公開

研 究 終 了 報 告 書

「多変数間に潜む高次相互作用の探索と分解」

研究期間: 2018 年 10 月～2022 年 3 月
研 究 者: 杉山 麿人

1. 研究のねらい

多変量データにおける変数間の相互作用の発見や解析は，多数の変数から構成されるシステムを解析する際の基本的かつ本質的な課題である．近年では，センサ技術や IoT などの情報技術が急速に発達し，遺伝学や脳科学，社会科学などの基礎科学から，医療などの応用科学まで，幅広い領域で多様な変数に関する多変量データが獲得されている．それらの多変量データにおける変数間の相互作用の解析は，データの背後に潜む現象を明らかにするための記述的（descriptive）分析における最も基本的な解析手順の1つであり，ありとあらゆる記述的データ分析で必須といっても過言ではない．

最も単純なケースである 2 変数間の相互作用を扱う場合は，その強度を測定すればよく，問題は単純である．線形の関連性を見つけることができる Pearson の相関係数にはじまり， 非線形の関連性を見つける相互情報量や MIC, MID など，数多くの尺度が提案され研究が進んできた．しかし，3 つ以上の変数が関わる高次相互作用（higher-order interaction）の発見は，これらの手法の naïve な一般化では解決することができない．任意の次数を考慮するためには，n 個の変数に対して 2 の n 乗個の高次相互作用の候補が存在し，その探索空間が組合せ爆発を起こしてしまうためである．さらに，3 つ以上の変数が関わる高次相互作用の働きを明らかにするためには，単に強度を測定するだけでなく，3 次以上の相互作用から単一の変数（1 次）や変数のペア（2 次）といった低次の相互作用への分解を可能とする理論が新たに必要となる．

そこで本研究では，高次相互作用の探索と分解を実現する解析技術を確立することをねらいとした．具体的には，多数の変数からなる多変量データやそのモデルにおいて，効率的に高次相互作用を探索するためのアルゴリズムを設計した．そして，発見された高次相互作用に対して変数のノックアウトをモデル上で適用することで相互作用を分解し，精緻な解析の実現を目指した．本研究で構築した基礎理論および実践的手法は，遺伝学，脳科学，社会科学，ヘルスケアなどの社会システムの根幹に関わる幅広い領域において，それらを形成する変数間の高次相互作用の解析を初めて可能にするものであり，今後の社会システム形成において重要な役割を果たすことが期待される．

2. 研究成果

（1）概要

まず，変数間の高次相互作用を含む可能性があるシステムにおいて，そこから生成される多変量データから高次相互作用を探索するための効率的アルゴリズムを構築した．この問題は，確率モデルの観点からは期待値が大きい相互作用の組を見つける，という問題として定式化することができるが，総当りで調べることは計算量の観点から実現不可能である．また，見つけた相互作用は必ず偽陽性の可能性がある．そこで，アイテム集合マイニングのアルゴリズムを利用して効率的な探索を実現するとともに，仮説検定と多重検定補正を組み合わせ

ることで偽陽性相互作用を制御しつつ，統計的に有意に出現する変数間相互作用をすべて発見するための手法の構築に成功した．

次に，情報幾何学の理論に着目し，変数間の相互作用を含むシステムから生成されるデータを確率分布として扱うことで，相互作用の分解を確率分布の集合である多様体上での操作として実現するアプローチを構築した．より具体的には，相互作用の集合がなす離散的な階層構造を半順序とみなし，半順序集合に対して導入される隣接代数を用いて確率モデルを構築することで，相互作用の分解をパラメータ空間での射影として実現することができることを示した．機械学習や統計物理で知られるボルツマンマシン（イジングモデル）や，エントロピー正則化最適輸送問題は，この提案モデルの特殊な場合に対応していることが理論的に明らかとなった．さらに，テンソルの分解や低ランク近似が同じくこの提案モデルの特殊な場合となっていることを明らかにし，平均場近似と組み合わせることで，テンソル低ランク近似を勾配法なしで解析的に求める高速なアルゴリズムの構築に成功した．

社会システムデザインにおける応用として，構築した分解手法をブラインド信号源分離へと適用し，既存手法と比較して高精度で信号源が分離できることを示した．さらに，多体波動関数の推定についての共同研究を進め，高次相互作用を導入したボルツマンマシンを用いることで，より高精度な推定が可能となることを示した．提案手法を応用することで，機械学習手法の実装の信頼性担保についても取り組み，Lasso 実装のバグを検出する手法の構築にも成功した．これらの手法は，今後の社会システムデザイン構築において欠かせない技術となることが期待される．

