データサイエンス』, 第 3 巻, 岩波書店, retrieved from *here*, OCLC: 1082871961.
大塚淳 (2012)「因果と実在, Judea Pearl, Causality, 第二版書評」,『科学基礎論研究』, 第 39 巻, 第 2 号, 109–115 頁, DOI： 10.4288/kisoron.39.2_109.
大森義明 (2008)『労働経済学』, 日本評論社.
大湾秀雄 (2017)『日本の人事を科学する: 因果推論に基づくデータ活用』, 日本経済新聞出版社, 東京, OCLC: 992155946.
岡田章 (1996)『ゲーム理論』, 有斐閣.
奥村綱雄 (2015)「部分識別とその応用: 処置効果を中心に」,『日本経済学会春季大会』, 5 月, retrieved from *here*.
——— (2018)『部分識別入門: 計量経済学の革新的アプローチ』, 日本評論社, 東京, retrieved from *here*, OCLC: 1057483434.
戒能一成 (2017)「政策評価のための横断面前後差分析 (DID) の前提条件と処置効果の安定性条件 (SUTVA) に問題を生じる場合の対策手法の考察」, RIETI ディスカッションペーパー 17-J-075, 産業経済研究所.
神嶌敏弘 (2017)「公平配慮型データマイニング技術の進展」,『第 31 回人工知能学会全国大会論文集』, 一般社団法人人工知能学会, DOI： 10.11517/pjsai.JSAI2017.0_1E1OS24a1.
川野秀一・松井秀俊・廣瀬慧 (編) (2018)『スパース推定による統計モデリング』, 第 6 号, 共立出版, retrieved from *here*.
楠田康之 (2019)『経済分析のための構造推定アルゴリズム』, 三恵社, 名古屋, OCLC: 1129899352.
工藤教孝 (2007)「動学的最適化入門」, Technical report, retrieved from *here*.
久保川達也・江口真透・竹村彰通・小西貞則 (1993)「統計的推測理論の現状」,『日本統計学会誌』, 第 22 巻, 第 3 号, 257–312 頁, DOI： 10.11329/jjss1970.22.257.
黒木学 (2014)「統計的因果推論における原因の確率とその評価」,『統計数理』, 第 62 巻, 第 1 号, 45–58 頁, retrieved from *here*.
——— (2017)『構造的因果モデルの基礎』, 共立出版, 東京都文京区, retrieved from *here*.
黒木学・小林史明 (2012)「構造的因果モデルについて」,『計量生物学』, 第 32 巻, 第 2 号, 119–144 頁, DOI： 10.5691/jjb.32.119.

黒澤昌子 (2005)「積極労働政策の評価–レビュー」,『フィナンシャル・レビュー』, 第 77 巻, 197–220 頁, 7 月, retrieved from *here*.
経済セミナー編集部 (2016)『経済セミナー増刊進化する経済学の実証分析』, 日本評論社, retrieved from *here*.
小西貞則・北川源四郎 (2004)『情報量規準』, 予測と発見の科学, 第 2 号, 朝倉書店.
小宮山純平 (2019)「機械学習に潜む公平性の問題」,『経済セミナー E-Book 機械学習は経済学を変えるか』, 日本評論社,『経済セミナー』711 号より.
五島圭一・高橋大志・山田哲也 (2019)「自然言語処理による景況感ニュース指数の構築とボラティリティ予測への応用」,『金融研究』, 第 38 巻, 第 3 号, retrieved from *here*.
清水昌平 (2012)「構造方程式モデルによる因果推論: 因果構造探索に関する最近の発展」,『行動計量学会第 40 回大会』, 9 月, retrieved from *here*.
——— (2017)『統計的因果探索』, 機械学習プロフェッショナルシリーズ, 講談社, 東京都文京区, retrieved from *here*.
末石直也 (2009)「セミ・ノンパラメトリックモデルと内生性」,『経済論叢』, 第 183 巻, 第 2 号, 87–97 頁, 4 月.
杉山将・山田誠・ドゥ・プレシマーティヌス・クリストフェル・リウソン (2014)「非定常環境下での学習：共変量シフト適応, クラスバランス変化適応, 変化検知」,『日本統計学会誌』, 第 44 巻, 第 1 号, retrieved from *here*.
杉山将 (2013)『イラストで学ぶ機械学習: 最小二乗法による識別モデル学習を中心に』, 講談社, 東京, retrieved from *here*, OCLC: 860876211.
高井啓二・星野崇宏・野間久史 (2016)『欠測データの統計科学』, 岩波書店.
竹村彰通 (1991)『現代数理統計学』, 創文社現代経済学選書, 創文社, retrieved from *here*.
田中章詞・富谷昭夫・橋本幸士 (2019)『ディープラーニングと物理学原理がわかる、応用ができる』, 講談社, 東京, OCLC: 1112379761.
田中司朗 (2019)「医学のための因果推論の基礎概念」,『計量生物学』, 第 40 巻, 第 1 号, 35–62 頁, 8 月, DOI： 10.5691/jjb.40.35.
津川友介 (2014)「疫学の「因果関係ダイアグラム（Causal Diagram）」」, 12 月, retrieved from *here*.
——— (2015)「差分の差分分析（Difference-in-Differences Design）」, 7 月, retrieved from *here*.
中村知繁 (2019)「統計的因果推論とデータ解析 ~ 適切な運用を目指して ~」, 5 月, retrieved from *here*.
林岳彦 (2011)「確率と因果を革命的に架橋する:Judea Pearl の do 演算子 - Take a Risk: 林岳彦の研究メモ」, 12 月, retrieved from *here*.
——— (2017)「バックドア基準の入門」, retrieved from *here*.
——— (2019)「環境分野における”EBPM” の可能性と危うさ: 他山の石として」, retrieved from *here*.
保木邦仁・渡辺明 (2007)『ボナンザ vs 勝負脳: 最強将棋ソフトは人間を超えるか』, KADOKAWA, 東京; 東京, OCLC: 676002553.
星野崇宏 (2009)『調査観察データの統計科学 –因果推論・選択バイアス・データ融合』, 岩波書店.
宮川雅巳 (2004)『統計的因果推論–回帰分析の新しい枠組み–』, シリーズ予測と発見の科学, 第 1 号, 朝倉書店, Tokyo, retrieved from *here*.
森田果 (2014)『実証分析入門: データから「因果関係」を読み解く作法』, 日本評論社, 東京, OCLC: 881836881.
山口慎太郎 (2017)「動学的離散選択モデルの構造推定」,『第 20 回労働経済学カンファレンス』, 東京, 9 月, retrieved from *here*.

山名一史 (2017) 「「エビデンスに基づく政策形成」とは何か」，財務省広報誌「ファイナンス」 68, retrieved from *here*.
山本勲 (2019) 『人工知能と経済』，勁草書房，OCLC: 1122753468.
横尾真・岩崎敦・櫻井裕子・岡本吉央 (2012a)「『計算機科学者のためのゲーム理論入門』シリーズ第 1 回非協力ゲーム (基礎編)」,『コンピュータソフトウェア』, 第 29 巻, 第 2 号, 69–84 頁, DOI： 10.11309/jssst.29.2_69.
——— (2012b) 「『計算機科学者のためのゲーム理論入門』シリーズ第 2 回非協力ゲーム (発展編)」,『コンピュータソフトウェア』，第 29 巻，第 3 号，3_39–3_53 頁，DOI： 10.11309/jssst.29.3_39.
渡辺澄夫 (2012) 『ベイズ統計の理論と方法』，コロナ社.