論 文 の 内 容 の 要 旨

論文題目　Data-driven approach to images in natural science
（自然科学画像に対するデータ駆動的アプローチ）

氏　名　中西（大野）　義典

自然科学の様々な分野で画像データが得られるようになり，中には天文学の画像解析手法が生命科学でも有効であるという事例が報告されている．天文学と生命科学とでは，対象が大きく異なるにもかかわらず，共通のデータ解析が行われるということは驚くべきことである．そこには，自然科学画像の解析には分野によらない普遍的な原理の存在が示唆される．本論文の目指す所は，多様な視点の導入により，その原理を探り，自然科学全体を刷新することである．

自然科学では，計測データから仮説を立て新たな実験により検証するということが行われてきた．この当然ともいえる行いが，皮肉なことに，技術の進歩によるデータの高次元化に伴い，研究者の人手では立ち行かなくなっている．一方で，計算機や情報科学の発展は目覚ましいものがある．自然科学と情報科学が密接に連携し，伝統的な仮説検証の営みを再興する枠組みをデータ駆動科学と呼ぶことにする．

自然科学者と情報科学者が連携しようとすると，しばしば次のような問題が生じる．情報科学の手法があまりにも多すぎて，自然科学者が適切な手法を選択できないという問題や，その逆に，自然科学の分野があまりにも細分化されすぎて，情報科学者が全ての専門知識を把握することはできないという問題である．こうした状況を打開するために，図１に示す，データ駆動科学の指導原理ともいうべき，データ駆動科学の三つのレベルを提唱する．

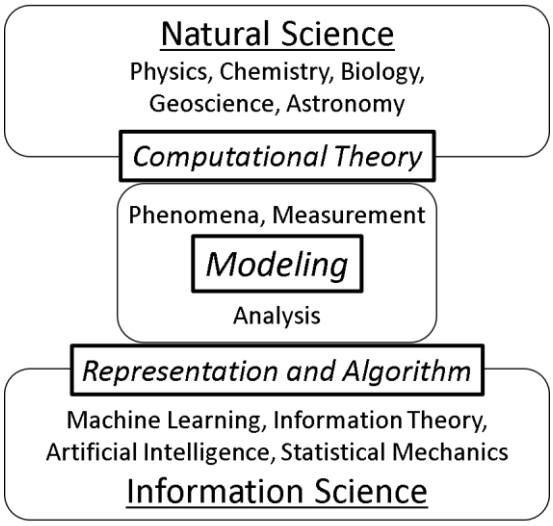

図１　データ駆動科学の三つのレベル．計算理論，モデリング，表現・アルゴリズム．

データ駆動科学の三つのレベルは，理論神経科学者の David Marr が提唱した，情報処理を司る機械を理解するために議論されるべき三つのレベルに影響を受けており，Marr の三つのレベルに含まれる計算理論のレベルと表現・アルゴリズムのレベルの間にモデリングのレベルが挿入されている．計算理論のレベルでは，自然科学者が，そのデータ解析でなすべきことを，先見知識に基づき議論する．表現・アルゴリズムのレベルでは，情報科学者が，数理的に困難な問題を，計算機を用いて効率的に解く手法を開発する．要となるモデリングのレベルでは，自然科学者

がデータ解析の目的や戦略を自身の専門用語に頼ることなく明言する．これは，情報科学者の理解だけでなく，他分野の自然科学者が抱える類似の問題に対する気付きを促す効果も合わせもつ．こうして浮かび上がる，自然科学のデータ解析における本質的な問題に関しては，万難を排してでも解く意義があるということを，自然科学者と情報科学者が共有することが重要である.そのようなモデリングに関する話題として，本論文はマルコフ確率場（MRF)，圧縮センシング（CS)，解空間解析の三つに焦点を絞って議論する．

自然科学で得られる画像データから，その系を説明するのに支配的な潜在構造を抽出することは重要である．とりわけ，画像データから，そのデータに記録された現象の拡散係数を推定することは本質的な問題である．画像処理でよく用いられる MRF が拡散方程式に対応することに着目し，筆者らが提案した画像データから拡散係数を推定する手法について説明する．

MRF とは，画像が滑らかであるという知識に基づいてモデリングする手法である．たとえば画像修復では，隣接する画素の値が離れている場合に罰則を与える正則化項を導入して，原画像を推定する．正則化項の係数はハイパーパラメータ（HP）と呼ばれ，画像修復の性能を左右するのだが，実は，この HP が拡散係数に対応する．ここで注意すべきは，HP を推定する目的は，情報科学と自然科学とで異なることである．情報科学では，データ処理の性能を向上させる HP を求めればよく，むしろ一定の値に決めることが後の処理を考えても都合が良い．一方で，自然科学では，拡散係数そのものが重要な量であり，その推定値は信頼度を含めて評価されるべきである．筆者らは自然科学の要請に対応するため，ベイズ推定の枠組みを用いて HP を推定する手法を提案した．

