科学研究費補助金新学術領域研究「スパースモデリングの深化と高次元データ駆動科学の創成」
2017 年度チュートリアル

スパースモデリングとデータ駆動科学

本武 陽一[A], 五十嵐 康彦[A,B,C], 市川 寛子[D], 竹中 光[A], 桑谷 立[B,E], 永田 賢二[B,F],
中西 (大野) 義典[B,G], 赤井 一郎[H], 岡田 真人[A,C 1]
A. 東京大学 大学院新領域創成科学研究科, B. 科学技術振興機構 さきがけ,
C. 物質・材料研究機構 統合型材料開発・情報基盤部門, D. 東京理科大学 理工学部,
E. 海洋研究開発機構 地球内部物質循環研究分野, F. 産業技術総合研究所 人工知能研究センター,
G. 東京大学 大学院総合文化研究科, H. 熊本大学 パルスパワー科学研究所

1 はじめに

今年度で，新学術領域研究「スパースモデリングの深化と高次元データ駆動科学の創成（略称 疎性モデリング，領域代表者 岡田 真人，2013〜2017 年度)」の研究活動が終了する．計測データに隠された規則性を抽出するデータ解析の系統的技術の開発は，来るべき「データ科学時代」における全ての科学分野に共通する喫緊の課題である．本領域では，多くの自然科学分野の計測データに普遍的にスパース性が存在することを基本原理としたスパースモデリング (SpM: Sparse Modeling) に注目し，生命分子からブラックホールに至る，幅広い自然科学分野の実験・計測研究者と情報科学者の連携により，この課題の解決を目指した．本チュートリアルでは，本新学術領域の目的であるデータ駆動科学の創成と，スパースモデリングの深化について，著者の一人の岡田が代表をつとめる計画研究 B01-2 スパースモデリング班の研究成果を中心として述べる．

次節 §2 では，本新学術領域の先行プロジェクトである，特定領域研究「確率的情報処理への統計力学的アプローチ（略称 SMAPIP，領域代表者 田中和之，2002〜2005 年度)」と特定領域研究「情報統計力学の深化と展開（略称 DEX-SMI，領域代表者 樺島祥介，2006〜2009 年度)」の情報数理基盤である情報統計力学について紹介する．Kabashima らは情報統計力学の枠組みを用いて，スパースモデリングの一種である圧縮センシングを理論的に取り扱った [1]．磁性体などの物質の性質を説明する理論的な枠組みであった統計力学は，情報統計力学により計測科学をも取り扱えるようになった.「計測はマザーオブサイエンス」とも言われており，スパースモデリングを取り扱うことで，情報統計力学は科学全体を射程内に収めたわけである．これが新学術 SpM と先行研究の関連である．

§3 では，データ駆動科学について述べる．データ駆動科学とは，データ科学の一分野であり，実験/計測/計算データの背後にある潜在的構造の抽出に関して，データが対象とする分野に依存しない普遍的な学問体系である．本新学術領域の活動により，David Marr が指摘した三つのレベルが [2]，データ駆動科学を創成するためにも重要な着眼点となることを確信した．データ駆動科学においては，データ解析の目的である計算理論のレベルと，データ解析手法である表現・アルゴリズムのレベルの間に，モデリングのレベルが存在することに気づいた．これらの三つのレベルを新たに“データ駆動科学の三つのレベル”と名づけ，データ駆動科学の学理の原点に位置付けた [3]．このデータ駆動科学の三つのレベルを考察することにより，データ駆動科学を効率的に推進するためには，普遍性の高い情報処理手法をキーテクノロジーにすることが有効であることがわかる．そのキーテクノロジーのひとつがスパースモデリングである．LASSO に代表される L1 最適化の枠組みのスパースモデリングについては，今回だけでなく過去のチュートリアルでも取り扱われているので [4–6]，§4 はもっとも基礎的な一変数の線形回帰モデル，つまり一次関数をベイズ推論の枠組みで取り扱う．続く §5 では，スパースモデリングの深化の一例として，著者たちの共同研究の成果である，ES-DoS(状態密度付き全状態探索法) の紹介をする [7]．最後に，§6 では，本チュートリアルのまとめと，スパースモデリングとデータ駆動科学の今後の展望を述べる．

2 情報統計力学とスパースモデリング

2.1 ニューラルネットワークとスピン系

情報統計力学とは，ベイズ推論と統計力学の数理的等価性に基づき，情報科学の幅広い分野を統計力学的手法で取り扱う枠組みである [8]．情報統計力学では，ランダム相互作用を持つスピン系を理論的に取り扱

[1] E-mail: okada@edu.k.u-tokyo.ac.jp

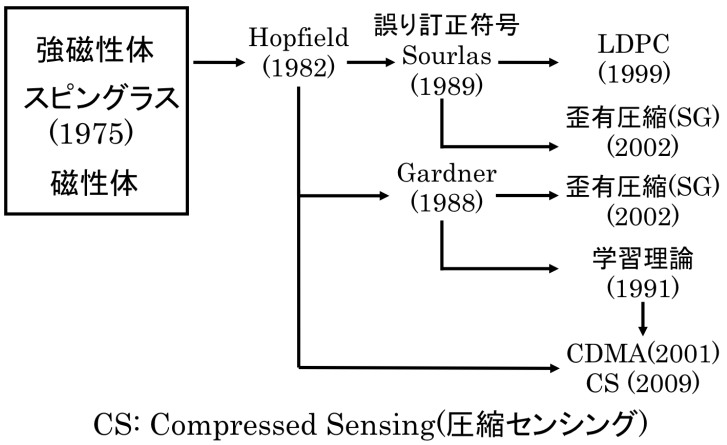


Figure 1: 情報統計力学の体系

う手法であるレプリカ法 [9,10] をもちいて，以下に説明するような情報科学的な対象の性質を解析的に取り扱うことができる [11].

現在の人工知能 (AI) ブームをニューラルネットワーク (神経回路モデル) の研究から見ると，今は第 3 次ニューラルネットワークブームと考えることができる．情報統計力学は，今から 30 年前の第 2 次ニューラルネットワークブームの中に生まれた．第 2 次ニューラルネットワークブームのきっかけのひとつは，物理学者の Hopfield による連想記憶モデルの Hopfield モデルの提案である [12]．Hopfield は 1982 年に"Neural networks and physical systems with emergent collective computational abilities." という魅力的な題名の論文を書いた．この論文で Hopfield は，スピングラスに代表されるランダムスピン系と，脳の神経回路網 (ニューラルネットワーク) とが深く関係していると主張した．Hopfield モデルの提案を契機に，多くの統計物理学者がニューラルネットワークの研究に参入した．情報統計力学は Hopfield モデルによって始まったといっても過言ではない．なお通常 Hopfield モデルと呼ばれているこのモデルは，Hopfield の提案より 5 年前の 1977 年に Amari が提案しており [13]，Amit はこのモデルを Amari-Hopfield モデルと呼ぶべきであると述べている．

Amari-Hopfield モデルは，物性物理学における磁石のモデルと数理的な対応がある．Ising スピン s は量子力学のスピンを簡略に表現したものであり，± 1 の値をとる．スピンが上を向いているアップの状態が $+1$ に対応し，ダウンの状態が -1 に対応する．我々の大脳皮質には百億以上の神経細胞がある．電位の状態により，神経細胞は発火・非発火の二状態をとる．これらの二状態を Ising スピンの二状態に対応させることにより，神経細胞のネットワーク (ニューラルネットワーク) を Ising スピン系として取り扱うことが可能になる．さらに，デジタル情報処理は 0 と 1 の二値で情報をあらわす．ビットの 0-1 と Ising スピンの ± 1 を対応させることで，情報科学の多くの対象を Ising スピン系として取り扱うことができる．これが情報統計力学の基本的な戦略である．この戦略をスパースモデリングに用いたものが，§5 で述べる ES-DoS(状態密度付き全状態探索法) である．ES-DoS では説明変数を使う使わないの二状態を Ising スピンで表現して，スパースモデリングを行う．

図 1 は，情報統計力学で取り扱う内容を 1 枚の絵にしたものである．ここでは，図 1 にもとづき，情報統計力学がどのように発展して，スパースモデリングにつながるかを説明する．N 個の Ising スピンからなる系を考える．i 番目の Ising スピン s_i は ± 1 の値をとるものとする．スピンアップの状態が $+1$ に対応し，ダウンの状態が -1 に対応する．つぎに，以下のようなエネルギー $E(\boldsymbol{s})$ を導入する，

$$E(\boldsymbol{s}) = -\sum_{(i,j)} J_{ij} s_i s_j - \sum_{i=1}^{N} h_i^0 s_i \tag{1}$$

ここで，J_{ij} は i 番目のスピンと j 番目のスピンの相互作用であり，相互作用は対称であるとし $J_{ij} = J_{ji}$ とする．また (i,j) に関する和は i と j に関するすべての組に関する総和を意味する．h_i^0 は i 番目のスピンへの磁場である．エネルギー $E(\boldsymbol{s})$ は式 (2) の非同期状態更新の際に非増加であることを示すことができる．

$$s_i = \mathrm{sgn}(h_i) = \mathrm{sgn}(\sum_{j \neq i}^{N} J_{ij} s_j + h_i^0). \tag{2}$$

ここで sgn(·) は符号関数であり，引数の符号に対応させて +1 または −1 を返す関数である [11]．式 (2) から明らかなように，この系はニューロンの結合荷重が J_{ij} で，入力または閾値が h_i^0 のニューラルネットワークである．

2.2 強磁性体の伏見-Temperly モデル

図 1 の強磁性体 (磁石) の伏見-Temperly(FT) モデルを説明する [14]．FT モデルでは，スピンへの磁場は一様であるとし，$h_i^0 = h^0$ とする．スピンは残りの $N-1$ 個のスピンと相互作用するとし，その大きさは一様であるとする．式 (1) の右辺の第一項の相互作用の大きさを $O(N)$ にするために，相互作用の大きさを $O(1/N)$ とする．これらの条件から $J_{ij} = J_0/N$ とする．まとめると FT モデルでは以下のようにおき，

$$J_{ij} = \frac{J_0}{N}, \quad h_i^0 = h^0, \tag{3}$$

モデルの相互作用と磁場に関して i 依存性がない．FT モデルにおいては，Ising スピン s_i が全て同じ方向を向く時，つまり $\sigma_i\sigma_j = 1$ の場合、式 (1) のエネルギー $E(\boldsymbol{s})$ は最小値を取る．つまり，FT モデルではスピンは揃うのである．

2.3 スピングラスの SK モデル

次にスピン間の相互作用に乱雑さが存在する場合を議論しよう．スピン間の相互作用に乱雑さが存在する系をスピングラスという．ここでは Sherrington と Kirkpatrick によって提案された SK モデルを説明する [10] SK モデルでは，相互作用 J_{ij} が平均 J_0/N 分散 J^2/N のガウス分布に従う，

$$J_{ij} = \frac{J_0}{N} + \frac{J}{\sqrt{N}} z_{ij}, \quad z_{ij} \sim \mathcal{N}(0,1). \tag{4}$$

ここで $\mathcal{N}(0,1)$. は平均 0 分散 1 のガウス分布をあらわす．$J = 0$ が FT モデルに対応している．この相互作用の形を見ただけで，スピングラスの解析がいかに難しいかがわかる．このような状況で，レプリカという概念を導入することにより，この系に存在する対称性を鋭く見抜いたのが Edwards と Anderson である [9]．Sherrington と Kirkpatrick はレプリカ法をもちいることで，SK モデルを解析的に取り扱うことに成功した [10].

