研 究 終 了 報 告 書

「再標本化による情報計測のためのデータ駆動診断法開発」

研究期間： 2017年10月～2021年3月
研 究 者： 中西　義典

1. 研究のねらい

本研究のねらいは情報計測の限界を見極め、その成否を診断するためのデータ駆動診断法を開発することである。情報計測の研究が一筋縄でいかない理由は、高度な計測・解析技術に最先端の情報科学・統計数理手法が組み合わさるとき、できることとできないこととの境界が曖昧であり、このことによりさまざまな弊害が生じるためである。情報計測の限界を過小評価した場合は、目先の課題解決に囚われて、新しい方法の導入や異なる対象への応用による更なる発展の可能性が見過ごされてしまう。その一方で、情報計測の限界を過大評価した場合は、不十分なデータから得られる過誤を見落とし議論が迷走してしまう。さらに悪いことには、個々の情報計測の研究がそのどちらの状況に陥っているのかの判断が難しいことにより議論が衝突することもある。本研究で開発するデータ駆動診断法により無理なことは無理であると適切に主張するための道具立てを与えることを目指す。こうした主張は一見悲観的にも思えるが、長期的視野で見れば情報計測の進化を加速させるために極めて重要である。情報計測の最前線においてでさえ依然として計測不可能な対象が明らかになることにより、将来の実験計画を洗練させることができるようになると考えられる。

2. 研究成果

（1）概要

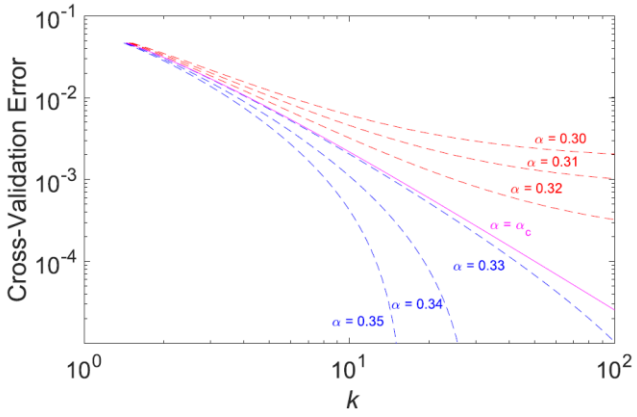
データ駆動診断法を開発するために、計測・解析課題に対して情報科学・統計数理手法を適用するという情報計測の枠組みについて、その成否を診断する手法やその限界を定量化する手法を提案し、提案手法の性能を評価した。代表的な研究成果の一つは、圧縮センシングと呼ばれる情報計測の枠組みに対して、データ駆動的にその成否を診断する手法を開発したことである。圧縮センシングとは、スパースモデリングに基づいて推定に必要なデータ数を削減する枠組みである。スパースモデリングとは計測対象にスパース性が内在することを想定することをいう。計測対象のスパース性が自明に確認される場合は少数のデータからでも計測対象の信号復元に成功することが知られているが、現実的な計測・解析課題ではそのような場合は稀である。そこで、データさえあれば実行可能な交差検証を応用することにより、計測対象のスパース性に関する事前知識を用いることなく信号復元の成否を診断する手法を開発した。また、計測対象に顕著なスパース性が確認できる場合に、簡便に推定アルゴリズムのハイパーパラメータを決定する手法の開発という副次的な成果が得られた。もう一つの代表的な研究成果は、結晶構造解析にベイズ推論を適用するという情報計測の枠組みに対して、計測データに基づいて構造モデルを最適化するとともに、最適な構造モデルの不確かさを定量化する方法を開発した。結晶構造解析とは、試料となる物質に量子ビームを照射しその散乱・回折パターンから物質の構造モデルを推定することをいう。ベイズ推論では、計測装置の性質や計測対象の事前知識を活用して、構造モデルを記述する温度因子や原子座標等

