2024 年 8 月 28 日(水)

計測解析双方向相互作用

東京大学　大学院新領域創成科学研究科　複雑理工学専攻

岡田真人

1. はじめに

ポストプロセスとしてのデータ解析の限界

従来の実験/計測のデータ解析の手順は，まず験/計測をおこない，そこで得たデータをデータ解析する．図 1 の上図にあるように，データの情報処理は実験/計測からデータ解析への一方向的に流れる．また従来は，データ解析は，その実験/計測の研究者がある意味片手間で行うか，その分野の理論家が行うことが多い．そのため，ここの分野のデータ解析は，他の分野のデータ解析で発見された知見などを利用することはほとんどない．つまり，分野ごとに分離された縦割り構造なのである．これらの理由で従来のデータ解析は，研究者にとって副次的なことがあることが多く，データ解析の学理を構築すべきという機運は稀有であった．

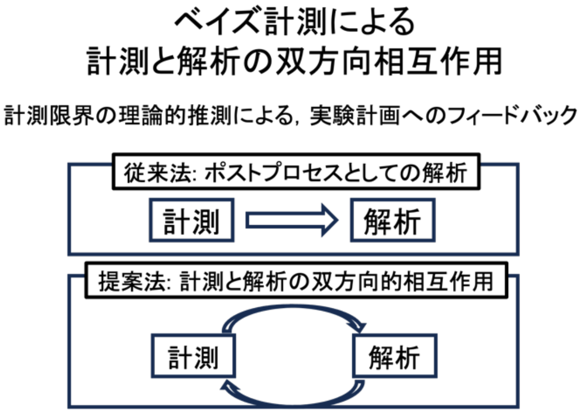

図 1 計測と解析の双方向相互作用

階層的自然観に基づけば，データの生成を記述するための数理モデルは，ミクロなレベルからの理論的演繹だけでは決定することができず，ミクロからの演繹ができない故におこる複数モデルを，データだけから決定する理論的枠組みが必要となる．

この複数モデルの決定プロセスは，すべての分野の共通課題であるのにも関わらず，何らかの方法で複数モデルを決定する理論的枠組みを構築しようとする機運は生じない．これが，ポストプロセスとしてのデータ解析の限界である．

2. ベイズ計測による計測限界と計測解析双方向相互作用

最小二乗法などによるパラメータの点推定に対して，ベイズ計測ではパラメータの事後確率分布を求めることができる．　パラメータの事後確率分布を求めることができると，従来法では手が届かなかった，対観測ノイズなどの計測限界を

理論的に取り扱うことができる．この計測限界の理論は図 1 の上図に示すように，データ解析を実験/計測の単なるポストプロセスと捉える一方向的な情報の流れではなく，下図のようなデータ解析から実験/計測へのフィードバックをかけられる双方向的な情報の流れになる．この双方向的な情報の流れを，我々のグループは計測と解析の双方向相互作用と呼んでいる．この双方向的な枠組みでは，データ解析は実験計画に大きな影響を与える．つまり，実験家も積極的にベイズ計測を習得なければ，1.で述べたポストプロセスとしてのデータ解析の限界を越えることができずに，時代に取り残されてしまう．

2.1 ノイズ強度の増加によるパラメータの事後確率分布の定性的変化

Nagata らはベイズ的スペクトル分解[1]において，ノイズ強度を増加させると，パラメータの事後確率分布の形が定性的に変化することを数値実験で示した[2]．

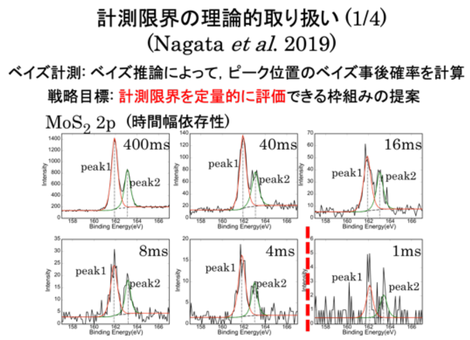

図 2: 計測時間幅を変化させた場合のスペクトル

Nagata らは MoS_2 の S の 2p 軌道の X 線光電子放出スペクトル(XPS: X-ray Photoelectron Spectroscopy)の光電子の計測時間を変化させて計測した．図 2 が，その一連の実験結果である．計測時間幅が 400msec の場合は，スペクトル分解するまでもなく，スペクトルは 2 ピーク構造である．図 2 からわかるように，計測時間幅を 40msec，16msec, ..., 1msec まで減らしていくと，スペクトルがギザギザしてくるがわかる．これは，計測時間幅が減少すると，その時間窓で検出される光電子の数が減り，光の量子性が顕在化し，ポアソンノイズが増えるからである

