PRESS RELEASE（2022/12/05）

計測データの量や質に対するベイズ推定のスケーリング則を解明

~複雑現象の計測と数理モデリングをつなぐ新たな指針に~

ポイント

① 複雑現象の理解には所与の計測データを過不足なく表す関数や方程式（数理モデル）が有用

② データの質や量に応じて最良の数理モデルを選択するベイズ推定のスケーリング則を解明

③ データに根ざした数理モデルの簡略化や複雑現象の計測を効率化する指針につながると期待

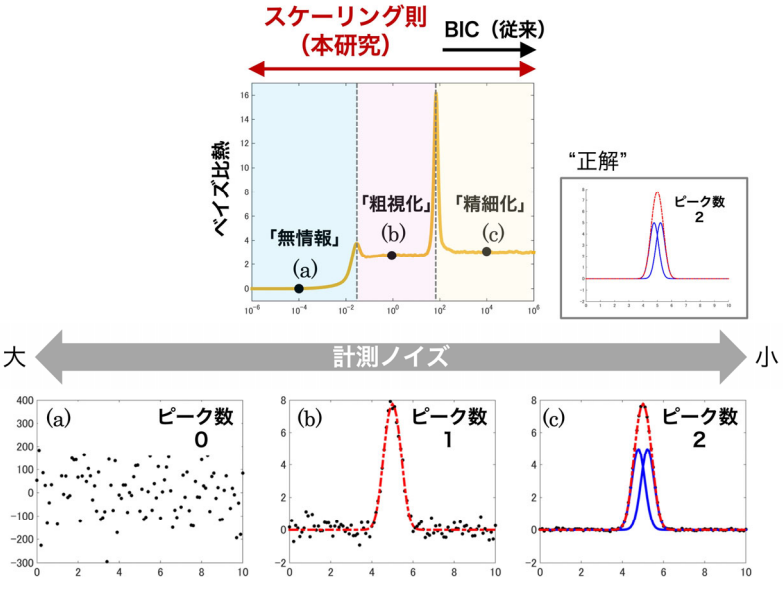
概要

古くは惑星の運動を司るケプラーの法則が象徴するように、単純な関数や方程式を用いて計測データを表す数理モデリングは様々な現象に対する理解を深めてきました。ベイズ情報量規準（BIC）は所与のデータを過不足なく単純に表す数理モデルを選ぶための指標であり、近年のデータ駆動科学を支える標準的なツールの一つです。IT 分野などで幅広く用いられているベイズ推定[※1]を数学的に近似した統計学の公式として、BIC は導かれます。しかし、同近似はデータの量や質による影響を無視しており、本来それらがベイズ推定にどう影響するかは BIC の発見から 40 年以上に渡り未解決問題のままでした。

九州大学情報基盤研究開発センターの徳田悟助教、東京大学大学院新領域創成科学研究科の岡田真人教授らの共同研究グループはベイズ推定と統計物理学の数学的な対応に着目し、理論解析を進めることで、計測データの量や質に対するベイズ推定のスケーリング則[※2]を初めて明らかにしました。これを元にした数値シミュレーションを行うことで、ベイズ推定が計測データの質や量に応じた複数の「状態」を取り、状態毎に異なる数理モデルを最良とみなす性質を発見しました。データの量が多く質が高い状態であるほど、より多くのパラメータを持つ複雑な数理モデルを最良とみなすこともわかりました。これらは BIC では説明がつかず、今回発見したスケーリング則によって初めて明らかになった性質です。

今回の発見はこれまで研究者の洞察に頼ってきた数理モデルの簡略化を計測データに根ざして客観化・自動化することを可能にし、様々な複雑現象の実態を捉えるために役立つと期待されます。見方を変えれば、所与の数理モデルの妥当性を実証するために「どれくらいの量や質の計測データが必要か？」という問いに答えるものでもあり、計測の効率化の指針につながることも期待されます。

本研究成果は米国物理学会発行の学術誌「Physical Review Research」に米国東部時間 2022 年 12 月 6 日（火）に掲載されました。



計測データの質（計測ノイズの大きさ）に応じたベイズ推定の三態

ベイズ比熱という量を新たに定義し、計測ノイズの大きさに対するスケーリング則を導きました。これを元に、2 つのピーク関数（青線）の重ね合わせ（赤破線）を”正解”とする計測データ（a-c の黒点）を想定した検証を行い、計測ノイズの大きさに応じた 3 つの「状態」（a-c）がベイズ推定にあることを突き止めました。各状態では異なるピーク数の関数が最良とみなされました。これらは正解が当てられる状況を仮定した BIC では説明がつかない結果です。