提案手法は，事後分布自体に着目するという点で独創的である．特に，拡散現象の理解に最低限必要なモデルに対して，図 2 左に示すように，解析的に HP の事後分布を計算し，その分布の広がりから推定の信頼度を評価できることを示した．また，図 2 右に示すように，変分ベイズ法により求めた近似分布を用いると，信頼度を過大評価することを示した．

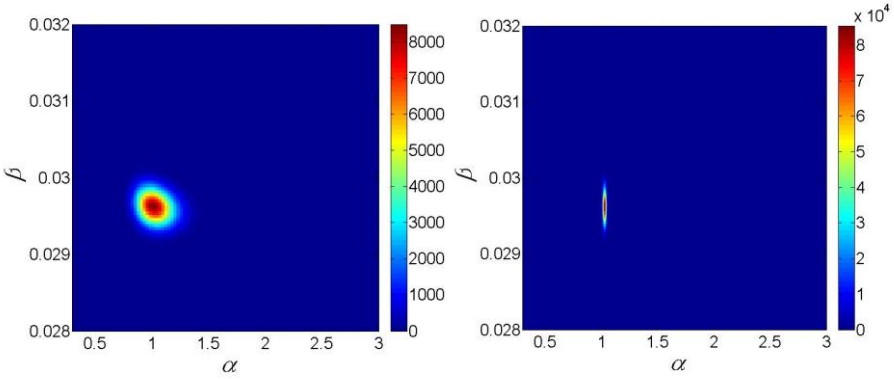

図 2　MRF の HP の事後分布．左が厳密解，右が変分ベイズ法により得られる近似分布．

CS の目的は，データ量を減らし測定時間を短くしながらも，原信号を得ることである．これを行うためには，未知変数の数よりデータ数が少ないという劣決定の問題を解く必要がある．ここでの戦略は，原信号のスパース性を仮定することである．原信号がスパースであるとは，その非ゼロ成分が少数であることをいう．実際に原信号がスパースであるとき，観測データを説明しうる多数の解候補から最もスパースなものを選ぶことにより，原信号を得ることができる．

CS を走査型トンネル顕微・分光法による準粒子干渉観測に適用する．準粒子干渉観測に CS が適用できるのは，干渉パターンがフーリエ空間上でスパースであると考えられるからである．この観測は，物性物理学で重要な役割を担うだけでなく，CS の視点からみても興味深い対象である．というのも，相補的な役割を担う実験手法が存在したり，実験系が制御しやすく，考案された実験計画に柔軟に対応できたりするからである．

具体的には Ag(111)表面の電子に特有なフーリエ空間上での円形パターン（図 3(a)）を，間引いたデータを用いても再構成できるかを検証した．対象のスパース性を活用せずに，従来通り擬似逆行列を作用させてフーリエ変換像を得た場合は，図 3(b)のように，再構成に失敗する．計測点を間引いたデータに対して，スパースな解を選好する LASSO（Least Absolute Shrinkage and Selection Operator）を用いた場合に円形パターンが再構成できるかを調べると，等間隔に間引いたデータを用いると失敗した（図 3(c)）が，ランダムに間引いたデータを用いると成功した（図 3(d)）．計測手法と解析手法をともに改善させることにより目的を達成できることを示した．

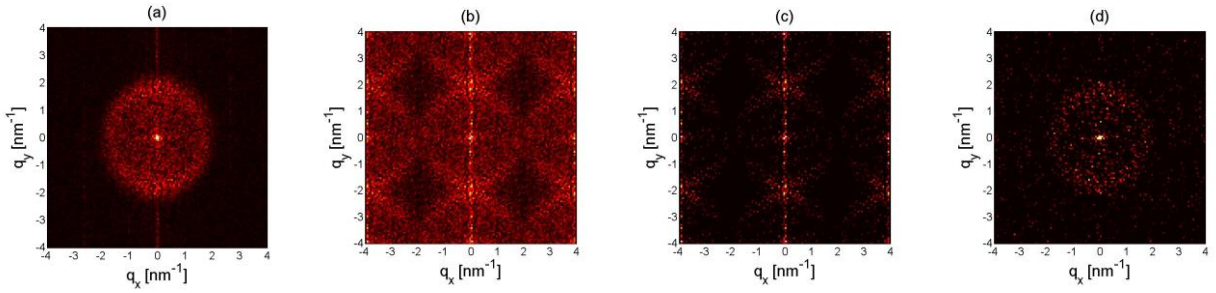

図 3　走査型トンネル顕微・分光法による準粒子干渉に対する CS の適用結果．(a)通常用いられる 256×256（=65536）の画素数をもつ画像データのフーリエ変換像．(b)等間隔に計測点を間引いて 64×64（=4096）の画素数をもつ画像データに擬似逆行列を適用して得られるフーリエ変換像．(c) 等間隔に計測点を間引いて 64×64（＝4096）の画素数をもつ画像データに LASSO を適用して得られるフーリエ変換像．(d)ランダムに計測点を間引いて 4096 の画素数をもつ画像データに LASSO を適用して得られるフーリエ変換像．