（2）詳細

研究テーマ A「高次相互作用の探索」

データから高次相互作用を効率的に探索するためのアルゴリズムを構築した．データに n 個変数がある場合，とりうる相互作用の個数は 2^n となるため，計算量が組合せ爆発を起こしてしまい，総当りで調べることは不可能である．また，データはあくまで有限のサンプルであるため，相互作用があるように見えたとしても，データに偶然に出てきているだけ，すなわち偽陽性の可能性がある．そこで，アイテム集合マイニングのアルゴリズムを利用して効率的な探索を実現するとともに，仮説検定と多重検定補正を組み合わせることで偽陽性相互作用を制御しつつ，統計的に有意に出現する相互作用をすべて発見するための手法を構築した．これまでに，0 や 1 のみからなる離散データに対する手法はすでに研究が進んでいるが，連続値データに対する手法は存在していなかった．本研究では初めて，連続値データから統計的に有意な相互作用を発見するアルゴリズムを実現した．

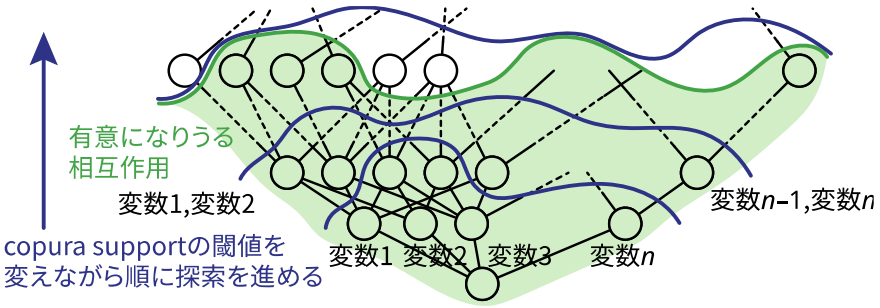

図 1: 相互作用の探索．

鍵となるアイデアは，copura support と呼ばれる一種の期待値をデータから求めることで，この値を用いた分枝限定法を実現することである（図 1）．さらに，copura support を用いた尤度比検定を抽出された各相互作用に適用し，Tarone 法に基づく多重検定補正を実施することで，偽陽性の発生する確率（FWER; Family-Wise Error Rate）を任意の値に制御することを可能とした．実データを用いた実験によって，連続値データを離散化してから既存手法を適用するアプローチと比較して，より精度良く相互作用が発見できることを示した．この研究は，ETH Zürich の Prof. Karsten Borgwardt 氏との共同研究に基づく．得られた成果は，The 28th International Joint Conference on Artificial Intelligence (IJCAI 2019) で発表した．

研究テーマ B「高次相互作用の分解」

まず高次相互作用を表現しているデータ形式としてテンソルに着目し，テンソルをより柔軟に分解するための手法を構築する研究に取り組んだ．これまでに，CP 分解や Tucker 分解といった手法が提案され，幅広い分野で利用されている．しかし，任意の次数のテンソルに対して一貫した枠組を与えるための統計的理論は，未だ発展途上である．そこで，この問題を解決するために情報幾何学の理論に着目した．情報幾何学では，対象を確率分布として扱うことで，様々な処理を確率分布の集合である多様体上での操作として実現する．各テンソルを，離散構造を持った確率分布として扱うことで，任意の次数に対するテンソル分解を，多様体上の射影として実現できることを発見した（図 2）．そこで，この性質に基づくルジャンドル分解という新たなテンソル分解手法を確立し，理論的な解析をおこなうとともに，実データでその性能を検証した．特に，画像データを用いた実験によって，CP 分解や Tucker 分解といった既存手法よりもテンソル再構成の精度が高いことを示した．この成果は，The 32nd Annual Conference on Neural Information Processing Systems (NeurIPS2018) にスポットライト発表として採択された（採択率 168/4856 = 3.5%）．また，ルジャンドル分解はボルツマンマシンの学習の一般化として捉えることができ，より一般の高次相互作用分解手法の研究に繋がる理論的基盤となっている．

