科研費

科学研究費助成事業 研究成果報告書

平成 30 年 6 月 4 日現在

機関番号: 12608

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: 25120013

研究課題名(和文)大規模なスパースモデリングへの統計力学的アプローチ

研究課題名(英文)Statistical mechanics approach to massive sparse modeling

研究代表者

樺島 祥介 (KABASHIMA, Yoshiyuki)

東京工業大学・情報理工学院・教授

研究者番号:80260652

交付決定額(研究期間全体):(直接経費) 41,300,000円

研究成果の概要(和文):スパースな構造を抽出する統計モデルによる多変量解析の方法(=スパースモデリング)について系統的な開発/整備を行った.統計モデルを利用したデータからの情報抽出は尤度や相互情報量など情報量概念に関する最適化問題として系統的に定式化することができる.しかしながら,こうした系統的定式化に沿った方法は,多くの場合,計算量的困難を伴うため具体的に実施することが難しい.本研究では,この困難を統計力学の概念/技術を用いて克服することで,現状の打破を目指した.特に,1)圧縮センシング,2)潜在変数モデリング,3)モデル選択の課題に関し,多数の事例に対して性能解析,実際的アルゴリズム開発を行った.

研究成果の概要(英文): The objective of this project is to develop and systemize methods of multivariate statistics that extract sparse structures lying behind observed data, utilizing the notion of information quantities. In general, one can systematically formulate the statistical-model-based information extraction as optimization problems concerning the information quantities. However, such methods are often computationally difficult to perform in practice. In this project, we intended to practically overcome this difficulty by employing methods and notions of statistical mechanics. By analyzing various concrete models, we aimed to construct a methodology for "systematic" and "practically performable" sparse modeling. Our achievements include performance analysis and algorithm development for various problems that arise in 1) compressive sensing, 2) latent variable models, and 3) issues of model selection.

研究分野: 統計力学,情報理論,機械学習

キーワード: 圧縮センシング 潜在変数モデリング モデル選択 レプリカ法 平均場近似 特異モデル

1.研究開始当初の背景

(1) 学術的背景

計測技術の革新による高次元大量データの 出現により,科学的発見をデータ駆動で行う 「データ中心科学」への期待が高まっている 高次元データから情報抽出を行う方法論は 一般に多変量解析とよばれる.こうした方法 は系統的に定式化できる一方で,計算量的困 難を伴うため具体的に実施することが一般 に難しい.その結果,具体的な実施のための アルゴリズム開発のみならず,そうした方法 に本来備わる分析能力や限界についても議 論の余地が多数残されている.

(2) 国内・国外の研究動向及び位置づけ 多変量解析の従来法は現実的計算量での処 理の要請から、そのほとんどが"多次元正規分 布"を生成モデルとして仮定している.しかし ながら,現実のデータは必ずしも正規性を満 たさない.正規性を補完する概念として国内 外の研究で近年注目されているのが「スパー ス性」である.ほぼすべての自然科学は,観 測データは背後にある簡潔な法則から生成 されている、という考え方を共有している、 スパースモデリングとは、この考え方にもと づき,データの生成モデルに簡潔性(=スパ ース性)を課すことで「観測データ→簡潔な 法則」を半自動的に推定する枠組みであるが, 内在する計算量的困難が情報量概念にもと づく系統的方法の活用を阻んでいる.

2.研究の目的

上記の認識のもと,本研究では大自由度性に起因する計算困難の問題を大自由度結合システムの取り扱いに長けている統計力学の概念/技術を用いて克服し,具体的な事例を多数検討することを通じて,「システマティック」かつ「実践できる」方法論の構築を目指した.具体的には,研究開始以降5年間の期間内に大きな進展の見込まれる以下の3つの課題に取り組んだ.

(1) 圧縮センシング

スパース性を利用して,少数の観測値からの信号復元を可能にする圧縮センシングはMRIやNMRなど物理的制約のため観測コストの高い問題に有用である.しかしながら,スパースな推定は一般に計算コストが高い.統計力学の平均場理論にもとづき低計算量,かつ,推定精度の高いスパース推定アルゴリズムを開発する.さらに事例研究を通じてその可能性と限界を解明する.