【研究の背景と経緯】

関数や方程式を用いてデータを表現する数理モデリングは多くの科学技術分野で重要な役割を果たしています。現在の人工知能の中心を担う深層学習も端的に言えば、学習データに合わせて調整された合成関数の一種です。深層学習は優れた予測性やドメインを問わない万能性から重宝される一方、膨大な数のパラメータを持つ複雑な関数であることから、その解釈や理解が困難であることが問題視されています。他方、物理学などの分野ではパラメータの数が少ない単純な数理モデルと計測データを照らし合わせることで、様々な現象に対する理解を深めてきました。このことは惑星の運動を司るケプラーの法則によって象徴されます（図 1)。「所与のデータを表すために必要以上に複雑な数理モデルを用いるべきではない」とする指針はオッカムの剃刀と呼ばれ、情報量規準や疎性モデリングなどの統計学的方法にも反映されています。研究者の洞察に頼ってきた伝統的な数理モデリングをこうした統計学的方法によって客観化・自動化する試みは近年のデータ駆動科学における一つの潮流です。

ベイズ情報量規準（BIC）はオッカムの剃刀を具体化した統計指標であり、データ駆動科学を支える標準的なツールの一つです。パラメータの数が多い複雑な数理モデルほど、データへの当てはまりが良いというトレードオフを定量化するものであり、両者のバランスが取れた最良の数理モデルを選ぶことを可能にします（図 2)。BIC は統計学を出自とする公式であり、IT 分野などでも幅広く用いられているベイズ推定を近似することで導かれます。しかし、この近似は所与のデータの量が十分に大きい極限を仮定しており、計測データの量や質の有限性がベイズ推定にどう影響するかは 1978 年に BIC が発見されて以来、長らく未解明のままでした。

【研究の内容と成果】

今回、九州大学、東京大学らの共同研究グループは上記の問題を解決するため、多粒子系の統計性を記述する統計物理学とベイズ推定の数学的な対応に着目しました。統計物理学における有限サイズスケーリングや自己平均性といった概念を導入し、理論解析を進めることで、計測データの量や質に対するベイズ推定のスケーリング則を導くことに成功しました。研究グループは理論解析を適用する問題設定として、条件付き独立な観測に基づく非線形回帰を選びました。スペクトルや画像、時系列のようなデータセットの多くは等間隔に配置された計測点毎に独立に計測されたものとみなせ、それらに数理モデルを当てはめる非線形回帰も典型的なアプローチです。研究グループは計測点の個数をデータの量、各データに加わる計測ノイズの大きさをデータの質として定量化し、両者の積として「観測の精細度」という量を新たに定義しました。加えて、ベイズ推定に内在し、オッカムの剃刀を実現する正則化効果に注目し、その定量化に対応する「ベイズ比熱」という統計量を新たに定義しました。理論解析を進めることで、観測の精細度に対するベイズ比熱のスケーリング則を導きました。

さらに、これを元にした数値シミュレーションを行い、「相転移」とも呼ぶべき変化の存在を発見しました。元々、相転移とは水が氷に凝固するような物質の状態変化を指します。研究グループはベイズ推定にも複数の「状態」があり、状態毎に異なる数理モデルを最良とみなすことを突き止めました。特に、観測の精細度が高い（データの量が大きいまたは質が高い）状態であるほど、より多くのパラメータを有する複雑な数理モデルを最良なものとみなすことがわかりました。このような相転移は BIC を初めとした従来の理論では説明がつかず、今回発見したスケーリング則によって初めて明らかになった性質です。観測の精細度が着目する現象のスケールに対応するものとみなせば、今回の発見はこれまで研究者の洞察に頼ってきた数理モデルの簡略化を、計測データに根ざして客観化・自動化できる可能性を示しており、データ駆動科学を推進する新たな方向性を示すものです。

【今後の展開】

今回、研究グループは数理モデルの簡略化を計測データに根ざして客観化・自動化するための理論的な裏付けを与えました。今後は実際の計測データを元に、理論の実証を進めていきます。今回の発見は様々な複雑現象の実態を捉えるために役立つものであり、物性物理学、化学、生命科学や地球科学など、幅広い分野への波及効果が期待されます。

一方、見方を変えれば、今回の発見は所与の数理モデルの妥当性を実証するために「どれくらいの量や質の計測データが必要か？」という問いに対して一つの答えを与えるものとも捉えられます。今後はマシンタイムが限られる大型実験施設における計測の効率化や計測上の原理的限界から S/N 比が制限される時間分解計測の最適設計などへの指針となることが期待されます。