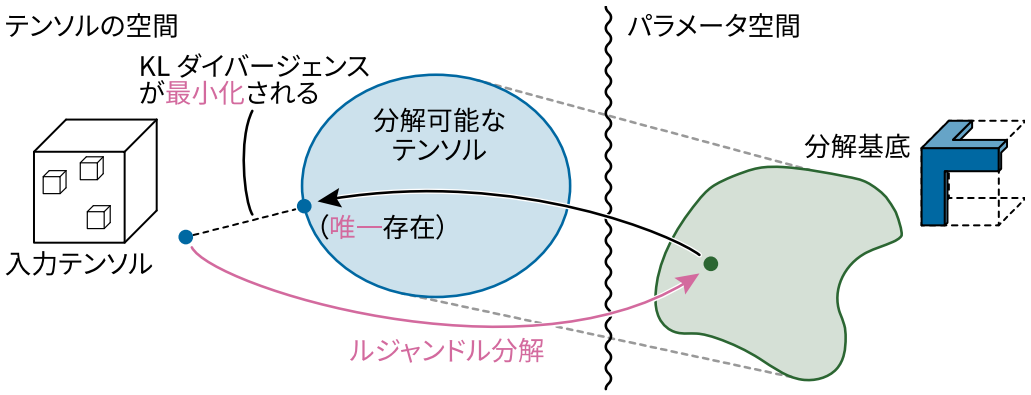

図 2: 多様体上でのテンソル分解．

テンソルの Tucker 分解では，Tucker ランクを落として低ランク近似を実現しているが，この

操作を情報幾何学的に解釈することで，テンソルの低ランク近似がルジャンドル分解と同様の枠組みで多様体上での射影として実現できることを発見した．この発見によって，平均場近似を用いた非勾配法による低ランク近似が可能となり，初期値や学習率に影響されない，高速かつ安定したテンソル低ランク近似アルゴリズムの構築に成功した．この成果は，The 35th Annual Conference on Neural Information Processing Systems (NeurIPS2021)に採択済みである．

情報幾何学を用いたテンソル分解手法は，東京大学の津田宏治氏及び理化学研究所の中原裕之氏との共同研究である．さらに，津田氏が研究代表者を務めている CREST「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」領域の研究テーマ「離散構造統計学の創出と癌科学への展開」とも密接に連携をしている．

研究テーマ C「高次相互作用分解の応用」

研究テーマ B で構築した手法に基づき，ブラインド信号源分離や点過程推定を可能とするアルゴリズムの構築に成功した．特に，ブラインド信号源分離については，観測された信号を可視変数，推定したい信号源を隠れ変数と過程し，それらの変数間の相互作用を含む階層的な情報幾何学的モデルを構築することで，学習によるデータからの推定を可能とした(図 3)．ICA に基づくような既存手法では，信号源間の高次相互作用を取り扱うことができないため，それらの既存手法と比較して，提案手法はより高精度で信号源を復元することができる．この成果は，The 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021)で発表した．

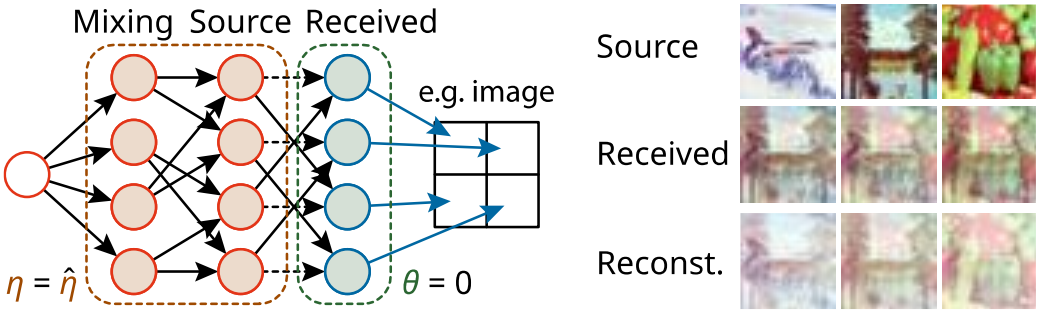

図 3: ブラインド信号源分離で用いる階層モデルと得られた結果の例．

上記の研究と平行して，同様の階層的な確率モデルを利用した多体波動関数の推定についての共同研究を進め，高次相互作用を導入したボルツマンマシンを用いることで，より高精度な多体波動関数推定が可能となることを示した．この結果は，Journal of Chemical Theory and Computation 誌で発表した．