2.4 Amari-Hopfield モデル

Amari-Hopfield モデルでは，p 個の記憶パターンを Ising スピン系の平衡状態にするようにスピンの相互作用 J_{ij} を決める [12,15,16]．μ 番目の記憶パターン $\boldsymbol{\xi}^\mu$ は，各要素が +1 または −1 をとる N 次元ベクトルである．μ 番目の記憶パターン $\boldsymbol{\xi}^\mu$ の第 i 番目の成分 ξ_i^μ を以下の確率で独立に決める．この記憶パターンをもちいて式 (1) のスピン間の相互作用 J_{ij} を以下のように決める，

$$\mathrm{Prob}[\xi_i^\mu = \pm 1] = \frac{1}{2}, \quad \mu = 1, \cdots, p \tag{5}$$

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \tag{6}$$

ここで p はモデルに憶えさせる記憶パターンの数である．

このモデルは §2.2 で述べた強磁性体の FT モデルの拡張になっている．パターン数を $p = 1$ としよう．1 番目の記憶パターン $\boldsymbol{\xi}^1$ を用いて新たなスピン状態 $\tau_i = \xi_i^1 s_i$ を導入する．この変換をゲージ変換と呼ぶ．式 (1) に従ってこの系のエネルギーを書き下し，ゲージ変換を用いると，

$$E(\boldsymbol{s}) = -\frac{1}{2N} \sum_{j \neq i} \xi_i^1 \xi_j^1 s_i s_j = -\frac{1}{2N} \sum_{j \neq i} \tau_i \tau_j, \tag{7}$$

となり，FT モデルと等価になる．また p が N に比べて $O(1)$ のときも，Amari-Hopfiled モデルは FT モデルとほぼ同じ性質を持つことが知られている [17].

Hopfield は計算機シミュレーションをもちいて，記憶パターン数 p が $p \sim 0.15N$ までは記憶パターンを系の定常状態であるが，それ以上であると記憶パターンが不安定化することを示した [12]．ここから，p が N に対して 1 の場合 ($p \sim O(1)$) と，N の場合（$p \sim O(N)$）で，系の挙動が定性的に異なることがわかる．$p \sim O(N)$ の場合の Amari-Hopfiled モデルは，SK モデルと同じ性質を持つ．そこで Amit らは，SK モデルで提案されたレプリカ法をもちいて，記憶容量が $p_C = 0.138N$ であることを示した [18]．これが統計力学が情報科学に進出した第一歩である．

2.5 Sourlas 符号

Sourlas は Amari-Hopfield モデルを参考にして，Sourlas 符号と呼ばれる誤り訂正符号を提案した [19]. Sourlas は $p = 1$ の場合の式 (6) の J_{ij} を送信メッセージを符号化したものだと考え，それがガウス通信路で送られ，以下の J_{ij} が受信されるとした．

$$J_{ij} = \frac{1}{N}\xi_i^1\xi_j^1 + w_{ij}, \quad w_{ij} \sim \mathcal{N}\left(0, \frac{\sigma^2}{N}\right). \tag{8}$$

ここで w_{ij} は平均 0 分散 σ^2/N の独立なガウス分布に従うとする．この J_{ij} を使って Sourlas はベイズ推論に基づき，元信号 $\boldsymbol{\xi}^\mu$ の復号に，式 (8) の J_{ij} を式 (1) に代入した系を用いれば良いことを示した．ゲージ変換すると，式 (8) の J_{ij} は，

$$J_{ij} \sim \mathcal{N}\left(\frac{1}{N}, \frac{\sigma^2}{N}\right), \tag{9}$$

で示される平均 $1/N$ 分散 σ^2/N の独立なガウス分布に従う．この系は SK モデルと等価である．

2.6 情報統計力学の深化と展開

図 1 は，情報統計力学で取り扱う内容を 1 枚の絵にしたものである．これで全てとは言えないが，情報統計力学で取り扱う内容の関係性をほぼ描けていると思う．たとえば，物理学会の情報統計力学のセッションにこの図を持っていけば，聞いている講演内容が情報統計力学全体の中でどのように位置づけられるかがわかるはずである．

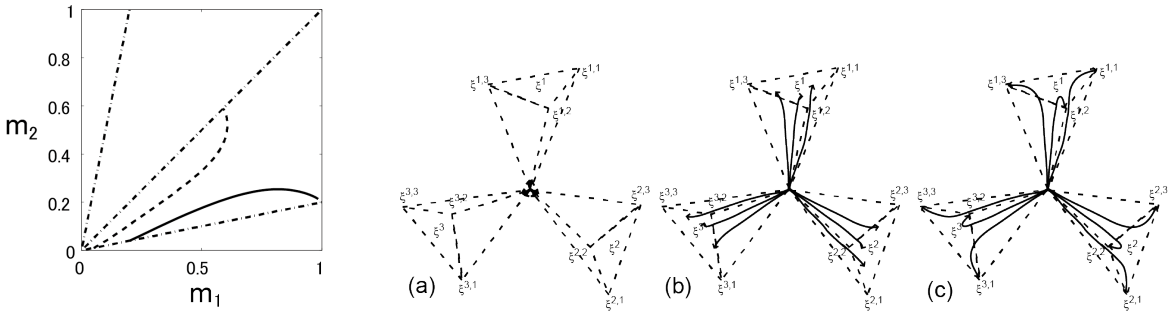
情報統計力学の起点は，強磁性体の伏見-Temperly モデル [14] と Sherrington-Kirkpatrick モデル [10] である．これらのモデルの解析手法がすべてのモデルに対する基礎になる．Hopfield による提案がスピン系と情報科学の出会いのきっかけである [12]．Sourlas の提案により Amari-Hopfield モデル [13] [12] は誤り訂正符号に発展し [19]，それは情報理論の最先端である低密度パリティ検査符号 (LDPC) を取り扱うところまで発展した [20,21]．Murayama と Okada は，誤り訂正符号と歪有データ圧縮の双対性から Kabashima らによる希釈 Sourlas 符号の知見 [20] をもちいて，歪有データ圧縮に関してもレプリカ解析を行った [22].

ここでは，解説しなかったが，ニューラルネットークのパーセプトロンの記憶容量の統計力学的な解析は Gardner によってなされた [23]．この解析結果は Cover らの結果と一致している [24]．パーセプトロンの記憶容量の統計力学的解析は，パーセプトロンを用いた歪ありデータ圧縮の研究に発展している [25,26]. Gardner によってなされたパーセプトロンの記憶容量の計算は，パーセプトロンの学習理論へと発展した [27]. Tanaka は，CDMA のマルチユーザー復調をレプリカ法をもちいて解析した [28,29]．ここでは詳細は述べないが，CDMA は線形パーセプトロンの一種と考えることもできる．また，CDMA は Amari-Hopfield モデルと数理的に似た構造をしている．Tanaka と Okada は，この数理的等価性に注目し，CDMA のマルチユーザー復調のダイナミクスを，Amari-Hopfield モデルの記憶想起のダイナミスの理論を用いて解析している [15,30,31].

スピングラス理論は，スピングラスという一つの物理的対象の理論的枠組みから，学習理論と情報通信理論へと対象を広げた．これが本新学術領域の先行プロジェクトである，特定領域研究「確率的情報処理への統計力学的アプローチ (略称　SMAPIP)」と特定領域研究「情報統計力学の深化と展開（略称　DEX-SMI)」へと繋がっている．

スパースモデリングの先駆は，1980 年代の後半に石川真澄が提案したニューラルネットワークの忘却付き構造学習である [32]．樺島は，この研究に触発されて，いまから 20 年以上前の 1996 年に，レプリカ法を用いて L1 および L2 拘束を導入した線形パーセプトロンの学習理論を構築している [33]．1996 年の Tibshirani の LASSO の提案を経て [34]，2000 年代半ばからは Donoho らが提唱している圧縮センシングが計測工学，通信工学，医用工学など幅広い分野で革新的情報抽出技術として大きな注目を集めている [35]. Kabashima らは，レプリカ法をもちいて，圧縮センシングを理論的に取り扱った [1]．この数理構造は，樺島が 1996 に行った，L1 拘束付きの線形パーセプトロンのレプリカ解析に対応している [33].

圧縮センシングは，センシングの名の通り，計測に関する数理的枠組みの一つである．その対象は，本新学術領域の課題であるブラックホール，MRI，NMR に代表される医学や自然科学など，計測工学が取り扱う全ての対象である．つまり，スピングラス理論にもとづく情報統計力学は，学習理論と情報通信理論を経由して，自然科学全般を含む幅広い対象を取り扱える体系に成長した．これが，SMAPIP と DEX-SMI の後継プロジェクトとしての，本新学術領域スパースモデリングの立ち位置である．なお，情報統計力学によるスパースモデリングの詳細は，竹田によるチュートリアルを参考のこと [36].