【参考図】

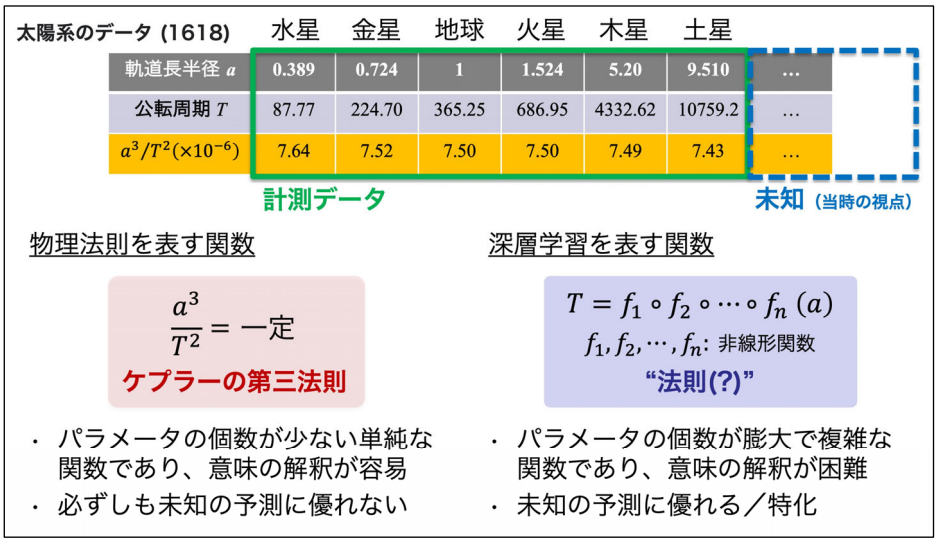


図１　数理モデルとしての物理法則と深層学習の比較

上段の表は太陽系惑星の軌道長半径と公転周期を記録したデータ（1618 年）。当時、ケプラーはこのデータを後にケプラーの第三法則（左下）と呼ばれる単純な関数で表した。現代的には同じデータを深層学習（右下）のような複雑な関数で表すこともできるが、これを”法則”と呼ぶかは議論が分かれるであろう。この例は未知の現象に対する優れた予測性が必ずしもその現象の理解に直結しないことを示唆している。

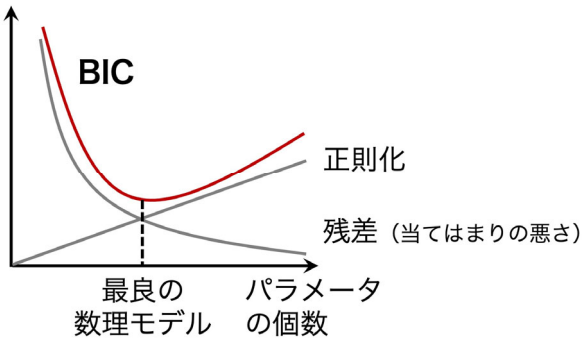


図２　ベイズ情報量規準（BIC）の仕組み

BIC は数理モデルと計測データの組に対して定義され、パラメータの個数とデータへの当てはまりの度合いの間に成り立つトレードオフを定量化するものである。候補となるいくつかの数理モデルの中で BIC が最小となるものが、最良の数理モデルとみなされる。

【用語解説】

(※1) ベイズ推定

パラメータ推定は計測データを表す数理モデルを立て、モデルのパラメータの値をデータに合うように求めるデータ分析の一つである。特に、計測データとパラメータが共にランダムに値が決まるもの（確率変数）とみなし、計測データが与えられた下でパラメータが従う条件付き確率分布を求める手続きをベイズ推定と呼ぶ。パラメータの値だけでなく、その値の不確かさを定量化できることが一つの特徴である。ベイズ推定は条件付き確率の連鎖律（ベイズの定理）をその基礎とし、数理モデルの不確かさも定量化できる。ベイズ情報量規準（BIC）は計測データが与えられた下で数理モデルが従う条件付き確率分布（モデルの事後分布）を近似することで導出される。今回、この近似で無視される計測データの量や質に応じたモデルの事後分布の変化を明らかにした。

(※2) スケーリング則

2 つ以上の興味のある量の間に成立する変換則。例えば、球の半径rと体積Vに着目すると、その間には$V = 4\pi r^3/3$という関係が成立する。つまり、これはVがr^3に比例するという変換則である。このことから、rを 2 倍するとVは 2^3=8 倍になることがわかる。今回、新たに定義した「ベイズ比熱」という量Cと計測データの量（計測点の個数）n、計測データの質（計測ノイズの小ささ）βの間に$C = f(n\beta)$という関係が成立することを、それを満たす関数fの詳細と共に明らかにした。

【謝辞】

本研究は JSPS 科研費 （JP20K19889, JP25120009）、JST-CREST（JPMJCR1761）などの助成を受けたものです。

【論文情報】

掲載誌：Physical Review Research

タイトル：Intrinsic regularization effect in Bayesian nonlinear regression scaled by observed data

著者名：Satoru Tokuda, Kenji Nagata, and Masato Okada

DOI：10.1103/PhysRevResearch.4.043165

【お問合せ先】

＜研究に関すること＞

九州大学情報基盤研究開発センター　助教　徳田 悟（トクダ サトル）

TEL：092-583-8423

Mail：s.tokuda.a96@m.kyushu-u.ac.jp

＜報道に関すること＞

九州大学広報室

TEL：092-802-2130　FAX：092-802-2139

Mail：koho@jimu.kyushu-u.ac.jp

東京大学大学院新領域創成科学研究科広報室

TEL：04-7136-5450

Mail：press@k.u-tokyo.ac.jp