さらに，本研究の発展として，機械学習実装の信頼性担保にも取り組んでいる．特定の機械学習アルゴリズムでは，変数間の相互作用に基づき，多面体領域として可能な入出力の範囲が特定されることが知られている．もともとこの性質は，機械学習で発生したバイアスを補正した分布を作成する selective inference を実現するために必要となる polyhedral lemma として知られていたが，この機械学習アルゴリズムが持つ性質を利用することで，この範囲から逸脱する入出力のペアが検出された際に，実装にバグが混入していると断定することができる．

本研究では，このアプローチを実装バグ検出手法として確立し（図 4），機械学習手法 Lasso の実装に混入したバグを高精度で検出できることを示した．この成果は，The 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2020) の Visions and Reflections Track で発表した．

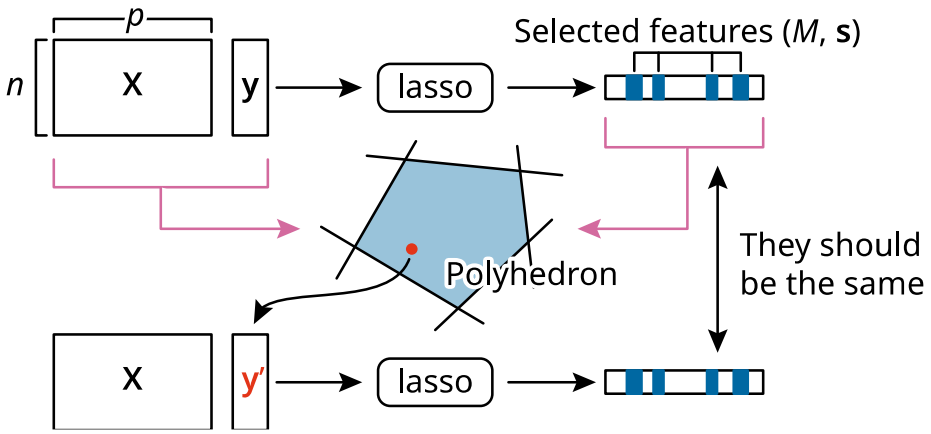

図 4: Lasso 実装のバグ検出．

ブラインド信号源分離や点過程推定といった研究は，シドニー大学の Prof. Lamiae Azizi 氏，Simon Luo 氏，Prasad Cheema 氏らとの共同研究に基づく．また，モデルの学習アルゴリズムについては，東京大学の松島慎氏との共同研究を進めている．さらに，量子化学への応用は，名古屋大学の柳井毅氏及び東京大学の津田宏治氏との共同研究に基づいている．特に，柳井氏はマテリアルズインフォ領域のさきがけ研究者であり，異なるさきがけ領域間での共同研究となっている．

3. 今後の展開

本研究での中心的課題として取り組んだ高次相互作用分解のアルゴリズムについて，実応用領域との共同研究により，社会実装をすすめる．

本さきがけ研究では，主にデータとして与えられている変数について精緻な解析を実現する，というねらいのもと研究を進めてきた．一方，最先端の機械学習・データ解析手法では，主に深層学習などにおいて，隠れ変数と呼ばれるデータには存在しない変数を大量に使うことで予測精度を向上させるアプローチが主流である．特に，隠れ変数（パラメータ数に対応）を大量に増やすことを過剰パラメータ化（overparameterization）と呼び，過剰パラメータ化でのモデルが積極的に用いられている．これまでのバイアスバリアンストレードオフのレジームを超えて，過剰パラメータ化でなぜ良い性能が出るのか，理論的にも不明な点が多い．そこで，本研究で構築した情報幾何学的な枠組みを利用して，モデルが持つ表現力を幾何学的に特徴づけることで，これらの課題の解決を目指すとともに，実用的な手法の開発に着手する．これらの内容をねらいとした研究計画が，すでに JST の創発的研究支援事業に採択されており，この支援のもとで研究を継続する．

また，本研究で取り組んだ機械学習実装の信頼性担保について，社会実装を実現するためには，より広範なアルゴリズムについての適用可能性を模索する必要がある．現在，多様なソース

コードについて統計的性質での特徴づけに取り組んでいるが，ソースコードが持つ自然性というより本質的かつ重要な問題を含むことが判明しつつあり，この課題については長期的な研究として取り組む予定である．

