様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　　６　日現在

機関番号：　１２６０１
研究種目：　新学術領域研究（研究領域提案型）
研究期間：　2013～2017
課題番号：　２５１２０００９

研究課題名（和文）スパースモデリングによる潜在構造の抽出

研究課題名（英文）Extraction of latent structure by sparse modeling

研究代表者
岡田　真人（Okada, Masato）

東京大学・大学院新領域創成科学研究科・教授

研究者番号：９０２３３３４５

交付決定額（研究期間全体）：（直接経費）　　75,700,000円

研究成果の概要（和文）：スパースモデリング班（B01-2）による研究は3つの課題について行われた．課題1ではベイズ推定によるスペクトル分解法の実データ適用の他，ノイズ分散推定法や，L1正則化つき回帰を利用した計算量が軽く高次元スペクトルにも適用可能な手法を開発した．課題2では基底が不明な場合にSpDMDを用いた基底の推定と選択を行う手法の開発，実データへの適用を行った．課題3では，全数探索を用いて複数の基底の組み合わせの適切さを評価する手法を開発し，実データへの適用を行った．これら3課題の研究によって，SpMを用いて潜在構造を抽出する普遍的な手法が開発され実データによる検証が行われた．

研究成果の概要（英文）：The sparse modeling team (B01-2) sets three tasks. Task 1 is applications of Bayesian spectral decomposition method to actual data. We developed a noise variance estimation method and a fast calculation method using L1 regularization and verified its effectiveness with actual data. In task 2, we developed a basis estimation and selection method using Sp-DMD for time series data, and applied it to actual data and verified its effectiveness. In task 3, a method of evaluating the appropriateness of the basis combination using an exhaustive search was developed, and this method was applied to actual data and the effectiveness was verified. Through research on these three tasks, we developed a universal method to extract latent structures using SpM and verified its effectiveness by actual data.

研究分野：　高次元データ駆動科学

キーワード：　データ駆動科学　潜在構造抽出　スペクトル分解　ブラインドセンシング　スパースDMD　全状態探索　ES-DoS

１．研究開始当初の背景

今日の科学は，仮説に基づき実験・観測を行い，得られたデータの潜在的な物理特性を説明する少数の変数（説明変数）からなる仮説(モデル)の検証を行う，という仮説の提案／検証ループの不断の繰り返しにより発展してきた．近年の実験・観測技術の発展は研究者に大量の高次元観測データを与える一方で，データの肥大化により研究者の直感的行為である思索や試行錯誤，さらにはデータと説明変数の比較に基づく仮説の提案／検証ループを著しく困難にしている．

本計画研究では，実用的時間で大量の高次元データから系の低次元の潜在構造を効率よく抽出するスパースモデリング(SpM)という考え方を指導原理として，生物学・地学・天文・医療・物質科学などの分野で得られる実験・計測データから，系の潜在構造としての物理特性を抽出する普遍的手法を開発し，仮説の提案／検証ループの繰り返しを促進する枠組みを目指した．

２．研究の目的

本計画研究は, SpM により潜在構造を抽出する普遍的手法を開発した. SpM によりデータの潜在構造を抽出することで，研究者による直感が働かないような高次元，大規模なデータであっても解釈可能になり，仮説／検証ループを働かせることが可能になる．本計画研究では，分野の個別性を超えた類似性／共通性に基づく高次元データ駆動科学におけるモデリング原理の確立を目的とした．

スパースな構造を抽出する SpM の問題は，問題背後の構造(基底)がわからない場合に基底を推定するブラインドセンシングの問題と，基底の候補がわかっていて最適な候補の組み合わせが一つに決まる場合と，一つに決まらない場合とに分類できる．こうした考察に基づき，三つの課題を設定した．まず最も簡単なケースとして，基底の候補がわかっていて最適な基底の組み合わせを選ぶ問題として，スペクトル分解を用いたモデリングを【課題 1】に設定した. 次に,【課題 2】では，基底がわからない，もしくは適切と思われる基底の候補が定まらない場合に用いるブラインドセンシング（BS）を用いたモデリングを行った．さらに,【課題 3】では，全数探索によるアプローチで，最適な基底の組み合わせが複数ある場合について取り組み，また，モンテカルロ法による高速全数探索手法を開発した．以上 3 課題により特定の分野に留まらない高次元データ駆動科学におけるモデリング法が確立されることを目指した．