(2) 潜在変数モデリング

圧縮センシングでは,スパース表現を可能にする基底が既知であることを前提としているが,より一般には観測データからこうした基底(=簡潔な法則)を求めることが重要になる.このような課題は「潜在変数モデリング」として系統的に定式化できる.ただし,潜在変数モデリングは一般に最適化すべき

目的関数が推定パラメータに複雑に依存するため実施が難しい、統計力学のレプリカ理論,平均場理論や確率推論研究の知見にもとづいてこの問題の解決をはかる.

(3) モデル選択

実データを解析する際には必ずデータの生成過程に関する何らかのモデル化を行わなければならないが、その結果はモデルごとに異なるため何らかの指標を用いて適切なモデルを選択することが必要になる.こうした指標としては赤池情報量規準(AIC)が有名であるが、スパースモデリングではスパース性に伴う離散性、不連続性に由来して既存指標の適用がしばしば適切でない.そこで、それらに代わる指標を構成し、適切なモデル選択のための実際的アルゴリズムを開発する.

3.研究の方法

(1) 圧縮センシング

圧縮センシングは,解がスパース(多くの成分がゼロである)という仮定の下,劣決定方定式 y = Ax($x \in \mathbb{R}^N$, $y \in \mathbb{R}^M$, $A \in \mathbb{R}^{M \times N}$, M < N を x について解く問題として,定式化される.我々は先行研究において,観測行列で与えられる場合に,Lp ノルム最小化min $\{ \|x\|^p \}$ subj. to y = Ax によって正しいスパース解を探索できる条件を統計力学の上の先行研究では,同様の条件の下で,効率的に解を探索するアルゴリズムの開発を決ける.これら圧縮センシングに関する性能の形成で、対対の場合、ノイズの存在する場合に広げていく.

(2) 潜在変数モデリング

圧縮センシングでは,観測される信号 y をス パースなベクトルxによって表現する行列 Aが既知であることを前提としている.しか しながら,一般のデータ解析では,そうした スパース表現を可能にする行列は既知とは 限らず,観測されたデータにもとづいて発見 する必要がある.この問題はデータ $Y \in \mathbb{R}^{M \times P}$ が与えられた際,スパースな行列 $X \in \mathbb{R}^{N \times P}$ をもちいてY = AXと表現しなお すことができる辞書行列 $A \in \mathbb{R}^{M \times N}$ を見つ ける問題として定式化できる.この問題はし ばしば辞書学習と呼ばれるが,辞書学習がど のような条件であれば成功するのかは自明 ではない、この条件を統計力学のレプリカ法 をもちいて分析する.また,辞書学習のように, 観測データYを ,規則性を表す行列Aと何ら かの特徴を持った表現Xに分解する問題は 一般に行列分解問題と呼ばれるが,因子分析, 低ランク行列近似,ロバスト PCA,クラスタ リングなど多くの潜在変数モデリングがこ の形式に定式化される.しかしながら,行列

分解に関する統一的な理論的知見は十分ではない.この問題を解決するために,ベイズ統計の枠組みに沿って推定問題を定式化し,統計力学のレプリカ法および平均場近似の方法をもちいて,ベイズ法によって達成される限界性能の特徴づけ,および,近似的行列分解アルゴリズムの開発を行う.

(3) モデル選択

AIC は最尤推定によって求めたパラメータ が有する予測誤差 (新しいデータに対して評 価される負の対数尤度の期待値)を手元のデ ータにもとづいて解析的に評価したもので ある.ただし,その際,最尤推定量が漸近的 に正規分布にしたがうことを仮定している ため,推定量が正規分布にしたがうことが必 ずしも保証されないスパースモデリングで は根拠を失う場合がある.この問題を解決す る方法として広く使える情報量規準 (widely applicable information criterion: WAIC) が 近年提案されている.WAICでは推定量の漸 近正規性が保証される正則モデルに対して は AIC と等価な結果を与え,正規性の保証が ないモデルに対しても予測誤差に関する数 学的な根拠のある推定量を与えている.この 規準の有効性をスパースモデリングに対し ても実験的に確認する.より汎用的なモデル 選択の方法としては、手元のデータにもとづ いて予測誤差を数値的に評価する交差検証 法が知られている.ただし,交差検証法は訓 練データにもとづく推定とテストデータに 対する誤差評価を訓練/テストデータの役 割を変えながら,何度も繰り返す必要があり, 計算量の高さが利用に対するボトルネック となっている.この問題を実際的に解決する ため,物理学で多用される摂動法にもとづい て交差検証を実際に行うことなしに交差検 証誤差を近似的に評価する公式を開発する.