(a) (b) (c) (d)

Figure 2: (a) は連想記憶モデルの想起過程を示す．(b) から (d) は想起過程の概略図である．

3 高次元データ駆動科学の創成

3.1 ニューラルネットワークへの実証的アプローチ

§2 で述べたように，統計力学とベイズ推論の数理的等価性を指導原理として，Amari-Hopfield モデルは情報統計力学として発展した．一方，脳の記憶のモデルとして Amari-Hopfield モデルはどのように発展したのであろうか．情報統計力学の華々しい発展に比べて，以下の論点から，それはあまりにも険しい道のように思える．

Amari-Hopfield モデルによる理論的予測を，脳の記憶のモデルとして検証しようとした時，それを，どのように実証していけばよいであろうか．少し考えると，これを直接行なうのは難しいことに気づく．物理や化学等における数理モデルのほとんどは，その数理モデルの構成要素は物理的実体を直接表現している．この場合，モデルの成否を知るには，物理的実態に即した直接的な計測を行ない，それらを比較すれば良い．理科第一分野に属する物理や化学は，第一原理から演繹を行いやすく，この直接計測のパラダイムで著しい進歩を遂げた．では，理科第二分野の生物学や地学ではどうであろうか．分子生物学の導入により，今日では，生物学でも第一原理的な視点による直接計測のパラダイムは有効である．しかしながら，このような第一原理的なアプローチをとりにくい，生物学の分野は数多く存在する．

その一つは，脳のモデルとしてのニューラルネットワークの研究である．ここでは，ニューラルネットワークを構成するニューロン自体が研究の対象であるため，構成要素のニューロンの性質をアドホックに仮定する場合が多い．例えば，ここで紹介した Amari-Hopfield モデルでは，ニューロンの活動を，神経スパイクを出す発火状態と，神経スパイクを出さない非発火の二状態をとる Ising スピンで表現している．ここでは発火非発火の二状態をとる内部変数である膜電位には，時間的な履歴効果はないとしている．これらの性質は，ニューロンの生物物理学的な莫大な知見のほとんどを説明しない．そのような背景のため，AI ブームが日本に押し寄せるついこの間までは，Ising スピンやアナログ素子を用いたニューラルネットワークは，理論神経科学の研究において流行遅れとされていた．そのため，現在では，積分発火型ニューロンと呼ばれる，膜電位の履歴効果を取り入れた二値素子で盛んに神経スパイクが議論されているが，積分発火型ニューロンもまた，ニューロンの生物物理学的な莫大な知見のほとんどを説明しない．しかも，AI ブームの影響で，神経科学の分野でも手のひらを返したように，アナログニューラルネットワークが評価されている [37,38]．第一原理からの演繹が難しい分野では，日々このような迷走が起こっている可能性がある．そのほかにも，タンパク質のフォールディングは，原子論的な定式化の分子動力学で議論できるが，その長時間的な挙動を追う計算を行なうのは難しい場合がある．そのような場合は，タンパク質を弾性体と近似した粗視化モデルが議論されている．この粗視化モデルを実際の計測と結びつけるにはどうしたらよいのであろうか．また地震のバネブロックモデルを，実際の計測と結びつけるのにも同種の問題があるはずである．次に，この種の問題への普遍的なアプローチを示唆する知見を紹介する．

3.2 スパース化による潜在構造抽出 [11]

Amari は Amari-Hopfiled モデルに，相関のある似たパターンを記憶させると，それらの真ん中にある混合状態もアトラクターになること指摘して、それを概念形成と呼んだ [13]．岡田らのグループは，この系の記憶パターンの想起のダイナミクスを理論的に取り扱い，図 2 に示すように，想起の初期は系の状態が混合状態に近づき，その後，記憶パターンが想起されることを示した [16,39,40]．

Matsumoto らは，この理論的な予想を検証するために，Sugase らの電気生理学実験のデータを用い

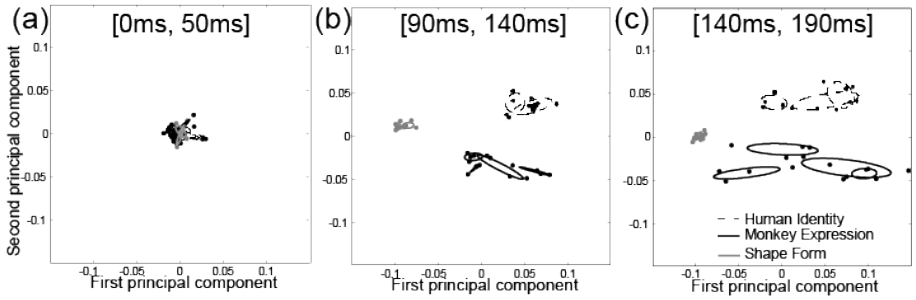


Figure 3: 側頭葉の顔応答細胞の神経集団ダイナミクス．(a) が初期状態，(b) が中間状態，(c) が終状態に対応する．

た [16, 41, 42]．Sugase らは顔画像のセットをサルに提示し，視覚パターン認識の責任領野として考えられている側頭葉の神経細胞の活動を測定した [41]．Sugase らが用いた刺激画像セットは顔画像とそれ以外の画像に分類できる．さらに顔画像のセットはヒトの顔画像とサルの顔画像に分類できる．さらにヒトの顔画像セットはヒト別に分類でき，さらに顔の表情別にも分類できる．サルの顔画像セットもヒトの顔画像セットと同様に分類できる．Sugase らの刺激画像セットは，われわれが外界について感じるような階層構造が埋め込まれているわけである．Sugase らは刺激と神経応答の相互情報量を計算することにより，応答の初期には，顔画像と顔画像以外の分類のような大分類に関する情報が含まれており，応答の初期につづく成分では，ヒト別や表情別のような詳細な分類に関する情報が含まれていることを示した．

Matsumoto らは細胞集団の挙動に対して，主成分分析 (Principal Component Analysis, PCA) を行った [42]. 図 3 は第一主成分と第二主成分で張られる空間で，神経細胞集団の挙動を視たものである．図 3 (a) は神経応答の初期状態を神経集団ベクトルを可視化したものである．神経集団ベクトルは，刺激提示時刻から [0-50msec] の時間幅で求めた発火率から求めた．各点が刺激画像に対応している．初期状態であるので，すべての点は原点に集まっている．図 3 (b) は [90-140msec] の時間幅に対応している．この時間幅は中間状態に対応する．図 3 (a)-(c) を通じて点線の楕円は，個人別にヒトの顔画像をそれぞれまとめて，ヒトごとに点の分布を囲むように書いた．全部でヒト別に関する三つの点線の楕円が存在する．実線の楕円は，表情別にサルの顔画像をそれぞれまとめて，表情ごとに点の分布を囲むように書いた．全部でサルの表情に関する四つの実線の楕円が存在する．灰色の楕円は，形別に顔以外の画像をそれぞれまとめて，形ごとに点の分布を囲むように書いた．全部で形に関する灰色の五つの楕円が存在する．図 3 (c) が最終状態に対応する．図 3 (c) では図 3(b) にくらべて，ヒトのヒト別に関する点線の楕円とサルの表情別に関する実線の楕円とがより分離している．図 3(b) ではそれぞれ一塊であったヒトとサルの顔画像のクラスターが，図 3 (c) でヒト別およびサルの表情別のサブクラスターに分離したわけである．まとめると，Matsumoto らは，神経集団ベクトルが [90-140msec] に対応する発火の中間状態では，ヒトの顔，サルの顔，顔画像以外に対応する三つのクラスターに分離し，[140-190msec] に対応する発火の後半では，それぞれのクラスターがサブクラスターに分かれることを示した [42].

この例から，実験データを PCA などをもちいてスパース化 (低次元化) することで，データの背後にある本質的な構造を抽出することが可能であることが示唆される．このアプローチは，ニューラルネットワークだけでなく，第一原理からの演繹がむずかしい生物学や地学などの理科第二分野全般に適用可能な研究戦略であることがわかる．

3.3 連立方程式とデータ駆動科学の三つのレベル [44]

データ駆動科学とは，データ科学の一分野であり，実験/計測データや数値計算データの背後にある潜在的構造の抽出に関して，データが属する学問分野に依存しない普遍的な学問体系である．同じアルゴリズムがスケールや対象を超えて，有用であることが多いという経験的事実を背景として [43]，その理由を問い，背後にある普遍性から，データ解析自体を学問的対象とする枠組みである [3]．ベイズ推論やスパースモデリングが，その数理科学や統計学的な背景である [44].

対象とする学問に依存しない普遍的な学問体系の構築が可能な理由の一つは，図 4(a) の連立方程式とその応用の三層構造である．連立方程式の文章題の鶴亀算，食塩水の問題，速度の問題は，対象としては全く異なる．これらの異なった対象が，連立方程式という一つの数学的に普遍的な表現を使って議論できることは，多くの人が経験している．これら多様な対象と，普遍的な数学的表現の対応が明確であれば，連立方程式で解をもとめるアルゴリムである加減法や代入法の開発を行えば，多くの対象を普遍的な立場で論じたことになる．ここで最も重要ことは，多様な対象と数学的な表現をどのようにむすびつけるかを議論する，

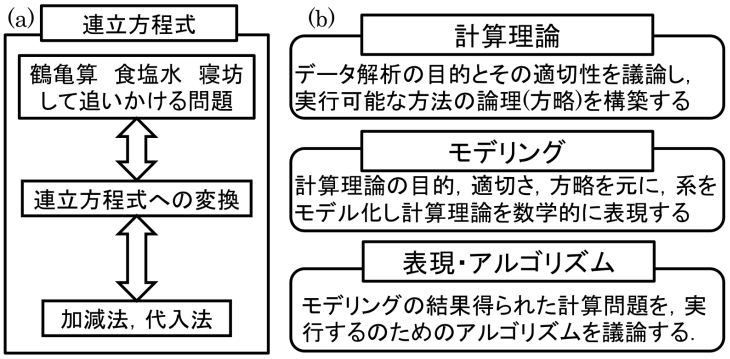


Figure 4: (a) 連立方程式とその応用 [44]．(b) データ駆動科学の三つのレベル [3, 44].

対象の数理科学的なモデリングが重要であるという視点である．図 4(a) の連立方程式とその応用の背景にある三層構造が重要なのである [44].

我々は上記の考察と神経科学者である David Marr が提案した，David Marr の三つのレベルを参考に，図 4(b) のデータ駆動科学の三つのレベルを提案した [3]．図 4(b) の上の第一のレベルの「計算理論」ではデータ解析に対応する情報処理の目標や，その目標を達成するための戦略である方略，それらの目標と方略の適切さなどを自然言語で議論する．この過程は，連立方程式の応用問題が適切に連立方程式に変換できるように文章を書くことに対応する．下の第三のレベルの「表現・アルゴリズム」のレベルは，機械学習やデータ解析手法に対応する．我々は，データ駆動科学の構築においては，データ解析の目的である計算理論のレベルと，データ解析手法である表現・アルゴリズムのレベルの間に，計算理論の表現・アルゴリズムへの変換を担う新たなモデリングのレベルが必要であることを指摘した (図 4(b)).