３．研究の方法

【課題 1】スペクトル分解を用いたモデリング

多峰性スペクトルを適切な個数の単峰性の基底関数の線形和に分解する問題は，数理統計学の難問の一つである特異モデルの一つであり，X 線光電子分光法(XPS)，NMR，光の反射スペクトル解析など，あらゆる自然科学の計測でも必要不可欠である．本課題では，このスペクトル分解の問題を多様な対象に適用するための系統的方法論を構築する．

【課題 2】ブラインドセンシング(BS)を用いたモデリング

基底関数自体が未知であるスペクトル分解の問題も数多く存在し，ブラインドセンシングと呼ばれる．基底関数が未知の中でも，もっともらしい候補は挙げられる状況について過完備基底による次元削減を行う．候補自体も挙げられずデータから基底関数を推定する必要がある状況に対してはと SpDMD を用いた基底の推定と選択を行う．電子構造などの物理特性を記述する潜在構造の自動抽出を行う枠組みを構築する．

【課題 3】モンテカルロ法による高速全数探索を用いたモデリング

データの高次元化による自由度の増加により最適な組み合わせの候補を複数選択する必要があるケースがある．これを全状態探索（ES）における状態密度（DoS）を可視化することにより，ある組み合わせの候補に対して適切性を評価する．さらに探索すべき基底の組み合わせが爆発する問題をレプリカ交換モンテカルロ法により解決する．これらの手法を人工データを用いて開発・検証を行い. さらに地球科学班(A02-1)の津波堆積物解析，神経活動データ，天文データなどに応用することで多くの説明変数の組み合わせからある組み合わせの適切性を評価する枠組みを構築する．

４．研究成果

実用的時間で大量の高次元データから系の低次元の潜在構造を効率よく抽出する SpM という指導原理により，生物学・地学・天文・医療・物質科学などの分野で仮説の提案／検証ループの繰り返しを促進する枠組みを確立し，これらの広範な分野で，系の潜在構造としての物理特性を抽出できることを実証した．これは，自然科学では，高次元データにスパース(疎)性が内在するという仮説が普遍的に成り立つことを意味する．

【課題 1】スペクトル分解を用いたモデリング

我々は，本研究課題の研究開始前に開発したレプリカ交換モンテカルロ法を用いたベイズ推定によるスペクトル分解法を，生命科学班(A01-2)の NMR データ[6]や地球科学班(A02-1)の鉱石データ[7]に適用した．これによって，スペクトルの由来となったアミノ酸の同定や，カンラン石に含まれる鉄・マグネシウム比といった物理的情報のスペクトルデータからの抽出に成功した．さらに，それまでのスペクトル分解法では既知として事前に与える必要のあった，スペクトルのノイズ分散を，それを確率変数としてベイズ的に扱うことで推定することを実現した[5]. その推定では新たな計算を行うことなく，レプリカ交換モンテカルロ法においてサンプリングの効率化の観点から導入されるレプリカ準位の情報を有効活用し，計算量の増大を伴わないノイズ推定を実現した．この手法を用いることで，事前にノイズ量がわからない実際の実験系におけるスペクトル分解が実現される．これが，赤井公募班が行なったメゾスコピックな量子系の計測データに対するスペクトル分解において確認された[2].

一方で，ベイズ的スペクトル分解法は，モデルパラメータの多い多峰スペクトルや，高次元スペクトルなどに適用することは，計算量の観点から困難である．そこで我々は，SpM を用いた L1VM スペクトル分解法を開発した[8]. 物理モデリング班(B01-3)の走査型トンネル型電子顕微鏡の 2 次元スペクトルデータに対する L1VM スペクトル分解法の適用が成功したことからも，それが有効に働くことが確認された[9].