通常は，ポアソンノイズを小さくするために，計測時間幅を十分大きく取る．しかしながら，計測対象が有機物などで，X 線で破壊されてしまいやすい場合は，できるだけ計測時間を短くしたい．また，計測対象が X 線で破壊されなくても，系のダイナミクスを XPS で追いたい場合は，ダイナミクス見るために計測時間幅を小さくしたい．これらの場合，できれば実験の前の実験計画の段階で，2 ピーク構造が保てる範囲の中で，できるだけ計測時間幅を小さくしたい．しかし，そのような実験計画を得ることができる理論的枠組みをこれまでは持っていない．

その計測時間幅の実験計画に使えるのがベイズ計測のパラメータの事後確率分布推定である．図 3 は，計測時間幅を 400msec, 40msec，16msec, …, 1msec まで減らした場合の，基底関数の中心位置の事後確率分布である．今の場合，2 ピークスペクトルであるので，一つのスペクトルあたり，2 個の基底関数の中心位置のパラメータの事後確率分布が存在する．各計測時間幅の赤と青が 2 個の基底関数の中心の事後確率分布を対数で表している．計測時間幅が 400mse や 40msec の場合は，分布の形が 2 次間数になっており，基底関数の中心の事後確率分布は，ガウス分布で良く表せることがわかる．計測時間幅が 4msec になると，分布は 2 次関数より幅広になり，基底関数の中心の事後確率分布はガウス分布より，裾を引く形になっている．計測時間幅が 1msec の場合，二つの底関数の中心の事後確率分布は重なったおり，計測時間幅 400msec～4msec までとは，定性的に形が違っているのがわかる．これをより詳細にみたのが，計測時間幅 4msec と 1msec のパラメータの事後確率分布を示す図 4 である．図 4 から明らかなように，計測時間幅 4msec の場合は，赤と青の二つのパラメータの事後確率分布は重なってない．一方，計測時間幅 1msec の場合は，赤と青の二つのパラメータの事後確率分布が重なっている．図 5 は，計測時間幅を 400msec, 40msec，16msec, …, 1msec の場合の，基底関数の中心位置の事後確率分布の 2 次元ヒートマップである．図 5 から計測時間幅 4msec までは，2 次元事後確率分布は 1 ピーク構造を持つが，計測時間幅 1msec で突然，それまではことなる幅広い分布形状を持つようになる．

この計測時間幅依存性，つまりノイズの大きさ依存性を使えば，以下のような

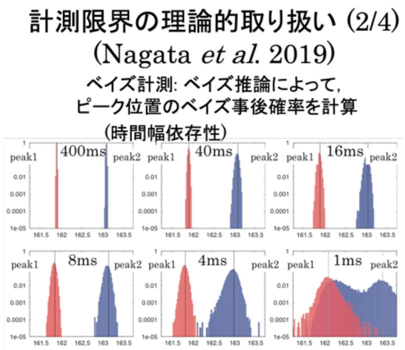

図 5: 計測時間幅を変化させた場合のパラメータの事後確率分布

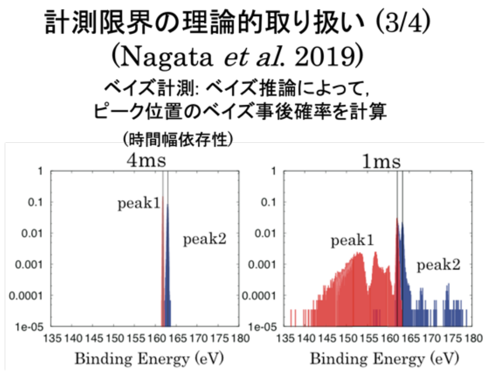

図 4 計測時間幅 4msec と 1msec のパラメータの事後確率分布

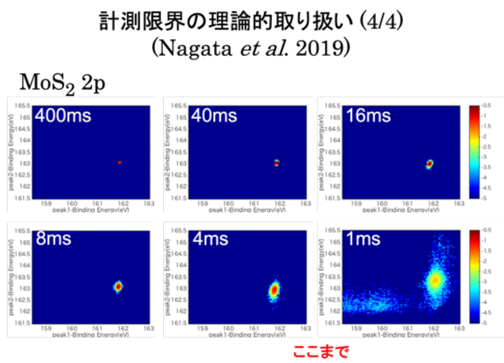

図 3 計測時間幅の変えた場合のパラメータの 2 次元事後確率分布のヒートマップ

手順で計測限界を実験する前にあらかじめ知ることができる．得られるデータはベイズ計測を使って解析するので，かならずデータ生成の順モデルがあるはずである．計算機を使って，その順モデルでデータを生成し，それにノイズを重畳する．そのノイズの大きさを何段かも用意して，それぞれのノイズに対して，ベイズ計測でパラメータの事後確率分布を求める．ここでは我々がパラメータの真値を知っている人工データを用いるので，重畳するノイズが小さければ，事後確率分布の中心は真値周りで，事後確率分布の幅も狭いはずである．そこで重畳するノイズを少しずつ大きくしていくと，事後分布の幅はそれに応じて広くなっていくはずである．それを繰り返していくと，あるノイズの値で，突然，事後確率分布の形状が定性的に変化すれば，そことそのひとつ前のノイズ値の間に，計測限界が存在する．