4. 研究成果

以下,3つの課題に対して得られた代表的な成果について記す.なお,文中の[論文]等は「5.主な発表論文等」の中での論文番号を表している.

(1) 圧縮センシング

直交観測行列

 行間で直交したランダム行列は要素間の独立性を仮定した標準的なランダム行列よりも信号復元精度を高めること,また,原信号の信号密度がいくつかのブロックにわかれてばらつきがある場合,それらのブロック構造に対応するようにランダム直交行列を連接させて観測行列を構成することで信号復元精度が向上すること,などを明らかにした.

要素間の独立性を仮定したランダム観測行列に関しては,近似的メッセージ伝搬(AMP)が高い信号復元性能を与えることが示されている.ただし,直交行列にもとづいた観測行列では,AMPの信号復元精度は必ずしも最良ではない.我々は,期待値整合近似(expectation consistent (EC) approximation)と呼ばれる平均場近似にもとづいてランダム直交行列に対し,AMPよりも優れた復元精度を与える近似的復元法を構成した[論文

1ビット圧縮センシング

標準的な圧縮センシングの理論では、観 測信号 v が任意の精度で採取できること 前提としている.しかしながら,一般に は,離散化による丸め誤差などにより 任意精度の実数を採取することはできな い.また,リモートセンシングなどでは, 観測を行うことに対するコストよりも、 得られた情報を遠隔地に伝達する際の通 信コストが技術的な律速となる場合もあ る.そういった状況を想定し,各観測値 の符号 $y = \operatorname{sgn}(Ax)$ のみからの信号復 元を試みる方法が Boufounos and Baraniuk (2008)により提案されている. ただし,実ベクトルに作用する $sgn(\cdot)$ は 成分毎にその符号を返す関数で符号関数 である .各観測時で土1の符号値に対応す る1ビットの情報にまで観測信号を圧縮 することから,この方法はしばしば1ビ ット圧縮センシングとよばれる.

Boufounos and Baraniuk (2008)では標準的な圧縮センシングに対し規定追跡法として知られている信号復元法にならい,制約条件 $y=\operatorname{sgn}(Ax)$,||x||,=N の下で, L_1 ノルム||x|| をを最小化することにより信号を復元する方法が提案れている.しかしながら,この方法には2つのチューニングパラメータが含まれていること,また,収束のために多数の反復を要することなど改善すべき点がまだ残されている.

我々は、予備的研究において、上記の最適化に統計力学のキャビティ法を適用することで先行研究のものとは異なる1ビット圧縮センシングの復元アルゴリズムを開発した、我々のアルゴリズムにはi)チューニングパラメータが1つのみである、ii)収束に要する反

復回数が少ない,iii)原信号の非ゼロ成分が比較的多い場合に,より高い精度の信号復元性能を示す,点で上記の方法よりも優位性がある.我々はさらに,1ビット圧縮センシングに関して,ベイズ推論方式に対する性能分析による性能限界の解明[論文],1ビット化に際し閾値を導入することで振幅情報の欠落を防ぐ方法の提案などの成果[論文]も得ている.

(2) 潜在変数モデリング

辞書学習のサンプル複雑度評価

圧縮センシングでは、原情報がスパースに表現される基底が既知である状況を想定している。しかしながら、一般には、対象となる信号をスパースに表現する基底が分かっているとは限らない。そういった場合には、データから同定する必要がある。この問題は辞書学習(dictionary learning)、あるいは、基底学習(basis learning)などとよばれる。