連立方程式の応用問題を解く前には，連立方程式を徹底的に解いておき，頭の中に連立方程式を叩き込んでおく必要がある．そのトレーニングの後に，連立方程式の数理構造を参照して，応用問題の背後にある連立方程式の数理構造を抽出することが応用問題を解く上では必須である．この考察から，表現・アルゴリズムのレベルで用いる統計学・機械学習の手法を一つに絞っておくことが，データ駆動科学を実践する際に重要であることがわかる．当然のことながら，一つに絞る手法は普遍的でなければならない．これは機械学習の本をいっぱい読んでから，データ駆動科学を推進せよと言っているのではない．アルゴリズムの開発の論文を書くために提案されたアルゴリズムに惑わされるのではなく，ミニマムな真に必要な手法に狙いを定めて，それだけをもちいて，眼の前にある実問題に潜む数学的構造を抽出するように取り組むことがデータ駆動科学の実践には必須であると述べているのである．その優れた情報科学的枠組みがスパースモデリングである．

4　最小二乗法のベイズ推論

§3 で述べたように，データ駆動科学の真髄は中学 2 年生程度でならう連立方程式にある．そこで本章では，連立方程式の構成要素である 1 次関数をベイズ推論で取り扱うことで，データ駆動科学の素養を養う．これを通して，自然の持つスパースな構造を抽出する基礎を説明する．

4.1　最小二乗法

1 次元データ $D = \{(x_i, y_i)\}_{i=1}^N$ を，以下のような 1 変数の線形モデルで回帰する問題を考える（図 5）．ちなみに，本節のみ N がサンプル数と定義されることに注意されたい．

$$y = ax + b \tag{10}$$

この問題の目標は，データ D をうまく表現できる a と b を決定することである．最小二乗法とはすなわち，この当てはまりの良さを表す指標として以下の二乗誤差のデータ平均を採用するものである．

$$E(a,b) = \frac{1}{2N}\sum_{i=1}^{N}\{y_i - (ax_i + b)\}^2 = \frac{1}{2N}\sum_{i=1}^{N}\{y_i^2 - 2(ax_iy_i + by_i) + a^2x_i^2 + 2abx_i + b^2\} \tag{11}$$

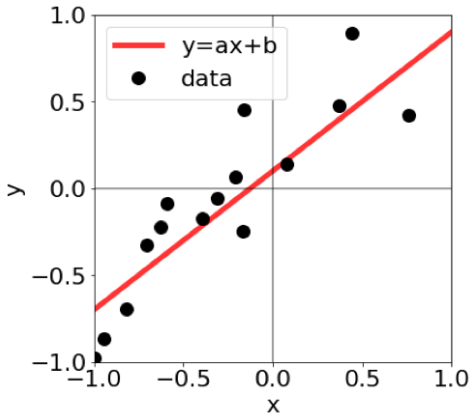


Figure 5: 単回帰問題

ここで次のような統計量を設定する．

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i,\ \ \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i,\ \ \overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2,\ \ \overline{y^2} = \frac{1}{N}\sum_{i=1}^{N} y_i^2,\ \ \overline{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i \tag{12}$$

すると，$E(a,b)$ は以下のように書き換えられる．

$$E(a,b) = \frac{1}{2}\left(\overline{y^2} - 2a\overline{xy} - 2b\bar{y} + a^2\overline{x^2} + 2ab\bar{x} + b^2\right) \tag{13}$$

ここで，$\bar{x} = 0$ となるように座標変換を行う．これによって a と b の依存関係が消去され，a と b のそれぞれの独立した二乗関数の線形和となる．その上で平方完成を行うと，

$$E(a,b) = \frac{1}{2}\left(\overline{x^2}\left(a - \frac{\overline{xy}}{\overline{x^2}}\right)^2 + (b-\bar{y})^2 - \frac{\overline{xy}^2}{\overline{x^2}} - \bar{y}^2 + \overline{y^2}\right) \tag{14}$$

となり，$E(a,b)$ を最小とする a_0, b_0 が $a_0 = \frac{\overline{xy}}{\overline{x^2}}, b_0 = \bar{y}$ と求まる．

ここで，式 (14) の a, b に関わる項をそれぞれ新しく，$\mathcal{E}_a(a) = \frac{1}{2}\overline{x^2}\left(a - \frac{\overline{xy}}{\overline{x^2}}\right)^2$，$\mathcal{E}_b(b) = \frac{1}{2}(b-\bar{y})^2$ と置き直す．すると，以下のように $E(a,b)$ を分解することができる（図 6）．

$$E(a,b) = \mathcal{E}_a(a) + \mathcal{E}_b(b) + E(a_0,b_0) \geqq E(a_0,b_0) \tag{15}$$

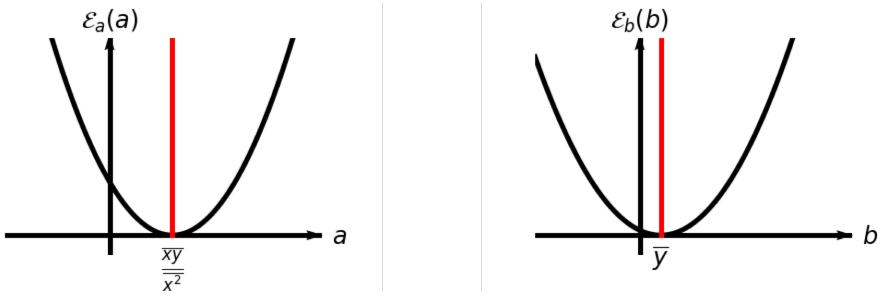


Figure 6: $\mathcal{E}_a(a)$，$\mathcal{E}_b(b)$ のランドスケープ

4.2 自然科学的視点からのベイズ推論の解説

図 7 のように，同じ直線で回帰される二つのデータを考えよう．一つは図 7(a) のように，データが直線からバラついている場合であり，もう一つは図 7(b) のように，データが直線上にほぼ乗っている場合である．

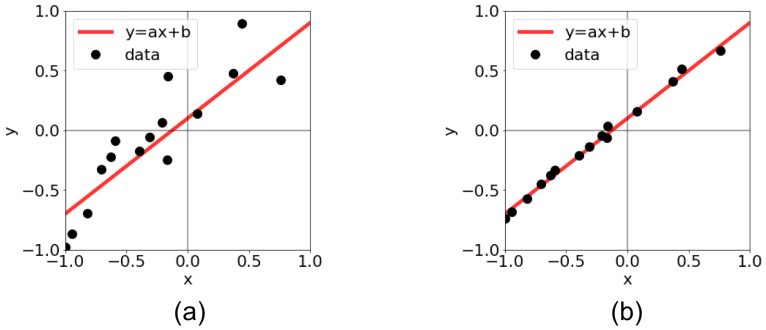


Figure 7: ノイズ分散の違うデータ

§4.1 の最小二乗法でもとめた両方の直線の傾きが等しいとする．この時，図 7(a) と (b) の違いは，どのように数学的に表現されるのであろうか? これに答えるのが，本節の目的である．直感的に考えると，図 7(a) のほうが傾き a や切片 b がばらついた中で決まりそうなのに対し，図 7(b) のほうが a や b がきちっと決まるような気がする．つまり，傾き a や切片 b の確率分布を知りたくなるわけである．この要望に応えるのがベイズ推論である．

私事で大変恐縮であるが，著者の一人の岡田がベイズ推論を用いた研究を最初に発表したのが 10 年前の 2007 年であった．発表のスライドを準備していると，ベイズ推論のことが全くわかってないことに気がついた．会議は京都駅近くの会場で行われて，それに間に合うためには，午前 6 時に出発する必要がある．今は，午前 0 時．シャレにならない状況であった．研究室にあるベイズ推論の本をいくつか斜め読みしても，まったく生じた疑問に答えてはくれなかった．この縦棒の位置がくるくる変わる意図がまったくわからない．このくるくる位置が変わる指導原理は何かがまったくわからなかった．はじめて真面目に考える条件付き確率である．本に頼るのは諦めて，自分で考えることにした．そこで，ステップ 1: 同時確率分布を出発点に議論を展開すれば良いのだと気づいた．ステップ 2: その同時確率分布を，因果律の形でまずは書き直す．ステップ 3: 次にその同時確率分布を，データを条件とした知りたい変数の条件付き確率で表現すれば理解できることに気づいた．こう考えると，どのような条件で条件付き確率を考えれば良いかがすぐにわかる．しかも，自然科学者が慣れ親しんだ因果律が，確率の形で見事に表現されている．ベイズ推論こそが，自然科学を表現する素晴らしい言葉だという確信を持ったのが午前 4 時であった．

これを §4.1 の最小二乗法を例にとり説明していこう．本節では簡単のため，$X \equiv \{x_i\}_{i=1}^N$ は実験者が系の外から決めることとする．つぎに，X 以外の残りの系を記述する変数である，傾き a，切片 b および出力 $Y \equiv \{y_i\}_{i=1}^N$ がすべて確率変数だと考え，それらの同時確率分布 $p(a,b,Y)$ を考える．これがステップ 1 である．つぎに出力 Y がどのように生成されるかを因果律の概念の元に考える．今，入力 X があらかじめ決まっているので，出力 Y は傾き a と切片 b が与えられると決まる．つまり a と b が原因となり， 結果 Y が因果的に決まるのである（図 8）．この因果律を用いて，確率の積の法則に基づくと，

$$p(a,b,Y) = p(Y|a,b)p(a,b) \tag{16}$$

となる．自然科学者が対象としている世界である同時確率分布 $p(a,b,Y)$ が，ヒトが理解しやすい因果律 $p(Y|a,b)$ を使って表現できたのである．つまり，自然科学者が慣れ親しんだ因果律が，確率の形で見事に表現されたのである．これがステップ 2 である．しかも，この視点を持つと，データの背後にある本質を抽出するという科学の営みは条件付き確率 $p(a,b|Y)$ で表現できることに気づき，因果律と科学の営みは等号で結ぶことが可能となることを見出した．

$$p(a,b,Y) = p(Y|a,b)p(a,b) = p(a,b|Y)p(Y), \quad p(a,b|Y) = \frac{p(Y|a,b)p(a,b)}{p(Y)} \tag{17}$$