4．　自己評価

研究目標の達成状況

申請時に掲げた高次相互作用の探索手法，及び分解手法について，どちらの課題に対しても手法を確立することに成功し，人工知能分野，及び機械学習分野での複数の論文として発表するなど，計画通り達成することができた．さらに，この研究を進めるうえで，研究開始時にはまったく計画していなかった，過剰パラメータ化モデルの解析との関連を見出し，今後の研究プロジェクトに繋げるなど，計画以上の進展があった．一方，社会システムでの実装や実現については，ブラインド信号源分離への応用や機械学習実装の信頼性担保など，いくつかの領域で萌芽的な結果を得ることに成功した．さらに，多体波動関数推定においては，他のさきがけ領域や CREST 領域との共同研究により，情報科学の枠を超えた成果を得ることがきた．以上，本研究の研究目的は，理論面では計画以上，応用面では計画どおりの内容が達成できたと考えている．

研究の進め方（研究実施体制及び研究費執行状況）

予定通り研究を進めた．3 年目および 4 年目は，新型コロナの影響があったが，研究体制を拡充するなど，柔軟に対応しつつ，各年度で研究費を計画通り執行することができた．

研究成果の科学技術及び社会・経済への波及効果

本研究で得られた多くの研究成果は，現時点では学術的な貢献に留まっているが，本研究で達成した研究成果は，国際的な水準に照らしても独創的かつインパクトのある先駆的な内容であると自負している．今後も引き続き研究に取り組むことで，よりインパクトの大きい社会実装へと繋げることができると考えている．

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：22件

1. Sugiyama, M., Borgwardt, K., Finding Statistically Significant Interactions between Continuous Features, *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (IJCAI 2019), 2019, 3490-3498
統計的に有意な変数間の高次相互作用を，多重検定補正を実施しつつ連続値データから網羅的に発見する効率的アルゴリズムを構築した．これまで，どのように多重検定補正を実施するのか，どのように組合せ爆発を回避して効率的列挙を実現するのか，の 2 点が困難な課題だった．本論文では，copula support と呼ばれる連続値データに直接適用可能な指標に着目し，Tarone の検定可能性による多重検定補正とアイテム集合マイニングのアルゴリズムを統合することで，この問題を解決した．

2. Sugiyama, M., Nakahara, H., Tsuda, K., Legendre Decomposition for Tensors, *Advances in Neural Information Processing Systems* (NeurIPS2018), 2018, 31, 8825–8835
情報幾何学を用いてテンソル分解に新たな理論的枠組みを与え，さらに新規のテンソル分解手法を構築し，一般的な手法と比較した際の優位性を明らかにした．これまで，行列を含む任意の次数のテンソルに対して，凸最適化での分解は困難であった．本論文では，テンソルを統計多様体上の要素として扱うという独自のアイデアを採用することで，分解を部分多様体への射影として凸最適化で実現することに成功した．
3. Ghalamkari, K., Sugiyama, M., Fast Tucker Rank Reduction for Non-Negative Tensors Using Mean-Field Approximation, *Advances in Neural Information Processing Systems* (NeurIPS2021), 34, 443–454, 2021
テンソルの Tucker ランクにおける低ランク近似操作を情報幾何学的に解釈することで，テンソルの低ランク近似が多様体上での射影として情報幾何学的に実現できることを発見した．この発見によって，平均場近似を用いた非勾配法による低ランク近似が可能となり，初期値や学習率に影響されない，高速かつ安定したテンソル低ランク近似アルゴリズムの構築に成功した．

（2）特許出願

研究期間全出願件数：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）

主要な学会発表

・Luo, S., Sugiyama, M., Bias-Variance Trade-Off in Hierarchical Probabilistic Models Using Higher-Order Feature Interactions, The 33rd AAAI Conference on Artificial Intelligence (AAAI-19), 2019, 33(01), 4488-4495.

・Luo, S. and Azizi, L. and Sugiyama, M., Hierarchical Probabilistic Model for Blind Source Separation via Legendre Transformation, Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021), 2021, 161, 312-321.

・Ghalamkari, K., Sugiyama, M., Fast Rank-1 NMF for Missing Data with KL Divergence, Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS 2022), 2022.

・Matsue, K., Sugiyama, M., Unsupervised Tensor based Feature Extraction and Outlier Detection for Multivariate Time Series, Proceedings of 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), 2021, 1–12.

受賞

・Cheema, P., Sugiyama, M., Best Paper Award of NeurIPS 2020 Workshop: Deep Learning through Information Geometry

・ガラムカリ　和，杉山　麿人，人工知能学会全国大会優秀賞，2021