実験データからピークの数を推定するには？ ―スペクトル分解とベイズ統計―

永田賢二 〈東京大学大学院新領域創成科学研究科 〉
杉田精司 〈東京大学大学院理学系研究科 〉
佐々木岳彦 〈東京大学大学院新領域創成科学研究科 〉
岡田真人 〈東京大学大学院新領域創成科学研究科 〉

あらゆる物理学の分野において，実験データから必要な情報を抜き出す作業は日常的に行われることである．特にデータの中から複数のピークを探し出し，その位置や広がりを評価することは，実に多くの場面で重要となる．実験データからピーク位置の情報をフィッティングなどで取り出すこと自体は，グラフソフトなどを使えばそれほど難しいことではない．ところが「いったい何個のピークがあるのか」ということを判断することは難しい．ほとんどの場合，何個のピークがあるかを判断するのは解析者の直感に委ねられる．しかし，時に何個のピークがあるか迷うデータに遭遇することもあるだろう．例えば，右下の図は複数のガウス関数の和にノイズを加えて生成した，人工的な実験データである．果たして何個のピーク（ガウス関数）があるのか，判断できるであろうか．

データのみからピークの個数を決定することは，理論的にも難しい問題である．例えば，データとフィッティング関数の差（誤差関数）を最小化してピークの個数を決定しようとすると，ピークの数を増やすことでいくらでも誤差を下げることができてしまう．このようなノイズまでフィットしてしまう「オーバーフィッティング」の問題を避けるためには，誤差関数だけでなく，モデルの複雑さとのトレードオフを兼ね備えた関数を考える必要がある．また同様の問題として，実験データを多項式でフィットする問題を挙げることができる．n 点のデータに対して，$n-1$ 次の多項式でフィットさせると，誤差なくすべてのデータをフィットさせることができるが，意味のないデータ解析であることは明らかであろう．このような，ピークの個数の決定や多項式の次数の決定の問題は，統計学の分野において「モデル選択」と呼ばれている．

モデル選択の問題に対しては，赤池情報量規準やベイズ情報量規準といった情報科学の分野で開発されたモデル選択規準が広く使われており，多項式フィッティングの問題をはじめとして，様々なモデル選択で一定の成功を収めている．しかし，ピーク個数の決定については，モデルに内在する数理的な構造の複雑さにより，これらのモデル選択基準の適用により決定することが困難である．最近になって，ベイズ推定とモンテカルロ法を組み合わせた新しい手法が開発され，ピーク個数の決定に応用されるようになった．この手法は，ベイズ推定で記述される評価関数に現れる量を「分配関数」「自由エネルギー」などに読みかえることで，モンテカルロ法を適用するといった特徴を持っている．

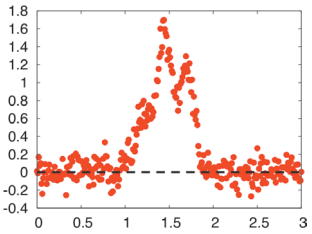
本稿では，なるべく専門性の高い内容は避け，ベイズ推定によるモデル選択の枠組みを概説し，実際に「右図は3つのピークが合成されている」と考えるのが最も自然であることを示す．

—Keywords—

ベイズ統計：
確率推論の枠組みの一つ．ベイズ統計による推定では，ベイズの定理（$P(B|A)=P(A|B)P(B)/P(A)$）を利用し，事象 A を観測したときにその要因となる事象 B の確率を推測する．$P(B)$ は事前確率，$P(B|A)$ は事後確率と呼ばれる．

オーバーフィッティング：
統計学の分野において，データを説明するモデルが不適切にもかかわらず，与えられたデータをよく再現してしまう現象．この現象が起きると，他のデータに対するモデルの説明能力が失われる．

赤池情報量規準：
統計モデルの選択に利用される，モデルの良さを評価する指標の一つ．Akaike's Information Criterion（AIC）とも呼ばれる．データへの適合度だけでなくモデルの複雑さも考慮に入れた指標であり，1973年に統計数理研究所に所属していた赤池弘次が発表して以降，広く使われている．



人工的なデータ例．何個のピークがあるでしょうか？