これがステップ 3 である．このベイズの公式の意味は，具体的な問題を取り扱うことで深みを増す．

4.3 ベイズ的なパラメータ分布 $p(a,b|Y)$ 推定

本節では，§4.2 で述べたデータの背後にある本質 $p(a,b|Y)$ を a,b の確率密度関数として書き下し，そこからパラメータ a,b についての情報を得ることを考える．そのために，ヒトが理解しやすい因果律 $p(Y|a,b)$ (図 8) に具体的なモデルを導入する．ちなみに，この $p(a,b|Y)$ は，データ Y が与えられた元での a,b の

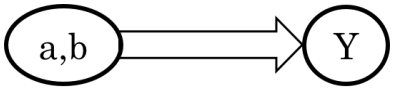


Figure 8: 単回帰の因果モデル

確率という意味で，よく事後確率と呼称される．同様に，§4.2 で出てきた $p(a,b)$ は，データ Y が与えられる前の a,b の確率という意味で事前確率と呼称される．

具体的な因果律のモデル $p(Y|a,b)$ を導入する．単回帰モデルではよく，パラメータを a,b とする線形モデルに，ノイズ n_i が加算されてデータ y が生成されると考える．

$$y_i = ax_i + b + n_i \tag{18}$$

ここで，ノイズ n_i は平均 0 で分散 σ^2 のガウス分布に従うとする．つまり，

$$p(n_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{n_i^2}{2\sigma^2}\right) \tag{19}$$

となる．$n_i = y_i - (ax_i + b)$ であるので，$p(n_i)$ は，a,b が与えられた元での y_i の条件付き確率 $p(y_i|a,b)$ と表現でき，y_i の生成モデルは，

$$p(n_i) = p(y_i|a,b) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-(ax_i+b))^2}{2\sigma^2}\right) \tag{20}$$

と書き直すことができる．今，複数のデータ $(X,Y) = \{(x_1,y_2),(x_2,y_2)\cdots(x_N,y_N)\}$ が独立に生成されると考えると，

$$\begin{aligned} p(Y|a,b) &= \prod_{i=1}^{N} p(y_i|a,b) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left(-\frac{\sum_{i=1}^{N}(y_i-(ax_i+b))^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left(-\frac{N}{\sigma^2}E(a,b)\right) \end{aligned} \tag{21}$$

となる．従って，a,b の事前分布 $p(a,b)$ を一様分布とした場合，式 (17) より，a,b についての事後確率は以下で与えられる．

$$\begin{aligned} p(a,b|Y) &= \frac{p(Y|a,b)p(a,b)}{p(Y)} \propto p(Y|a,b) \\ &= \exp\left\{-\frac{N}{\sigma^2}\Big(\mathcal{E}_a(a)+\mathcal{E}_b(b)+E(a_0,b_0)\Big)\right\} \\ &\propto \exp\left\{-\frac{N}{\sigma^2}\Big(\mathcal{E}_a(a)+\mathcal{E}_b(b)\Big)\right\} \\ &= \exp\left\{-\frac{N\overline{x^2}}{2\sigma^2}(a-a_0)^2+\frac{N}{2\sigma^2}(b-b_0)^2\right\} \end{aligned} \tag{22}$$

これより，a と b の事後確率は，a_0,b_0 を平均値とし，データノイズの大きさ σ^2 に比例した分散を持つガウス分布に従うことがわかる（図 9）．また，データ数の増大によって分散が減少することから，データ数の増大でその精度が向上することも表している．また，この事後確率から，パラメータ a,b の推定を行うことができる．例えば，事後確率が最大となる a,b の点を推定値とする最大事後確率（Maximum *a posteriori*, MAP）推定を行うと，事後確率がガウス分布となることから，その平均値の $a_{map} = a_0, b_{map} = b_0$ が最適値として得られる．

ここで，回帰の最終的な目標の 1 つである「未知の入力 x' に対する予測値 y' を推定する」ことを考えると，MAP 推定の結果得られたパラメータ a_{map}, b_{map} を用いて $y' = a_{map}x' + b_{map}$ と推定することが考えられる．さらにベイズ推定では，事後確率 $p(a,b|Y)$ を得たことを活かし，§4.5 で述べるようにパラメータ a,b を介さず直接予測分布 $p(y'|Y)$ を推定することができる．これにより，パラメータ a,b のゆらぎのような効果を加味した $p(y'|Y)$ の推定が可能となる．

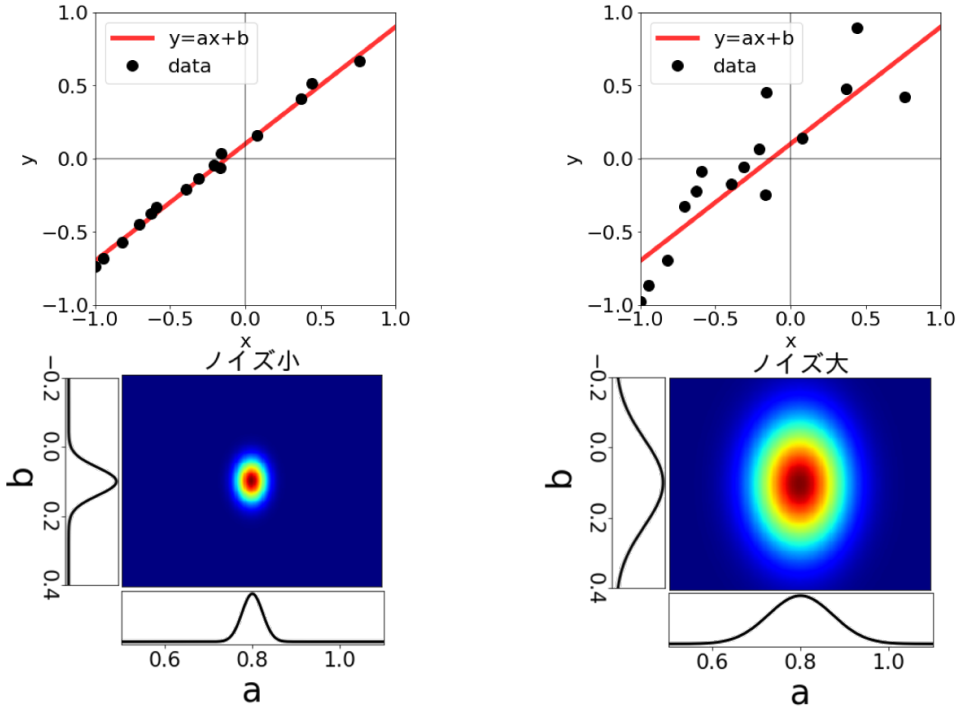


Figure 9: 事後確率分布とノイズの大きさの関係

4.4 ベイズ的なノイズ分散分布 $p(\sigma^2|Y)$ 推定

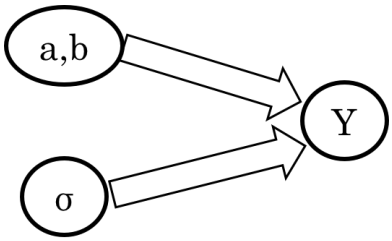


Figure 10: ノイズ分散 σ^2 を導入した単回帰の因果モデル

ここまでは，ノイズの分散が与えられていると考えてきたが，§4.3 節の a, b の推定でそうしたように，ここで，ノイズ分散 σ^2 を確率変数と考えその推定を行う．そこで，a, b 推定時と同様に，§4.2 節で述べた手続きに従い，σ^2 についての事後確率 $p(\sigma^2|Y)$ を算出する．まず同時確率を σ^2 を含む形で $p(\sigma^2, a, b, Y)$ と設定する（ステップ 1）．次に同時確率を図 10 にある因果律に基づき表現し直す（ステップ 2）．

$$p(\sigma^2, a, b, Y) = p(Y|\sigma^2, a, b)p(a, b)p(\sigma^2) \propto p(Y|\sigma^2, a, b) \tag{23}$$

最後に，同時確率 $p(\sigma^2, a, b, Y)$ を知りたい事後確率 $p(\sigma^2|Y)$ を用いて表現する（ステップ 3）．

$$\int dadb\, p(\sigma^2, a, b, Y) = \int dadb\, p(\sigma^2|a, b, Y)p(a, b|Y)p(Y) = p(\sigma^2|Y)p(Y) \propto p(\sigma^2|Y) \tag{24}$$

ここから，σ^2 の事後確率を求めるには，$p(Y|\sigma^2, a, b)$ を a, b について周辺化すれば良いとわかる．

$$\begin{aligned} p(\sigma^2|Y) &\propto \int dadb\, p(Y|a,b,\sigma^2) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \int dadb \exp\left\{-\frac{N}{\sigma^2}E(a,b)\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \left\{\exp\left(-\frac{N}{\sigma^2}E(a_0,b_0)\right) + \int da \exp\left(-\frac{N\overline{x^2}}{2\sigma^2}(a-a_0)^2\right) + \int db \exp\left(-\frac{N}{2\sigma^2}(b-b_0)^2\right)\right\} \\ &= (2\pi\sigma^2)^{-\frac{N-2}{2}}(N^2\overline{x^2})^{\frac{1}{2}} \exp\left(-\frac{N}{\sigma^2}E(a_0,b_0)\right) \end{aligned} \tag{25}$$

この事後確率を最大化する σ^2 を推定値とすること（MAP 推定）を考える．a, b の事後確率 $p(a, b|Y)$（式 (22)）のときのように，σ^2 の事後確率 $p(\sigma^2|Y)$（式 (25)）はそれを最大化する σ^2 がすぐにわかる形式ではないので，微分が 0 という極値条件を用いる．この微分を容易に行えるように，ここで事後確率 $p(\sigma^2|Y)$ の負の対数尤度をとった関数 $F(\sigma^2)$ を定義する．

$$F(\sigma^2) = -\log p(\sigma|y) = \frac{N}{\sigma^2}E(a_0,b_0) + \frac{N-2}{2}\log(2\pi\sigma^2) \tag{26}$$

さらに，式を簡略化するために，$\sigma^2 = s$ とおく．

$$F(s) = \frac{N}{s}E(a_0,b_0) + \frac{N-2}{2}\log(2\pi s) \tag{27}$$

これより $F(s)$ の極値条件は，

$$\frac{\partial F(s)}{\partial s} = -\frac{N}{2s^2}E(a_0,b_0) + \frac{N-2}{2s} = 0 \tag{28}$$

となる．従って，σ^2 は，

$$\sigma^2 = \frac{NE(a_0,b_0)}{N-2} = \frac{1}{N-2}\sum_{i=1}^{N}\{y_i - (a_0x_i + b_0)\}^2 \tag{29}$$