現在，本課題を通して確立されたスペクトル分解法は，構造材料や原子核実験といった幅広い分野で活用されるに至っている．

【課題 2】ブラインドセンシング(BS)を用いたモデリング

【課題 1】では現象を記述する基底が陽にわかっている状況において，得られたデータを個々の基底に分解する枠組みを構築した．一方で，自然科学におけるデータは基底関数自体が未知である場合も数多く存在し，ブラインドセンシング(BS)のアプローチが有効である．ブラインドセンシングとは直接観測できないデータの背後の構成要素（基底）をデータのみから推定及び選択する枠組みの総称である．基底関数がわからない状況は大きく 2 つに分けられる．1 つ目は基底関数が未知の中でも，もっともらしい候補は挙げられる状況，2 つ目は候補自体も挙げられずデータから基底関数を推定する必要がある状況である．それぞれの状況に対して，我々は過完備基底による次元削減と SpDMD を用いた基底の推定と選択を行った．

過完備基底による次元削減：自然科学ではデータが希少であり，基底を一から学習する，従来の BS を適用できない場合がある．代わりに，各分野の知見により複数の基底候補を用意できる場合，過完備基底による次元削減を BS の手法として適用できる．過完備基底の手法は基底選択を含むため，凸緩和法や貪欲法の適用を検討した．圧縮センシングでは，厳密解に一致することもある凸緩和法は性能が悪く，過完備基底の手法を実行するには貪欲法が優れていることが分かった[10].

SpDMD を用いた基底の推定と選択：現象が特定の基底関数の和で表されるが，その基底関数の形が重要となる典型的な対象としてコヒーレントフォノン(CP)が挙げられる. CP は分光法における巨視的量子現象の代表であり，物質科学の分野で注目されている．物質にパルスレーザーを照射し，励起された状態が緩和する過程を時系列データとして計測することで物質の構成分子の種類を非破壊で調べることができる．得られた原子振動は減衰振動の重ね合わせで記述され，この減衰振動の周波数と初期位相に物質の情報が

表現される．我々は，CP の典型的な対象である構造が既知の Bi(111)を解析した．Sparsity-promoting dynamic mode decomposition (SpDMD)を用いることで，物理的に重要な基底を抽出し位相を高い精度で推定できることを示した[3]．SpDMD は，時系列データを時間的にカップリングした動的成分（基底）の和に分解し，各基底に対してスパース性を仮定して刈り込みを行うという手法であり，基底の推定と選択を両方行うことができる．SpDMD で得られた基底には物理的に重要な基底だけでなく，これまでは実験研究者が恣意的に引いていたバックグラウンドの成分も含まれており，本手法によってバックグラウンド成分の除去も自動的に行えることを示唆していた．また，位相推定に従来用いられていたフーリエスペクトルを用いた手法は減衰振動や基底のスパース性を仮定していないため，SpDMD では各々の基底の位相がより高精度に推定できた．

【課題 3】モンテカルロ法による高速全数探索法を用いたモデリング

まず，与えられたデータに対して意味のある説明変数の組を選択する変数選択の問題において，Lasso や SLR などの既存のスパース推定手法は近似的な手法であり，それぞれ異なる結果を導く．我々はこの問題に対してすべての組み合わせについて情報量基準を評価する全状態探索（ES）を行った．また複数の情報量基準の頻度分布を可視化し，この分布を状態密度（DoS）として解釈することで，各組み合わせから得られる性能の全組み合わせ内での評価を定める事ができる．このように変数選択の解や近似手法の性能を評価する包括的な枠組みを ES-DoS として提案した．我々は ES-DoS を様々な分野へ応用した．地球科学班（A02-1）と共同し，堆積物の含有元素データから津波由来か否かを二値判別し，判別に重要となる元素の組を説明変数として選択する問題に適用した．また市川公募班（B01）とは発達障害児における ASD 群と ADHD 群の識別に ES-DoS を適用し重要なチャンネルの組み合わせの抽出に成功した．さらに，計算量が爆発し全状態探索を実行できない場合でも近似的に DoS を得る手法としてマルコフ連鎖モンテカルロ法を利用し，AES-DoS として提案した[1].