繰り返しになるが，このようなことは実験をする前に行えるので，事前に計測限界を求めておいて，その限界内で実験を行えば良い．これが図 1 の計測と解析の双方向相互作用を用いた枠組みの具体例です．

2.2 パラメータ事後確率に関する計測限界のメカニズム

2.1 で述べた計測限界が生じるメカニズムを議論する．まず，解析的にパラメータの事後確率分布を求めることができる,線形回帰モデル $y=ax+b$ を考える．このモデルでは，解析計算により，パラメータ a と b の事後確率分布はガウス分布に従い，そのガウス分布の分散は観測ガウス分布の分散を使って表現できる[3]．したがって，観測ガウス分布の分散を大きくしても，パラメータ a と b の事後確率分布はガウス分布のままです．つまり，線形回帰モデル $y=ax+b$ では，計測限界は存在しません．

この考察から，計測限界が存在するには，誤差関数が多峰性を持っていることが必要であることが定性的に理解できる．観測ノイズが小さい時は、誤差関数の大局的最小点の周りを REMC がサンプリングするので、パラメータの事後確立分布は近似的にはガウス分布であろう．観測ノイズが大きくなると、その大局的最小点から REMC のサンプリングが飛び出すことで，パラメータの事後確立分布の形が定性的に変化すると定性的に理解できる．

<u>計測限界とクロスオーバー(相転移)</u>

図 3 や図 4 のパラメータの事後確率分布の定性的な変化を観察すると，計測時間幅の少しの変化が，とても大きな変化を生むことから，この計測限界の現象が統計力学における相転移と関係あるかもしれないという仮説を思いつく．

これを理論的に取り扱ったのが Tokuda らである[4]．Tokuda らは，スペクトル分解をとりあげ，2 ピークで，二つの基底関数の中心が少しだけはなれた真のモデルを仮定した．この真のモデルに重畳する観測ガウスノイズの分散が大きい時には，中心位置の事後確率分布推定に失敗する．そこから，観測ガウスノイ

ズの分散を小さくしていくと，1 ピーク構造がモデル選択される．さらに，観測ガウスノイズの分散を小さくしていくと，2 ピーク構造を正しくモデル選択できる．つまり，この最後の段階が，正しく推定できる計測限界を表す．

Tokuda らは，統計物理学からのアナロジーも援用しながらベイズ比熱を定義した．観測ガウスノイズの分散が大きいところから減らしていくと、まず 1 ピークとモデル選択されるところで，ベイズ比熱がピークをとることがわかった．さらに，観測ガウスノイズの分散を減らしていくと，ただしく 2 ピーク構造とモデル選択されるところで，ベイズ比熱がピークをとることがわかった．Tokuda らは，これを漸近論で理論的に取り扱うことに成功した[4]．

2.3 ノイズ強度の増加によるモデル選択の計測限界

Nagata らはさらに，スペクトル分解におけるピーク数のモデル選択の対ノイズ性を議論している[2]．Nagata らは，パラメータの事後確率分布の計測限界を超えている場合でも，ピーク数は正しくモデル選択されることを示している．つまり，この事例だけを考えると，事後確率分布推定とモデル選択の計測限界は異なり，モデル選択の計測限界が高いことがわかる．これについては，今後さらなる検証が必要である．

3　まとめ

パラメータの事後確率分布推定が，対観測ノイズ性などの計測限界を理論的に取り扱うことにつながることを示した．計測限界の理論ができると，図 1 の上図に示すように，データ解析を実験/計測の単なるポストプロセスと捉える一方向的な情報の流れではなく，下図のようなデータ解析から実験/計測へのフィードバックをかけられる双方向的な計測と解析の双方向相互作用が生じる．この双方向的な枠組みでは，データ解析は実験計画に大きな影響を与える．

このようにベイズ計測は，単にデータ解析の性能を上げるだけでなく，ここで紹介したような計測と解析の双方向相互作用による実験計画へのデータ解析からのフィードバックのように，研究のやり方自体を，これまでの旧態依然としたものからモダンで効率的な枠組みに刷新できるパラダイム創成器であると考えられる．

参考文献

[1] Nagata, Sugita and Okada, “Bayesian spectral deconvolution with the exchange Monte Carlo method”, *Neural Networks*, 28, 82-89 (2012)

[2] Nagata, Muraoka, Mototake, Sasaki and Masato Okada, “Bayesian Spectral Deconvolution Based on Poisson Distribution: Bayesian Measurement and Virtual Measurement Analytics (VMA)”, *Journal of the Physical Society of Japan*, 88(4) 044003 - 044003 (2019)

[3] Katakami, Kashiwamura, Nagata, Mizumaki and Okada, “Mesoscopic Bayesian Inference by Solvable Models”
https://arxiv.org/abs/2406.02869

[4] Tokuda, Nagata and Okada, “Intrinsic regularization effect in Bayesian nonlinear regression scaled by observed data”, *Phys. Rev. Research*, 4, 043165 (2022)