と推定される．

4.5 ベイズ的な予測分布 $p(y'|Y)$ 推定

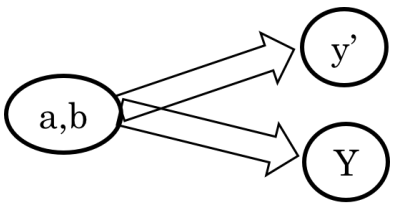


Figure 11: 予測値 y' を導入した単回帰の因果モデル

§4.3 で述べたように，回帰の最終的な目的の 1 つは，訓練データ Y にもとづき，未知の点 x' における予測値 y' を推定するというものである．そこで本節では，訓練データセット Y が与えられた元での未知の点 x' の予測値 y' が従う確率分布についての検討を行う．つまり，y' についての事後確率 $p(y'|Y)$ を算出する．これまでと同様に同時確率 $p(a, b, Y, y')$ から考える（ステップ 1）．この同時確率を因果律（図 11）で表現すると（ステップ 2），

$$p(y', a, b, Y) = p(y'|a,b)p(Y|a,b)p(a,b) \propto p(y'|a,b)p(Y|a,b) \tag{30}$$

となる．最後に同時確率 $p(a, b, Y, y')$ を，求めたい事後確率 $p(y'|Y)$ を用いて表現すると（ステップ 3），

$$\int dadb\, p(y', a, b, Y) = \int dadb\, p(y'|a,b,Y)p(a,b|Y)p(Y) = p(y'|Y)p(Y) \propto p(y'|Y) \tag{31}$$

となる．これより，事後確率 $p(y'|Y)$ は，

$$p(y'|Y) \propto \int dadb\, p(y'|a,b)p(Y|a,b) \propto \int dadb\, p(y'|a,b)p(a,b|Y) \tag{32}$$

となる．a,b についての事後確率 $p(a,b|Y)$ は式 (22) で与えられる．また，y' は推定された a,b の元で生成されるわけであるので，

$$p(y'|a,b) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y'-(ax'+b))^2\right) \tag{33}$$

で与えられる．これらを式 (12) へ代入する．その際，表記の省略のため，$\sigma_0'^2 = \frac{\sigma^2}{N}$，$\sigma_1'^2 = \frac{\sigma^2}{N\overline{x^2}}$ とおく．

$$\begin{aligned}
p(y'|Y) &= \int\int dadb \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_1'^2}} \frac{1}{\sqrt{2\pi\sigma_0'^2}} \exp\left(-\frac{1}{2\sigma^2}(y'-(ax'+b))^2 - \frac{1}{2\sigma_1'^2}(a-a_0)^2 - \frac{1}{2\sigma_0'^2}(b-b_0)^2\right) \\
&\propto \int db \left[\int da \exp\left\{-\left(\frac{x'^2}{2\sigma^2}+\frac{1}{2\sigma_1'^2}\right)\left(a-\left(\frac{x'^2}{2\sigma^2}+\frac{1}{2\sigma_1'^2}\right)^{-1}\left(\frac{x'y'}{2\sigma^2}+\frac{x'b}{2\sigma^2}+\frac{a_0}{2\sigma_1'^2}\right)\right)^2\right\}\right. \\
&\left.\times \exp\left\{\left(b-\left(\frac{1}{4\sigma^2\sigma_1'^2}+\frac{x'^2}{4\sigma^2\sigma_0'^2}+\frac{1}{4\sigma_0'^2\sigma_1'^2}\right)^{-1}\left(-\frac{x'a_0}{4\sigma^2\sigma_1'^2}+\frac{y'}{4\sigma^2\sigma_1'^2}+\frac{x'^2b_0}{4\sigma^2\sigma_0'^2}+\frac{b_0}{4\sigma_0'^2\sigma_1'^2}\right)\right)^2\right\}\right] \\
&\times \exp\left\{-\left(\frac{y'^2}{2\sigma^2}+\frac{a_0^2}{2\sigma_1'^2}+\frac{b_0^2}{2\sigma_0'^2}\right)+\left(\frac{x'^2}{2\sigma^2}+\frac{1}{2\sigma_1'^2}\right)^{-1}\left(\frac{x'y'}{2\sigma^2}+\frac{a_0}{2\sigma_1'^2}\right)^2\right\} \\
&\times \exp\left\{+\left(\frac{x'^2}{2\sigma^2}+\frac{1}{2\sigma_1'^2}\right)^{-1}\left(\frac{1}{4\sigma^2\sigma_1'^2}+\frac{x'^2}{4\sigma^2\sigma_0'^2}+\frac{1}{4\sigma_0'^2\sigma_1'^2}\right)^{-1}\left(-\frac{x'a_0}{4\sigma^2\sigma_1'^2}+\frac{y'}{4\sigma^2\sigma_1'^2}+\frac{x'^2b_0}{4\sigma^2\sigma_0'^2}+\frac{b_0}{4\sigma_0'^2\sigma_1'^2}\right)^2\right\}
\end{aligned} \tag{34}$$

このガウス積分を実行した後，y' について平方完成すると y' についての事後確率が得られる．

$$p(y'|Y) \propto \exp\left\{-\frac{1}{2\left(\sigma^2+x'^2\sigma_1'^2+\sigma_0'^2\right)}(y'-(a_0x'+b_0))^2\right\} \tag{35}$$

これが予測分布となる．したがって，予測分布の分散 $\sigma_{y'}^2$ は

$$\sigma_{y'}^2 = \sigma^2 + \frac{x'^2}{N\overline{x^2}}\sigma^2 + \frac{1}{N}\sigma^2 \tag{36}$$

と与えられる．最初に行った，訓練セットの平均 $\bar{x}$ が 0 となるような座標変換を元に戻すと，

$$\sigma_{y'}^2 = \sigma^2 + \frac{(x'-\bar{x})^2}{N\overline{(x-\bar{x})^2}}\sigma^2 + \frac{1}{N}\sigma^2 \tag{37}$$

となる．これは，予測分布の分散が訓練セットの平均 $\bar{x}$ で最小となる 2 次関数となることを意味する（図 12）．さらに，§4.4 節での分散の推定値（式 (29)）を代入すると，

$$\sigma_{y'}^2 = \frac{N}{N-2}E(a_0,b_0) + \frac{1}{(N-2)}\frac{x'^2}{\overline{x^2}}E(a_0,b_0) + \frac{1}{N-2}E(a_0,b_0) \tag{38}$$

が得られる．ここから，重複のない訓練データの数が増えるほど第 2 項と第 3 項が減衰し，サンプル数無限大の極限では，分散がノイズ分散 σ^2 と一致することがわかる．

ここで，切片なしモデル $y=ax$ と切片のみモデル $y=b$ を設定し，これらに対して予測分布を推定することを考える．

$$p_a(y'|Y) = \int da \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_1'^2}} \exp\left(-\frac{1}{2\sigma^2}(y'-ax')^2 - \frac{1}{2\sigma_1'^2}(a-a_0)^2\right) \tag{39}$$

$$p_b(y'|Y) = \int db \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_0'^2}} \exp\left(-\frac{1}{2\sigma^2}(y'-b)^2 - \frac{1}{2\sigma_0'^2}(b-b_0)^2\right) \tag{40}$$

$y = ax + b$ の場合と同様の計算を行うと，切片なしモデルの予測分布の分散 $\sigma^2_{y',a}$ と切片のみモデルの分散 $\sigma^2_{y',b}$ が得られる．

$$\sigma^{a2}_{y',a} = \frac{N}{N-2}E(a_0, b_0) + \frac{1}{(N-2)}\frac{x'^2}{\overline{x^2}}E(a_0, b_0) \tag{41}$$

$$\sigma^{b2}_{y',b} = \frac{N}{N-2}E(a_0, b_0) + \frac{1}{N-2}E(a_0, b_0) \tag{42}$$

これより，$\sigma^2_{y'}$ を構成する式 (38) の第 2 項が，傾き a のゆらぎに対応し，第 3 項が切片 b のゆらぎに対応するとわかる．さらに，第 2 項が $(x' - \bar{x})^2$ に比例することは，予測分布の標準偏差が $|x' - \bar{x}|$ に比例することを意味しており，ここから，第 2 項が，傾き a のゆらぎに応じた $\bar{x}$ を固定点とした回転自由度に応じた分散と解釈される．また，第 3 項も切片 b のゆらぎに対応した y 方向の平行移動の自由度に応じた分散であると解釈される．

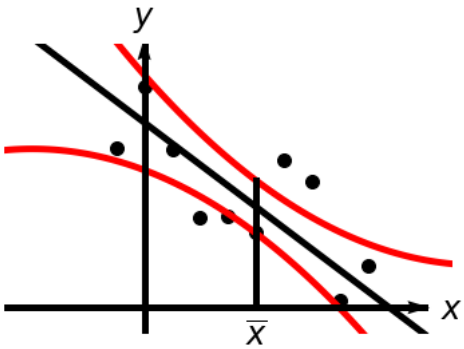


Figure 12: 予測分布：回帰式（黒線）を中心とした正規分布（赤線は分散）となる．

4.6 ベイズ的なモデル分布 $p(K|Y)$ 推定（ベイズ的モデル選択）

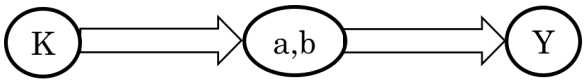


Figure 13: モデル K を導入した単回帰の因果モデル

ここまでは，$y = ax + b$ というモデルが与えられた元でのパラメータ推定を行ってきた．一方で，複数のモデルの中からどのモデルがデータ Y を説明するのに最適であるかを調べることも重要である．そこで本節では，モデルの選び方自体を確率変数と捉え，ベイズ推論に基づく定式化によって最適なモデルを選択することを考える．ここではモデルのラベルを K とし，$y = ax$ を $K = 1$ のモデル，$y = ax + b$ を $K = 2$ のモデルとしたうえで，2 つのモデルから最適なモデルを選ぶ問題を考える．そのための指標としてモデルの事後確率 $p(K|y)$ を計算し，この確率を大きくするようなモデルを選択することを考える．この事後確率を計算するために，まずは同時確率から考える（ステップ 1）．$K = 2$ の場合の同時確率の因果関係は $p(K)$ と $p(a, b, |K = 2)$ を一様分布として，

$$p(K = 2, a, b, Y) = p(Y|a, b)p(a, b, |K = 2)p(K = 2) \propto p(Y|a, b) \tag{43}$$