のパラメータの事後分布を数学的に計算する。これにより、最適な構造モデルを求めるだけでなく、その不確かさを定量的に評価する方法を開発した。

（2）詳細

研究テーマ A「交差検証による圧縮センシングに対するデータ駆動診断法の開発」

スパースモデリングに基づいて計測コストを低減する枠組みである圧縮センシングにおいて、どの程度取得するデータ数を少なくしてもよいかということは重要な問題である。しかしながら、計測対象やそのスパース性が未知であるとき、既に取得したデータ数が十分であるかを客観的に判断することは非常に難しい。そこで再標本化の一種である交差検証に着目した。交差検証とは、全データを訓練用と検証用とに分け訓練データのみを用いて推定した結果と検証データとの隔たりを交差検証誤差として定量化する手法である。本研究では、圧縮センシングの推定アルゴリズムとして L1 ノルム最小化に基づく基底追跡法を用いる場合について交差検証誤差を解析した。統計力学のレプリカ法により、全データ数／検証データ数（＝フォールド数 k）に対する交差検証誤差の依存性を調べた結果、データ数が圧縮センシングの成否を分ける転移点上にあるとき、交差検証誤差が漸近的に冪的に振舞うことが分かった（図）。このことから、交差検証誤差がフォールド数に関して指数関数的に減少するときは圧縮センシングが成功し、有限の値が残留するときは圧縮センシングが失敗することが分かった。このように交差検証のフォールド数依存性に着目する提案手法はデータさえあれば実行可能な手法でありデータ駆動的であるといえる。また解析で示されたシステムサイズ無限大の極限においてだけでなく、現実的なシステムサイズ有限の場合にも提案手法が有効に働くことを数値実験により示した。さらに、交差検証は様々な推定アルゴリズムと組み合わせて用いることができることから、基底追跡法に代表される凸緩和法だけでなく、貪欲法や閾値法を推定アルゴリズムとして用いる場合も提案手法が有効に働くことを数値実験により示した。[主な研究成果リスト 1、Nakanishi-Ohno and Hukushima, *Journal of Physics: Conference Series*, 2018, **1036**, 012014, doi: 10.1088/1742-6596/1036/1/012014]



さきがけ研究領域内の連携により得られた、本研究テーマに関する副次的な結果として、計測対象に顕著なスパース性が確認できる場合に適用可能な Least Absolute Shrinkage and Selection Operator (LASSO)のハイパーパラメータ決定法を開発した。L1 ノルム正則化に基づくLASSO は正則化パラメータというハイパーパラメータをもち、推定アルゴリズムとして用いる際にその値を決定する必要がある。前述のレプリカ法による交差検証誤差の解析を進める過程で、交差検証誤差を最小化するハイパーパラメータの値 λ_{CVE} と推定誤差を最小化するハイパーパラメータの値 λ_{MSE} とを比較した。結果として、計測ノイズの大きさによらず、比 λ

$\lambda_{CVE}/\lambda_{MSE}$ が一定になることが分かった。推定誤差を最小化するためには、交差検証誤差を最小化する値を補正する必要があるが、計測対象がスパース極限にあるときは解析により求めた一定値を乗じればよいことが分かった。数値実験により、この提案手法の性能が、従来手法である one-standard-error 則の性能を上回ることを示した。[主な研究成果リスト 3]

研究テーマ B「ベイズ推論による結晶構造解析に対する不確かさ評価法の開発」

情報計測の限界を見極めるために測定結果の不確かさを定量的に評価することは重要である。結晶構造解析では、物質中の原子座標や占有率、温度因子などの構造パラメータを推定する必要があり、限られた計測時間によるノイズの影響などによりパラメータ値を確定するだけのデータが得られない場合も少なくない。そこで本研究ではベイズ推論を活用することにより、不確かさの情報を付けてパラメータを推定する手法を開発し、本手法をX線CTR散乱データに基づくペロブスカイト酸化物の界面構造解析に適用した。ベイズ推論では、計測原理である回折理論を確率的に定式化し、ベイズの定理を用いて導出した事後分布を数値的に調べることになる。数値解析の観点から構造解析は非線形の最適化問題を内包し、その目的関数は局所最適解を多数もつ。また、特に構造解析の対象が界面であるときは、バルクのもつ並進対称性が破れているため構造パラメータの数が非常に多い。本研究で取り扱ったペロブスカイト酸化物の場合、パラメータの数は 60 余りであった。本研究では非線形かつ多変数の問題を取り扱うためマルコフ連鎖モンテカルロ法を用いてベイズ推論を実行した。マルコフ連鎖モンテカルロ法を用いると、つり合い条件を満たすように構成された遷移確率により、任意の確率分布へ収束する乱数を生成することができる。これによりベイズ事後分布からのサンプリングを可能にし、得られたサンプル列の平均値や分散を用いてパラメータの値やその不確かさを評価する。結果として、解析が難しい軽元素である酸素の振る舞いに関する情報を抽出し、ペロブスカイト酸化物薄膜に生じる金属絶縁体転移の機構を明らかにした。最近の進展として、高度なマルコフ連鎖モンテカルロ法の一種であるレプリカ交換法を実装した。これにより、よりロバストな構造解析が可能になり、多種多様な物質に対して適用できるようになると期待される。[主な研究成果リスト 2、Anada, …, Nakanishi-Ohno et al., *Physical Review B* 2018, **98**, 014105, doi: 10.1103/PhysRevB.98.014105]