辞書学習を理論的に考察するために、以下のような状況を想定する、 $y \in \mathbb{R}^M$ はある基底 $A^0 = (a_i^0) \in \mathbb{R}^{M \times N}$ により スパースな表現 $x \in \mathbb{R}^N$ をもちいて $v = A^0 x$ と表現されるとする. ただし, x の各成分 x_i (i=1,2,...,N) は分布 $\phi(x) = (1 - \rho)\delta(x) + \rho f(x)$ $(0 < \rho < 1$, f(x) は x = 0 に確率質量 を持たない密度関数)から独立に生成さ れているとする.このようなyのサンプ ルをP個集めてできる行列 $Y \in \mathbb{R}^{M \times P}$ が与えられたとき,対称性による重複を 除いて A^0 同定するために必要な最小の サンプル数 P はしばしばサンプル複雑 度 (sample complexity) とよばれる. 我々は基底Aの各カラムが長さ1に規 格化され、スパース行列Xの各カラムの 非ゼロ要素数が $N\theta$ ($0 \le \theta \le 1$)以下で ある,という制約の下,自乗誤差 $||Y-AX||_{L}^{2}$ を最小化する最小自乗学習 に対するサンプル複雑度を, $\alpha=M/N$, $\gamma = P/N$ をともにO(1)に保ちながら $N \to \infty$ とした極限において,統計力学

最小自乗学習は素朴であるものの,辞書学習として最適な学習法であるとは限らない.辞書学習の原理的限界を見極めるため,さらに我々は,真の辞書とスパース表現 A^0,X^0 が既知の事前分布からそれぞれ生成され, $Y=A^0X^0$ が与えられたとき,ベイズ推論によって辞書学

のレプリカ法をもちいて評価した.その

結果 $\theta \geq \rho$ であれば a, r, q によって

定まる臨界値 g_F よりもgが大きければ,

 A^0 を同定することが典型的に可能であ

ることを示す結果を得た[論文].この ことは,先行研究で知られている上界

 $O(N \ln N)$ には改善の余地があること

を示唆している.

習を行った場合のサンプル複雑度を評価した。その結果, $\gamma > \alpha/(\alpha - \rho)$ であれば A^0 を典型的に同定できるという結論を得た[論文].ベイズ推論の最適性から,この結果は,辞書学習に関するサンプル複雑度の原理的な下限を意味している。

行列分解の最適性能評価と近似的メッセージ伝搬

上述の辞書学習問題は,2つの行列の積 が与えられた場合に,与えられた制約を 満たすように2つの行列に分解する行 列分解問題と捉えることができる.同様 の形式で記述される問題は,ブラインド 行列較正,低ランク行列補完,スパース PCA, ブラインド信号源分離, ロバスト PCA, 因子分解, 非負値行列分解, クラ スタリングなど多変量解析におけるモ デル化技法に数多く見られる.これらは すべて,観測されるデータ(顕在変数) が隠れた変数(潜在変数)が原因で生成 されていると仮定し,生成過程の法則性 を表す行列と隠れ変数の両方を観測デ ータから推定しようとするモデル化の 方法とみなすことができる.こうした方 法はしばしば潜在変数モデル(latent variable model) とよばれる.

我々は,上述のベイズ推論の定式化による辞書学習の性能解析を拡張し,これら多くの行列分解問題の性能限界を統一的に解明する枠組みを与えた.また,この論文では,圧縮センシングに対して広く利用されているAMPを行列分解問題に一般化したアルゴリズムも導出している[論文].

(3) モデル選択

WAICに関する有効性の検証

ベイズ推論では,多くの場合,事前分布 や尤度関数を定めるハイパーパラメータ をデータから決定しなければならない. ハイパーパラメータを決定する際の規準 量としては,赤池情報量規準(AIC),最 小記述長 (MDL), 周辺尤度など, さま ざまな提案がこれまでなされている.し かしながら,その多くはデータ数が十分 に多い漸近領域において,推定されたパ ラメータが正規分布にしたがう正則モデ ルにもとづいて理論が構築されている. 残念ながら,スパースモデリングは,漸 近領域においてもパラメータの推定値が 正規分布にしたがわない特異モデルで記 述されることが多く、モデル選択規準に 関して再検討する必要がある.

この問題に対し、分担者の渡辺はベイズ予測分布に関する情報量の期待値(ベイズ予測損失)をモデルの良し悪しを定める尺度とした上で、その推定値をWAICとよび、モデル選択に利用するこ

とを提案している.WAIC とベイズ予測損失との差はデータ数nに対して $O(n^{-2})$ となることが数学的に保証されるため漸近領域においてその差は無視できる.また,WAIC はモデルが正則か否かによらず,データセットのみから数値的に評価できる量であるため広いモデルクラスに対して適用可能である.