のように書ける (ステップ 2)．最後に同時確率を事後確率 $p(K = 2|Y)$ で表現すると，

$$\int dadb\, p(K = 2, a, b, Y) = \int dadb\, p(K = 2|a, b)p(a, b|Y)p(Y) \propto p(K = 2|Y) \tag{44}$$

となる (ステップ 3)．これより事後確率は以下のように求められる．

$$\begin{aligned} p(K=2|Y) &\propto \int dadb\, p(Y|a,b) \\ &\propto \exp\left\{-\frac{N}{\sigma^2}E(a_0,b_0)\right\}\int da\exp\left\{-\frac{N\overline{x^2}}{2\sigma^2}(a-a_0)^2\right\}\int db\exp\left\{-\frac{N}{2\sigma^2}(b-b_0)^2\right\} \end{aligned} \quad (45)$$

この形から，パラメータの数が増えると誤差 $E(a_0,b_0)$ が小さくなるため，事後確率は大きくなる傾向にあることが分かる．ただし，パラメータの数が増えると積分するガウス分布も増えるため，これが結果として罰金項として生じることになる．ガウス積分を実行すると，

$$p(K=2|Y) \propto \exp\left\{-\frac{N}{\sigma^2}E(a_0,b_0)\right\}\times\sqrt{\frac{2\pi\sigma^2}{N\overline{x^2}}}\times\sqrt{\frac{2\pi\sigma^2}{N}} \quad (46)$$

が得られる．この事後確率を異なるモデル K 毎に計算し，事後確率が最大となる K を最適なモデルとすること（MAP 推定）を考える．この事後確率の負の対数をとった関数 $F(K=2)$ が以下のように計算できる．

$$\begin{aligned} F(K=2) &= -\log p(K=2|Y) \\ &= N\left\{\frac{1}{\sigma^2}E(a_0,b_0)+\frac{\log N}{N}\right\} \end{aligned} \quad (47)$$

この K をハイパーパラメータとして推定する際に定義される $F(K)$ は，物理学における自由エネルギーとの数式的な類似性から，よくベイズ自由エネルギーと呼称される．実は，ノイズ分散推定の際に出てきた $F(\sigma^2)$ もこのベイズ自由エネルギーと同様の表式となっていたため，関数名に F を用いていた．ちなみに，事後確率 $p(K|Y)$ を最大化することは，ベイズ自由エネルギー $F(K)$ を最小化することと等価となる．以上の計算を $K=1$ の場合も同様にして行うと，

$$p(K=1|y) \propto \exp\left\{-\frac{N}{\sigma^2}E(a_0)\right\}\times\sqrt{\frac{2\pi\sigma^2}{N\overline{x^2}}} \quad (48)$$

$$F(K=1) = N\left\{\frac{1}{\sigma^2}E(a_0)+\frac{\log N}{2N}\right\} \quad (49)$$

のようになる．このように，各モデルについて計算された事後確率 $p(K|Y)$ とベイズ自由エネルギー $F(K)$ を比較することにより，系統的なモデル選択を行うことが可能になる．例えば，今回 1 次元であった説明変数 x が高次元である場合は，切片を加えたその全ての組み合わせについてそれぞれ $F(K)$ を求めて比較すれば，取りうるモデルの中から最も最適なモデルが決定される．自然がスパースな構造を持っているならば，この時選ばれるモデルは少数の説明変数によって構成されるスパースなモデルとなることが期待される．

5　スパースモデリングの深化

§4 の議論から，多くの人が慣れ親しんでいる $y=ax+b$ の最小二乗法の背景に豊かな数理が潜んでいることがわかった．この議論から，説明したい変数である目的変数を説明する変数を決めることは，計算量の指数爆発を起こす問題であることがわかる．たとえば $y=ax+b$ の場合，実はそれが $y=ax$ であることや，$y=b$ であることを考慮すると，3 通りの事後確率 $p(K|y)$ とベイズ自由エネルギー $F(K)$ を計算して，比較する必要がある．これがもし $y=\sum_{i=1}^{N}a_ix_i$ とかける場合であると，2^N-1 通りの事後確率 $p(K|y)$ とベイズ自由エネルギー $F(K)$ を計算して，比較する必要があり，これは N に関して指数爆発する量である．この指数爆発を避ける方法の一つが，問題を L1 最適化で近似する枠組みである [34] である．L1 最適化にもとづくスパースモデリングについては，本領域のチュートリアルでも複数回取り上げられているので，ここでは議論しない [4–6]．

Cover と Van Campenhout は N 個の説明変数から適切な変数を選択するアルゴリズムの計算量は，指数爆発することを示している [45]．L1 最適化によるスパースモデリングは，あくまでも近似アルゴリズムなのである．L0 最適化と L1 最適化が同じ答えを出す条件は，実データを解析する上では，かなり特殊な条件であることを強調したい [35]．以下の節では，実データ解析に L1 最適化などのスパースモデリングの近似アルゴリズムを適用した際に，我々が実際に体験した困難を説明して，それをどう解決しつつあるかを，スパースモデリングの深化の一例として紹介する．

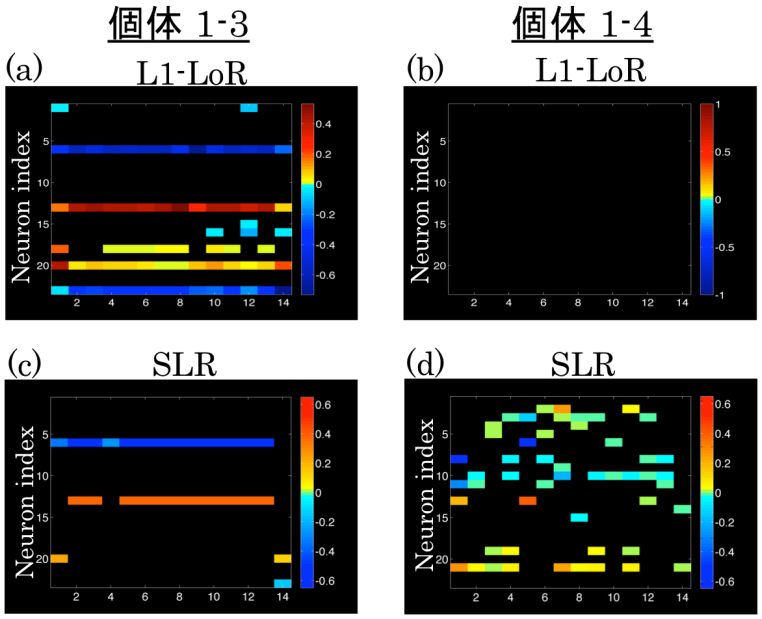


Figure 14: 実データにおける識別問題に，L1-LoR 及び SLR を適用した際に選択される変数のウェイトダイアグラム．横軸が交差検定の試行回数，縦軸がニューロン番号を表す．(a), (c) 個体 1 と 3 の識別問題に対して，L1-LoR と SLR を適用した結果．(b), (d) 個体 1 と 4 の識別問題に対して，L1-LoR と SLR を適用した結果．

5.1 近似アルゴリズムによるスパースモデリング

ここでは，Eifuku らが，顔観察時のサル側頭葉の単一ニューロン計測を行った際の電気生理学データを用いて識別におけるスパースモデリングを議論する [7, 46–48]．Eifuku らは，4 個体のヒトの顔それぞれについて，7 通りの角度で撮影した合計 28 枚の顔画像をマカクサルに観察させ，大脳皮質の AIT(Anterior Inferior Temporal cortex) の 23 個の神経細胞を記録した．これら 23 個の神経細胞の発火率を入力データとし，23 個の神経細胞のうち，どの神経細胞が角度によらない個体識別に寄与するかを議論する [7, 48]．

4 個体のうち 2 個体を選び，その識別を神経細胞の発火率データから行うことで，線形識別問題で取り扱う．なお，個体の組み合わせの識別問題は ${}_4C_2 = 6$ 通りであり，例えば個体 1 と個体 3 の組み合わせであれば，個体 1-3 と表記する．このような二値の識別問題における L1 最適化アルゴリズムとして，ここでは L1 ロジスティック回帰 (L1-LoR) を用いる [49]．N 次元の入力 $\mathbf{x} = (x_1, x_2, \ldots, x_N)^{\mathrm{T}}$ を用いてラベル $y \in \{-1, 1\}$ の二値識別問題を解く．p 組の学習データ $\mathbf{X} \equiv \{\mathbf{x}^\mu\}_{\mu=1}^p$, $\mathbf{Y} \equiv \{y^\mu\}_{\mu=1}^p$ が与えられているとする．このとき，二値の識別問題は，下記に表されるロジスティック損失を最小化する問題に帰着される．

$$\epsilon(\mathbf{w}; \mathbf{X}, \mathbf{Y}) \equiv \frac{1}{p} \sum_{\mu=1}^{p} \log(1 + \exp(-y^\mu f(\mathbf{x}^\mu; \mathbf{w})), \tag{50}$$

ここで，ロジスティックモデルを $f(\mathbf{x}; \mathbf{w}) = 1/(1 + \exp(-\sum_i w_i x_i))$ とし，$\mathbf{w} = (w_1, w_2, \ldots, w_N)$ はロジスティック回帰の係数である．

L1-LoR においては，L1 正則化項を導入して，下記の最適化を行うことで，重要な変数に対応するインデックスの $\mathbf{w}$ が非ゼロとなり，変数選択が行える．

$$\arg\min_{\mathbf{w}} \epsilon(\mathbf{w}; \mathbf{X}, \mathbf{Y}) + \lambda |\mathbf{w}| \tag{51}$$

この最適化は多項式オーダーの計算量で可能であり，スパースモデリングを近似的に行う枠組みである．ここで λ は正則化パラメータであり，λ が十分に大きい場合は係数 $\mathbf{w}$ がほとんど 0 となる．この λ は通常の L1 最適化と同様に，交差検証誤差の最小により決定する．

図 14 は L1-LoR をこのデータに適用した際に選択された，変数とその重みを図示したウエイトダイアグラム (WD) である．なお、交差検証は leave-one-out を用いた．図 14 の WD の各列は交差検証の各回に対応し，黒は，その変数が選択されてないこと意味する．黒以外で着色されている変数が選択されたことを示し, その変数に対する係数を図のようにカラーバーであらわす．個体 1-3 の識別においてはニューロン

6 とニューロン 13 が安定的に選択されているが，個体 1-4 の識別では選択されたニューロンがない．これは何を意味するのであろうか? そこで我々は，同じデータに対して，別のスパースモデリングのアルゴリズムである，SLR(Sparse Logistic Regression) を適用した [50]．その結果，図 14 に示すように，個体 1-3 の識別においては L1-LoR と SLR のスパース変数選択の結果はほぼ一致している．一方，個体 1-4 の識別の SLR の結果は，交差検証ごとに選ばれる変数が異なっている．これら L1-LoR と SLR の知見から，個体 1-3 と個体 1-4 の識別では何か定性的に状況が異なっているようである．

個体 1-4 では L1-LoR の際の交差検証誤差の最小化がうまく働かずに，λ が発散しているため，選択されたニューロンがなくなっている．つまり，これは神経細胞の活動データにそもそも個体 1-4 の識別をする能力がないことを意味するかもしれない．一方，SLR で交差検証ごとに選択されるニューロンが異なるのは，SLR が変分ベイズ推論に基づいており，最適化関数が多峰性を持つためかもしれない．いずれにしても，この二つは近似なので，そもそもデータに識別するための情報がないのか，それともたまたま近似手法がうまく動かなかったかの区別をつけることは原理的に不可能である．そこで我々は次節で導入する，変数の部分集合全てに対して交差検証誤差を計算する全状態探索法 (ES: exhaustive search) を提案した [7, 48, 51, 52].