さきがけ研究領域内の連携により、本研究テーマを応用して得られた成果として、予備的に得られた回折データを結晶構造解析の実験計画に役立てる手法を開発した。生物学や化学の分野においても与えられた化合物の分子構造を突き止める結晶構造解析に大きな関心がある。合成が難しい試料や不安定な試料を解析するときは、計測の無駄を省くために、早い段階で解析に適した結晶試料を選別したり、露光時間を決定したりする必要がある。予備的に得られた少数の回折データから計測対象の物性に影響を与える温度因子の値を推定するために、本研究では回折理論を平均化して得られるウィルソン統計を尤度関数として用いるベイズ推論を適用した。はじめに標準試料を用いて温度因子の推定が提案手法により正しく行われることを確認した。次に異なる溶媒分子を含んだ2種類の多孔性物質結晶に提案手法を適用したところ、温度因子の推定を介して溶媒分子の違いを判別することができた。これは、通常は数時間を要する計測を数分間の予備的な計測のみに置き換えることができたことを示す。[Hoshino, Nakanishi-Ohno et al., *Scientific Reports* 2019, **9**, 11886, doi: 10.1038/s41598-019-48362-3]

3.　今後の展開

本研究で開発してきたデータ駆動診断法を様々な情報計測の課題に応用展開していく。本研究ではスパースモデリングによる圧縮センシングやベイズ推論による不確かさ評価を軸に情報計測の課題を情報アプローチの観点から取り組んできた。スパースモデリングでは、今回対象としたアルゴリズムは、最も基本的な基底追跡法やLASSOであるが、それだけでなく他のアルゴリズムについてもデータ駆動診断を行えるようにすることが求められている。交差検証等の再標本化の技法を用いること自体は LASSO に限られたことではないため、幅広い応用が期待される。また、スパースモデリングは情報科学的な関心に留まらず、多くの計測データ解析に用いられる。データ駆動診断法を実際の計測の場に適用していくことも重要である。その際、再標本化の弱点である計算量の大きさを克服するため高性能計算等の計算科学との融合が必要になると考えられる。幸いなことに、交差検証はデータ分割を行った後はそれぞれの処理を独立に行えるので並列計算と親和性が高い。データ駆動診断を推進する目的は単に成功か失敗かを判断するのみならず、失敗と分かった後にその程度を定量化し、次のデータ獲得へと結びつけるための情報を得ることである。データ駆動診断が進展した後の長期的な展望として革新的な実験計画法の開発に繋ぐことを考えている。

4.　自己評価

本研究は情報計測領域における情報アプローチの研究として一定の成果を得た。特にスパースモデリングによる圧縮センシングに対してデータ駆動診断法を開発したことは理論的に重要な成果である。2000 年代半ばに行われた Candes や Donoho らの理論的研究に端を発した圧縮センシングは、これまで計測への応用やアルゴリズムの開発という方向性では数多くの進展が見られたが、その成否を客観的に判断するための方法論に乏しかった。本研究は圧縮センシングの分野にデータ駆動診断という新たな理論的研究の方向性を提示した点が画期的である。また、このことは戦略目標が掲げる「より少ないデータからの情報再構成技術」や「計測限界を定量的に評価できる枠組み」の構築に資するものである。データ駆動診断の研究は始まったばかりであり、これから発展する余地も広く残されている。まずデータ駆動診断の性能向上である。交差検証をはじめとする再標本化の性質を明らかにすることにより診断の精度を改善するだけでなく、高性能計算等の計算科学との融合により効率化を図ることが求められる。また、これらの技術を総合して実際の計測データ解析への適用を進めていく必要がある。

