

ベイズ推定の数理：計測データの量や質に対するスケーリング則

Scaling laws for Bayesian inference: Dependence on quantity and quality of observed data

徳田 悟

Satoru Tokuda

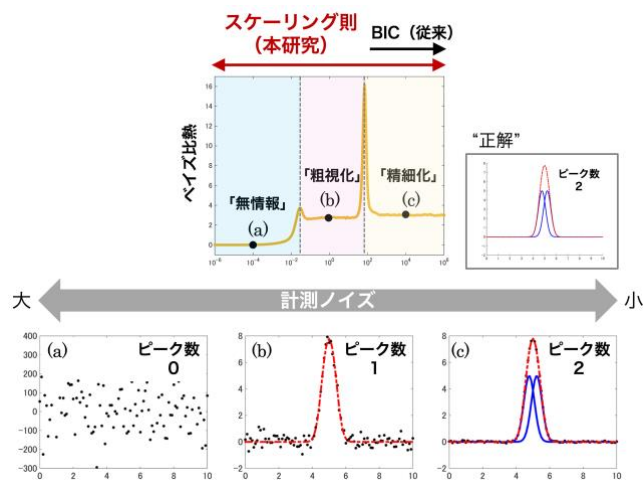
九大情基研セ

RIIT, Kyushu Univ.

古くは17世紀に発見された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た。ベイズ情報量規準 (BIC) はなるべく少ないパラメータで計測データを過不足なく表す単純な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ある。BICはベイズ推定における統計量の一つである「自由エネルギー」の近似した公式として導かれる[1]。ベイズ推定によるモデル選択はIT分野を始め、近年ではプラズマ物理にも活用されている[2]。しかし、BICの近似で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よるモデル選択に対してどう影響するかはBICの発見から40年以上も未解明だった。

我々はベイズ推定と統計物理の数理的対応に着目した理論解析を行い、計測データの量や質に依存するベイズ推定のスケーリング則を導出した[3]。さらに、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、状態によって異なるモデルを最適とみなす性質を発見した。データの量が多い／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モデルを最適とみなすこともわかった。これらの挙動はBICでは説明がつかず、導出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ある。

これらの知見は研究者の洞察に頼ってきたモデルの単純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、様々な現象の理解に役立つことが期待される。また、モデルの妥当性を実証するために「どの程度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る。



図：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依存するスケーリング則を導出した。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た。各状態で異なるピーク数の関数が最適とみなされた。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ある。

1. G. Schwarz, “Estimating the dimension of a model”, *Ann. Statist.* **6**(2), 461-464 (1978).
2. S. Tokuda, et al., “Bayesian inference of ion velocity distribution function from laser-induced fluorescence spectra”, *Scientific Reports* **11**, 20810 (2021).
3. S. Tokuda, et al., “Intrinsic regularization effect in Bayesian nonlinear regression scaled by observed data”, *Phys. Rev. Research* **4**, 043165 (2022).