論文 では,ラプラス事前分布 $\phi(w_i) \propto e^{-\lambda |w_i|}$ を用いたスパース線形回帰問題を取り上げ,WAIC がベイズ予測損失に関する精度の高い推定量として機能し,モデル選択規準の役割を果たすことを示した.

近似的交差検証誤差公式

交差検証法 (cross validation: CV)では 未知のデータに関する汎化能力(予測能 力とも呼ばれる)の高さをモデルの良さ を評価する際の規準とする.残念ながら 手元にあるデータのみから汎化能力を直 接知ることはできない、そこで、手元に あるデータセットDをパラメータ推定 に用いるための訓練セット D^{train} と汎化 能力を評価するためのテストセット $D^{ ext{test}}$ に分割し, $D^{ ext{train}}$ にもとづいて推定 された推定値が有する $D^{ ext{test}}$ への汎化能 力の高さによって,モデル(あるいは, モデルを特徴づけるハイパーパラメータ) の良し悪しを判断する.ただし,訓練セ ットとテストセットへの分割法には任意 性がある.そこで,Dをk個のサブセッ トに分割し,1つのサブセットを $D^{ ext{test}}$ 残りのk-1個のサブセットの和集合を D^{train} として,汎化能力をk通りの D^{test} の取り方に対し平均化して評価するk-fold CV が広く用いられている .k = |D| , すなわち,テストセットの大きさを1と した交差検証法は leave-one-out(LOO) CV ともよばれる.

CV は任意の推定法, 汎化誤差に関する任意の尺度に対して適用できるという意味で極めて汎用性の高い{モデル選択}の方法である.一方で, テストセットへの評価を行う毎に推定量の解き直しが必要となるため,計算量的なコストが高い.

以上の認識の下,我々はスパースモデリングの代表的な手法である L_i 正則化項付き線形回帰(LASSO)の LOO CV に関し,解き直しをすることなく,全データセット D に対する推定結果にもとづいて半解析的に LOO CV 誤差の近似値を求める公式を開発した[論文]. 少数データからの画像再構成でもちいられる全変動(total variation: TV)正則化[論文]や L_i -正則化項つき多項ロジスティック回帰に対しても開発した.

(研究代表者、研究分担者及び連携研究者に は下線)

[雑誌論文](計 50件)

T. Obuchi, S. Ikeda, K. Akiyama and Y. Kabashima: Accelerating cross-validation with total variation and its application to super-resolution imaging, PLoS ONE, 12, 12, e0188012(1-14) (2017) 査読有り

doi: 10.1371/journal.pone.0188012

Y. Kabashima, T. Obuchi and M. Uemura: Approximate cross-validation formula for Bayesian linear regression, in Proc. Allerton2016, pp. 596--600 (2016) 査読有り

doi:10.1109/ALLERTON.2016.7852286

Y. Xu and <u>Y. Kabashima</u>: Statistical mechanics analysis of thresholding 1-bit compressed sensing, J. Stat. Mech.: Theor. and Exp. Vol. 2016, pp. 083405(1--16) (2016) 査読有り

doi:10.1088/1742-5468/2016/08/083405

Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata and L. Zdeborová: Phase transitions and sample complexity in Bayes-optimal matrix factorization, IEEE Trans. on Inf. Theory, Vol. 62, pp. 4228-4265 (2016) 查読有り

doi: 10.1109/TIT.2016.2556702

M. Vehkaperä, <u>Y. Kabashima</u> and S. Chatterjee: Analysis of regularized LS reconstruction and random matrix ensembles in compressed sensing, IEEE Trans. on Inf. Theory, Vol. 62, No. 4, pp. 2100--2124 (2016) 査読有り

doi: 10.1109/TIT.2016.2525824

T. Obuchi and Y. Kabashima: Cross validation in LASSO and its acceleration, J. Stat. Mech.: Theor. and Exp., Vol. 2016, pp. 053304(1--36) (2016) 査読有り doi:10.1088/1742-5468/2016/05/053304