情報計測領域を通じて、様々な計測アプローチの研究者と連携・共同研究が進んだことは特筆すべきである。領域が開始する前から準備していた共同研究を継続して着実に進められたことにより得た成果もあるが、領域が開始した後に新たに着想した領域内共同研究も数多くあり、そのうち2件については査読有り英文学術雑誌に原著論文が掲載された。また、研究者個人としても、このように分野を跨いだ人脈を築けたことは大きな成果である。

5.　主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：10件

1. Yoshinori Nakanishi-Ohno and Koji Hukushima, Data-driven diagnosis for compressed

sensing with cross validation, *Physical Review E*, 2018, **98**, 052120, doi: 10.1103/PhysRevE.98.052120
交差検証を用いて圧縮センシングの成否を診断する手法を提案した。提案手法は、訓練データ数と検証データ数の比を変えた時の交差検証誤差の振舞いに着目する。レプリカ法により推定アルゴリズムとして基底追跡法を用いた時の交差検証誤差を解析し、圧縮センシングが失敗から成功へと丁度転移するとき交差検証誤差が漸近的に冪的に振舞うことを示した。合わせて、数値実験を行うことにより、提案手法が有効に働くことを示した。
2. Kazuki Nagai, Masato Anada, Yoshinori Nakanishi-Ohno, Masato Okada, and Yusuke Wakabayashi, Robust surface structure analysis with reliable uncertainty estimation using the exchange Monte Carlo method, *Journal of Applied Crystallography*, 2020, **53**, 387–392, doi: 10.1107/S1600576720001314
X 線 CTR 散乱法の計測データを用いて物質の界面構造を解析するために開発したベイズ推論を行うソフトウェアにレプリカ交換法を実装した。およそ 60 の構造パラメータをもつペロブスカイト型酸化物超薄膜を対象とした解析を行い、初期値に関するロバスト性が大幅に向上し、シンプルなモデルから解析を始めてもデータに即したモデルが得られた。また計測ノイズの相関を考慮に入れてパラメータの不確かさを評価できるようになった。
3. Yoshinori Nakanishi-Ohno and Yuichi Yamasaki, Multiplication method for fine-tuning regularization parameter of a sparse modeling technique tentatively optimized via cross validation, *Journal of the Physical Society of Japan*, 2020, **89**, 094804, doi: 10.7566/JPSJ.89.094804
スパース線形回帰の推定アルゴリズムである Least Absolute Shrinkage and Selection Operator の正則化パラメータを調節する方法を提案した。推定誤差を最小化するためには、交差検証誤差を最小化して得られる値をそのまま用いるのではなく値を補正する必要がある。レプリカ法により補正項を解析したところ、スパース極限においては計測ノイズの大きさによらない定数を掛ければよいことが分かった。また、数値実験により提案手法が従来法の one-standard-error 則の性能を上回ることを示した。

（2）特許出願

研究期間累積件数：0 件（特許公開前のものも含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1. 【招待講演】Data-driven diagnosis for compressed sensing with sparse modeling
   Yoshinori Nakanishi-Ohno
   Workshop ‘Development of next-generation quantum material research platform’ (Next QUMAT2017), Dec. 4, 2017, Tokyo, Japan.
2. 【招待講演】放射光 X 線回折とベイズモデリングの融合による表面・界面科学の新展開
   中西（大野）義典
   第 4 回放射光連携研究ワークショップ、2017 年 12 月 18 日、東京
3. 【招待講演】ベイズ推論が繋ぐ陽電子回折と表面構造解析
   中西（大野）義典

第 66 回応用物理学会春季学術講演会、2019 年 3 月 9 日、東京

4. 【解説記事】走査トンネル分光法の圧縮センシング―計測の効率に関する限界と可能性―
   中西（大野）義典、福島孝治
   固体物理、54(7)、343-351、2019 年 7 月 15 日
5. 【プレスリリース】熟練の研究者の「勘と経験」を誰でも簡単に再現～たった数分で単結晶構造解析の結果の事前評価が可能に～
   星野学、中西（大野）義典、橋爪大輔
   科学技術振興機構、理化学研究所、東京大学、2019 年 8 月 22 日