これらの全状態探索手法では，何らかの情報量基準に基づいて説明変数の組み合わせを評価する．我々はこの情報量基準として，従来よく用いられていた交差検証誤差（CVE）だけではなく，ベイズ推定に基づく自由エネルギー（FE）などの複数の情報量基準を合わせて DoS を評価する仕組みを提案した．我々は天文学班（A02-3），計測モデリング班（B01-1）と共同し，人工データおよび天文の実データを対象に手法を適用した．人工データによる検証の結果，データが少なければ全状態を探索したとしても真の潜在構造が得られない場合があることを発見した[4].

５．主な発表論文等

〔雑誌論文〕（計 22 件）

（主な成果 10 件を抜粋）

[1]. ES-DoS: Exhaustive search and density-of-states estimation as a general framework for sparse variable selection, Yasuhiko Igarashi, Hiroko Ichikawa, Yoshinori Nakanishi-Ohno, Hikaru Takenaka, Daiki Kawabata, Satoshi Eifuku, Ryoi Tamura, Kenji Nagata, *Masato Okada, Journal of Physics: Conference Series, in press (査読有)

[2].Bayesian Spectroscopy of Admixed Photoluminescence Spectra with Exciton, Biexciton and Electron Hole Droplet States in a GaAs/AlAs Type-II Superlattice, *K. Iwamitsu, Y. Furukawa, M. Nakayama, M. Okada, I. Akai, J. Lumin., 197, 18-22, 2018 (査読有)

[3]. Analysis of Coherent Phonon Signals by Sparsity-promoting Dynamic Mode Decomposition, S. Murata, S. Aihara, S. Tokuda, K. Iwamitsu, K. Mizoguchi, I. Akai, *M. Okada, J. Phys. Soc. Jpn., 87, 054003-1-5, 2018 (査読有)

[4]. Exhaustive search for sparse variable selection in linear regression, Yasuhiko Igarashi, Hikaru Takenaka, Yoshinori Nakanishi-Ohno, Makoto Uemura, Shiro Ikeda, *Masato Okada, Journal of the Physical Society of Japan, 87, 44802, 2018 (査読有)

[5].Simultaneous estimation of noise variance and number of peaks in Bayesian spectral deconvolution, Satoru Tokuda, Kenji Nagata, *Masato Okada, Journal of the Physical Society of Japan, 86(2), 024001, 2017 (査読有)

[6].NMR spectral analysis using prior knowledge, *T. Kasai, K. Nagata, M. Okada, *T. Kigawa, Journal of Physics: Conference Series, 699(1), 012003, 2016 (査読有)

[7].An automatic deconvolution method for Modified Gaussian Model using the Exchange Monte Carlo method: application to reflectance spectra of synthetic clinopyroxene, *P.K. Hong, H. Miyamoto, T. Niihara, S. Sugita, K. Nagata, J.M. Dohm, M. Okada, Journal of Geology & Geophysics, 5: 3, 1-15, 2016 (査読有)

[8].Three levels of data-driven science, Yasuhiko Igarashi, Kenji Nagata, Tatsu Kuwatani, Toshiaki Omori, Yoshinori Nakanishi-Ohno, *Masato Okada, Journal of Physics: Conference Series, 699, 012001, 2016 (査読有)

[9]. Compressed sensing in scanning tunneling microscopy/spectroscopy for observation of quasi-particle interference, Yoshinori Nakanishi-Ohno, Masahiro Haze, Yasuo Yoshida, Koji Hukushima, Yukio Hasegawa, *Masato Okada, Journal of the Physical Society of Japan, 85(9), 093702, 2016 (査読有)

[10]. Sparse approximation based on a random overcomplete basis, Yoshinori Nakanishi-Ohno, Tomoyuki Obuchi, Masato Okada, *Yoshiyuki Kabashima, Journal of Statistical Mechanics: Theory and Experiment, 2016, 063302, 2016 (査読有)

〔学会発表〕（計 48 件）

1. 岩満一功, 相原慎吾, 溝口 幸司, 五十嵐 康彦, 村田 伸, 岡田 真人, 赤井 一郎, SpDMD によるコヒーレントフォノンの減衰振動モード分解, 2017 年度人工知能学会全国大会 (JSAI2017), 2I2-1, 2017 年.