宮崎大,<u>渡辺澄夫</u>:情報量規準 WAIC を用いた LASSO 学習の最適化と都市 データ解析への応,信学技法, IEICE-114, No. 514, pp. 331--335 (2015) 査読無し https://ci.nii.ac.jp/naid/110010022072/

Y. Kabashima and M. Vehkaperä: Signal recovery using expectation consistent approximation for linear observations, in Proceedings of ISIT2014, pp. 226--230 (2014) 香読有り

doi: 10.1109/ISIT.2014.6874828

Y. Xu, <u>Y. Kabashima</u> and L. Zdeborová: Bayesian signal reconstruction for 1-bit compressed sensing, J. Stat. Mech.: Theor. and Exp. Vol. 2014, pp. P11015(1--23) (2014) 音読有り

doi:10.1088/1742-5468/2014/11/P11015

A. Sakata and Y. Kabashima:
Statistical mechanics of dictionary learning, EPL, Vol. 103, pp. 28008(1--6) (2013)
査読有り

doi: 10.1209/0295-5075/103/28008

[学会発表](計54件)

Yoshiyuki Kabashima, "Statistical mechanics approach to compressed sensing", Workshop on "Theoretical Comp. Sci. and Statistical Physics methods in Machine Learning" at International Center for Theoretical Sciences (ICTS-TIFR), Bengaluru, India, 2017

<u>Yoshiyuki Kabashima</u>, "Approximate cross-validation formula for Bayesian linear regression", the 54th Annual Allerton Conference 2016, Allerton, USA, 2016

Yoshiyuki Kabashima, "Overview of compressed sensing", the school & workshop "Spin Glass and Beyond: An old tool for new problems", Cargese, Corsica, France, 2014

Sumio Watanabe, "Discovery Phenomenon and Information Criteria", The Seventh Workshop on Information Theoretic Methods in Science and Engineering, Sheraton Waikiki Hotel, Honolulu, USA, 2014

<u>Sumio Watanabe</u>, "Model Selection for Non-Regular Statistical Models", 2013 Joint Statistical Meetings, Palais des congres de Montreal, Canada, 2013

[図書](計 2件)

渡辺澄夫(pp. 1--47), 永尾太郎, <u>樺島祥介(pp. 79--112)</u>, 田中利幸, 中島伸一, 「ランダム行列の数理と科学」(森北出版) 2014 年 4 月

鈴木譲,植野真臣,黒木学,清水昌平, 湊真一,石畠正和,<u>樺島祥介(pp.</u> 167-193),田中和之,本村陽一,玉田 嘉紀,「確率的グラフィカルモデル」(共 立出版)2016年7月

〔産業財産権〕

○取得状況(計 2件)

名称:信号処理システム及び信号処理方法 発明者: <u>竹田晃人</u>, <u>樺島祥介</u>, 李斗煥, 山田貴

之, 上原一浩 , 赤羽和徳.

権利者:日本電信電話株式会社,国立大学法

人東京工業大学 種類:特許

番号:特許 5761811

取得年月日:2015年6月19日

国内外の別:国内

名称:信号処理システム及び信号処理方法 発明者:<u>樺島祥介</u>, <u>竹田晃人</u>, 李斗煥, 山田貴

之, 上原一浩, 赤羽和徳.

権利者:日本電信電話株式会社,国立大学法

人東京工業大学 種類:特許

番号:特許 5761812

取得年月日: 2015年6月19日

国内外の別:国内

6. 研究組織

(1)研究代表者

樺島 祥介 (KABASHIMA, Yoshiyuki) 東京工業大学・情報理工学院・教授 研究者番号:80260652

(2)研究分担者

渡辺 澄夫 (WATANABE, Sumio) 東京工業大学・情報理工学院・教授 研究者番号:80273118

竹田 晃人 (TAKEDA, Koujin) 茨城大学・工学部・准教授 研究者番号:70397040

坂田 綾香 (SAKATA, Ayaka) 統計数理研究所・モデリング研究系・助教 研究者番号:80733071

(3)連携研究者

小渕 智之 (OBUCHI, Tomoyuki) 東京工業大学・情報理工学院・助教 研究者番号:40588448