HP の事後分布や，最もスパースな解を数値的に求めようとすると，必要な計算量が系の大きさに応じて指数関数的に増大する組合せ爆発の問題が生じる．また，こうした問題の多くは，図 4 左に示すように，多谷のエネルギー関数をもつため，素朴なアルゴリズムを適用しても，数ある局所化の一つに捕らわれ，大域解にたどり着くことはほとんどない．こうした場合にとられる手法は大きく二つに分けられる．一つは，探索空間を狭めることであり，ベイズ推定における変分ベイズ法が含まれる．もう一つは，評価関数自体を，取り扱いが容易なものに置換することであり，スパース解推定における凸緩和法が含まれる．解空

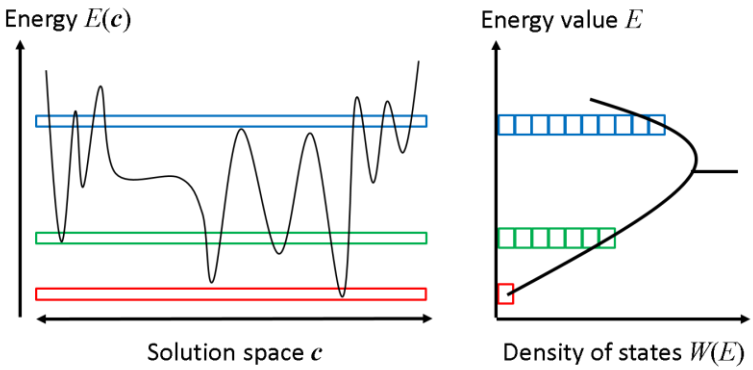

図 4　解空間解析の概念図．左がエネルギーの多谷構造，右が状態密度を表す．

間の構造を明らかにし，これらの手法の性質を理解することが，情報科学の知見を使用するうえで重要である．

解空間は一般に高次元でありその可視化は難しい．そこで状態密度に着目し解空間を解析した．状態密度とは，図 4 右に示すように，目的のエネルギー関数に対して，あるエネルギー関数値を実現する状態が解空間にいくつ存在するかを数え上げ，その数をエネルギー値の関数として表したものである．

具体的に，過完備スパース近似の問題に状態密度解析を適用する．過完備スパース近似とは,高次元ベクトルデータを，与えられた過完備な基底から，少数の基底ベクトルを選んで，できるだけ精度よく近似するという問題である．この問題は NP 困難である．解析には，情報統計力学のレプリカ法を用いて，大自由度極限での状態密度を導出した．この結果を図 5 中の実線により示す．さらに凸緩和法の一種である LASSO と貪欲法の一種である OMP（Orthogonal Matching Pursuit)の性能を状態密度上で比較した．貪欲法は，探索空間を狭める近似法の一つである．結果として，図 5 に示すように，OMP が LASSO より優れているが，理論的に達成可能な性能限界には及ばないことを明らかにした．

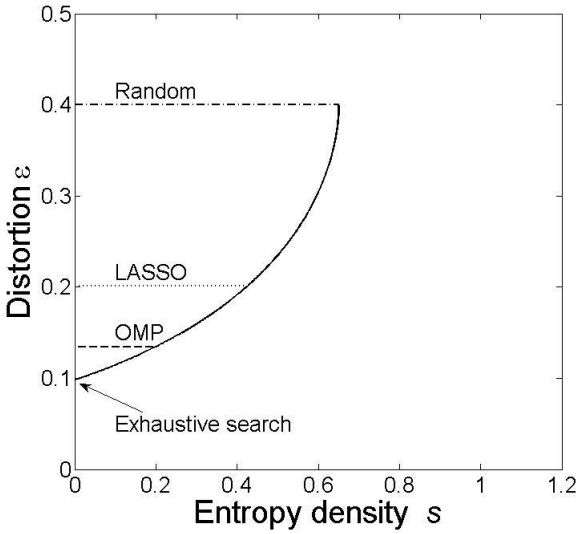

図 5　過完備スパース近似の解空間解析の結果. 縦軸がスパース近似におけるエネルギーに対応する歪値, 横軸が状態密度の対数（エントロピー）を表す．

結論として三つの研究の関連を述べる．自然科学で画像データを取得するということを三つのレベルに照らして考えると理解しやすい．データ取得の目的は，見えないものを見ようとすることである. その戦略は二つある. 一つは, 現象のモデルをたて，画像に直接表れない構造を抽出することであり，モデルを現実に如何に近づけるかが問題となる．これは MRF の研究に対応する．もう一つは，精緻に練られた実験により直接対象を計測することであり, 困難な実験をいかに効率的に行うかが問題となる. これは CS の研究に対応する．いずれの問題も，解空間を全て探索することが重要であり, 解空間解析の知見が鍵となることは言うまでもない. このように，現象・計測・解析からなる三角形を自然科学・情報科学で共有することにより，データ駆動科学を創成する．