2. 本武陽一, 五十嵐康彦, 竹中光, 永田賢二, 岡田真人, スペクトル分解における λ-スキャン法の提案, 信学技報, 117(293), IBISML2017-80, 325-332, 2017 年.

3. 徳田悟, 永田賢二, 岡田真人, 時間分解分光計測のベイズ的最適設計, 2016 年度人工知能学会全国大会 (JSAI2016), 2016 年.

4. Satoru Tokuda, Kenji Nagata, and Masato Okada, Phase transitions of statistical estimation, Statphys26（国際学会）, 2016 年.

5. Hikaru Takenaka, Kenji Nagata, Takashi Mizokawa, and Masato Okada, Bayesian Model Selection of NiGa2S4 Triangular Lattice with Boltzmann Factor, Statphys26（国際学会）, 2016 年.

6. 赤井一郎, 村田伸, 相原慎吾, 徳田悟, 岩満一功, 岡田真人, SpDMD によるコヒーレントフォノン信号のモード分解解析(I), 日本物理学会, 2016 年秋季大会, 2016 年.

7. 五十嵐康彦, 竹中光, 中西 (大野)義典, 植村誠, 池田思朗, 岡田真人, 全状態探索による線形回帰のスパース変数選択, 第 19 回情報論的学習理論ワークショップ (IBIS2016), 2016 年.

8. Yasuhiko Igarashi, Kenji Nagata, Tatsu

Kuwatani, Toshiaki Omori, Yoshinori Nakanishi-Ohno, Masato Okada, Three levels of data-driven science, International Meeting on “High-Dimensional Data Driven Science” (HD3-2015)（招待講演）（国際学会）, 2015 年.

9. Satoru Tokuda, Kenji Nagata, Masato Okada, A theory of phase transitions and crossovers in statistical estimation: Toward a data-driven approach for physical science, International Meeting on “High-Dimensional Data Driven Science” (HD3-2015)（国際学会）, 2015 年.

10. 村田伸, 永田賢二, 植村誠, 岡田真人, 時系列スペクトルデータからの潜在的動力学推定, ニューロコンピューティング研究会, 2015 年.

11. Kenji Nagata, Sparse modeling and variable selection with the exchange Monte Carlo method, Workshop on Mathematical Approaches to Large-Dimensional Data Analysis, 2014 年.

12. 徳田悟, 永田賢二, 渡辺澄夫, 岡田真人, 交換モンテカルロ法を用いた情報量規準 WBIC の計算機実験による検証, ニューロコンピューティング研究会, 2014 年.

13. 永田賢二, 村岡怜, 佐々木岳彦, 岡田真人, ベイズ推定に基づくスペクトル分解と必要最小計測時間の推定について, ニューロコンピューティング研究会, 2014 年.

14. Kenji Nagata, An Efficient Exhaustive Search for Variable Selection Using MCMC method, Inference, Computation, and Spin Glasses (ICSG2013), 2013 年.

15. 永田賢二, 北園淳, 中島伸一, 永福智志, 田村了以, 岡田真人, 交換モンテカルロ法を用いた変数選択問題における解の効率的全数探索, 第 16 回情報論的学習理論ワークショップ, 2013 年.

16. 若杉健介, 桑谷立, 永田賢二, 麻生英樹, 岡田真人, 連想記憶モデルを用いた確率的潜在構造抽出アルゴリズムの有効性の検証, ニューロコンピューティング研究会, 2013 年.

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）
○取得状況（計　0　件）

〔その他〕
ホームページ等
http://mns.k.u-tokyo.ac.jp/home.html

６．研究組織
(1)研究代表者
岡田　真人（Okada, Masato）
東京大学・大学院新領域創成科学研究科・教授
研究者番号：90233345

(2)研究分担者
田中 和之（Kazuyuki, Tanaka）
東北大学・大学院情報科学研究科・教授
研究者番号：80217017

村田 昇（Murata, Noboru）
早稲田大学・大学院先進理工学研究科・教授
研究者番号：60242038

井上 真郷（Inoue, Masato）
早稲田大学・大学院先進理工学研究科・教授
研究者番号：70376953

永田　賢二（Nagata, Kenji）
国立研究開発法人産業技術総合研究所・人工知能研究センター・主任研究員
研究者番号：10556062