5.2 全状態探索法によるスパースモデリング

全状態探索法 (ES) では変数の部分集合全てに対して交差検証誤差等による評価を行う．そこでまず，変数の部分集合の記述として，説明変数の集合を表すインディケータ $\mathbf{c} = (c_1, c_2, \ldots, c_N) \in \{0,1\}^N$ を導入する．$c_i = 1$ が i 番目の説明変数 x_i がその集合に属していることを，$c_i = 0$ は属していないことを表す．また，$\mathbf{c}$ の表すそれぞれの実現値を統計力学との類推から「状態」と呼ぶ．状態 $\mathbf{c}$ にあるときに選択された説明変数を $\mathbf{x}(\mathbf{c})$ と定義する．具体的には，アダマール積 $\circ$ を用いることで，$\mathbf{x}(\mathbf{c}) = (\mathbf{x} \circ \mathbf{c}) = (x_i c_i)_{1 \leq i \leq N}$ として状態 $\mathbf{c}$ の説明変数の集合を表現することができる．全状態探索法 (ES) においては，すべての状態 c における交差検証誤差などの評価を行う枠組みである．

全状態探索法 (ES) をさきほどの二値識別問題に適用することを考える．各状態 $\mathbf{c}$ における識別問題のための学習機械として，サポートベクターマシン (SVM)，特に線形 SVM をここでは用いる．我々は，学習機械として SVM を用いているので，ここで用いた手法を ES-SVM と呼ぶ．状態 $\mathbf{c}$ におけるサポートベクターマシンの識別関数及び誤差関数として次の関数を用いた．

$$y = \mathrm{sgn}(\mathbf{x}(\mathbf{c}); \mathbf{w}(\mathbf{c})) = \mathrm{sgn}(\mathbf{x} \circ \mathbf{c}; \mathbf{w}) = \mathrm{sgn}(w_0 + \sum_i^N c_i w_i x_i) \tag{52}$$

$$\epsilon(\mathbf{w}; \mathbf{Y}, \mathbf{X}) \equiv \frac{1}{p} \sum_{\mu=1}^{p} \max(0, 1 - y^\mu (w_0 + \sum_i^N c_i w_i x_i^\mu)) + \lambda \sum_{i=1}^N w_i^2 \tag{53}$$

ここでは，サポートベクターマシンの各説明変数に対する係数として，$\mathbf{w} = (w_1, w_2, \ldots, w_N)$ を定義している．

変数選択における評価基準としては説明変数の予測能力を評価する基準として，交差検証誤差 (cross validation error, CVE) を用いる．状態 $\mathbf{c}$ に対して CVE をどのように計算するかを説明する．交差検証（CV）の手続きは三つのステップからなっている．まず，全学習データ $(\mathbf{X}, \mathbf{Y})$ を訓練データ $(\mathbf{X}_{\mathrm{tr}}, \mathbf{Y}_{\mathrm{tr}})$ とテストデータ $(\mathbf{X}_{\mathrm{te}}, \mathbf{Y}_{\mathrm{te}})$ の二つに分ける．次に，訓練データのみを用いて線形 SVM を学習する．学習の結果得られるパラメータを $\hat{\mathbf{w}}_{\mathrm{tr}}(\mathbf{c})$ と書く．最後に，訓練された線形 SVM がテストデータの入出力関係をどの程度正しく説明できるかを，

$$\mathrm{CVE}(\mathbf{c}) = \frac{1}{p_{\mathrm{te}}} \sum_{\mu \in \mathrm{te}} \mathbf{1}(y^\mu (\hat{w}_{\mathrm{tr},0} + \sum_i^N c_i \hat{w}_{\mathrm{tr},i} x_i^\mu) < 0) \tag{54}$$

によって評価する．これは，テストデータにおける誤識別率を表す．p_{te} はテストデータの個数を表す．この CVE は漸近的には赤池情報量基準（AIC, [53]）に対応する [54]．CVE の値はデータの分け方によりばらつくため，データの分け方を変えて繰り返し CV を行うことにより CVE の典型値を評価する．たとえば M-fold CV では，全学習データを同じ大きさになるよう M 個の部分集合にランダムに分け，それぞれを一度ずつテストデータとして用いて M 回 CV を行い CVE の平均値を計算する．本節では，10-fold CV を利用して以下の交差検証誤差を求め，式 (53) 内の変数 λ については，今回は $\lambda = 5.0$ と固定してシミュレーションを行った．

5.3 状態密度つき全状態探索法 (ES-DoS)

p 個の学習データ $(\mathbf{X}, \mathbf{Y})$ に対して，SVM を学習するとき，説明変数の集合である状態 $\mathbf{c}$ を定めると，CVE($\mathbf{c}$) を計算することができる．したがって，2^N 個存在する状態 $\mathbf{c}$ すべてに対してこの CVE に基づいて ES を

行うと最適な説明変数の組が求まる．

従来の ES に関する研究を振り返ると，その目的は最適な説明変数の組合せを求めることに限られていた [55]．しかしながら，それだけでは必ずしも ES を最大限に活用しているとは言えない．本研究では ES の更なる進展を求めて「状態密度（DoS)」の重要性について議論する．DoS とは統計力学で用いられる考え方であり，全状態に対する評価値の度数分布

$$D_H(E) = \sum_{\mathbf{c}} \delta\left(E - H(\mathbf{c})\right) \tag{55}$$

である．ここで H は CVE といった評価基準を表す．また，$\delta(E)$ は，クロネッカーのデルタ関数である．ES を用いれば直ちに状態密度を求められることは明らかである．

5.4 近似的全状態探索法 (AES)

全状態探索では，変数の数に対する指数的オーダーで計算量が爆発してしまう．Nagata らは，この問題を情報統計力学で開発されたモンテカルロ法およびマルチヒストグラム法を利用した近似的全状態探索法 (Approximate Exhaustive Search, AES) を提案した [3, 48]．ここでは，CVE などの評価基準 H を「ハミルトニアン」と見なし，状態 $\mathbf{c}$ に関する「カノニカル分布」

$$p_\beta(\mathbf{c}) = \frac{1}{Z_{\beta_m}} e^{-\beta H(\mathbf{c})} \tag{56}$$

$$Z_{\beta_m} = \sum_{\mathbf{c}} e^{-\beta H(\mathbf{c})} \tag{57}$$

からの状態サンプルを得る．ここで β は逆温度を表す．複数の温度のカノニカル分布からサンプリングを行うために，交換モンテカルロ (EMC) 法を用いた [56]．EMC 法は parallel tempering と呼ばれることもある [57]．この方法では，評価関数 H が局所解を複数もつ系でも，複数の温度間での交換を行うアニーリング効果を発揮することで，効率的かつ網羅的に最適な変数の組み合わせを探索することができる．AES 法では EMC 法により得られたサンプルにマルチヒストグラム法を適用することにより状態密度分布 DoS を推定する [58]．マルチヒストグラム法では，EMC 法によりサンプリングされた結果から得られる各温度でのヒストグラム $H_{\beta_m}(E)$ を入力として，以下に示す状態密度分布 $D_H(E)$ とカノニカル分布の規格化定数 Z_{β_m} の反復方程式により状態密度関数 $D_H(E)$ を推定する．

$$D_H(E) = \frac{\sum_{m=1}^{M} H_{\beta_m}(E)}{\sum_{m=1}^{M} n_m \exp(-\beta_m E)/Z_{\beta_m}} \tag{58}$$

$$Z_{\beta_m} = \sum_{E} D_H(E) \exp\left(-\beta_m E\right) \tag{59}$$

ここで，n_m は m 番目のレプリカにおける総サンプル数である．

5.5 ES-DoS の人工データへの適用

次に適用した人工データについて述べる．ここで適用する人工データは，図 15 で示すように，ラベル -1 とラベル $+1$ の二値識別の問題であり，この問題を線形識別の枠組みで議論する．図 15(a) の場合は，識別に変数 1 と変数 2 の両方を用いた方が識別性能が向上するのは明らかである．一方，図 15(b) の場合は，変数 4 は識別に寄与しないため，変数 3 だけを用いた方が識別性能が向上するはずである．図 15(a) に関しては二つの変数を用いることで識別性能が上がり，(b) については一つの変数だけを用いる方が識別性能があがる．これは図のように二つの変数の分布の度合いを見れば明らかであるが，変数ごとの周辺分布については，図 15(a) と (b) の差はない．これは変数の選択において，周辺化した分布を用いてはいけないことを意味する．また図 15 の例は二次元入力であるが，入力が高次元になった場合どうすれば良いのであろうか? これを解決するのが ES 法である．

この結果を図 15(c) に表す．図 15(c) の左から，用いる変数の個数 K について，$K = 1$，$K = 2$，$K = 3$，そしてすべての変数を用いた $K = 4$ の結果である．全組み合わせ ($2^4 - 1 = 15$ 通り) の中で，CVE が最も低いものは変数 1，2，3 の組みであり，さきほどの考察の結果とも一致する．変数 4 を含む組み合わせは、その組み合わせから変数 4 を外した組み合わせに比べて識別率が低いことから，ラベルに対してランダムに応答する変数 4 はノイズとなり識別性能を悪化させていることがわかる．CVE は予測誤差最小基準であるので，データの背後にあるモデルであるインディケータを抽出することはできない．しかしながら，状況によっては，CVE を用いても変数選択できることがわかる．

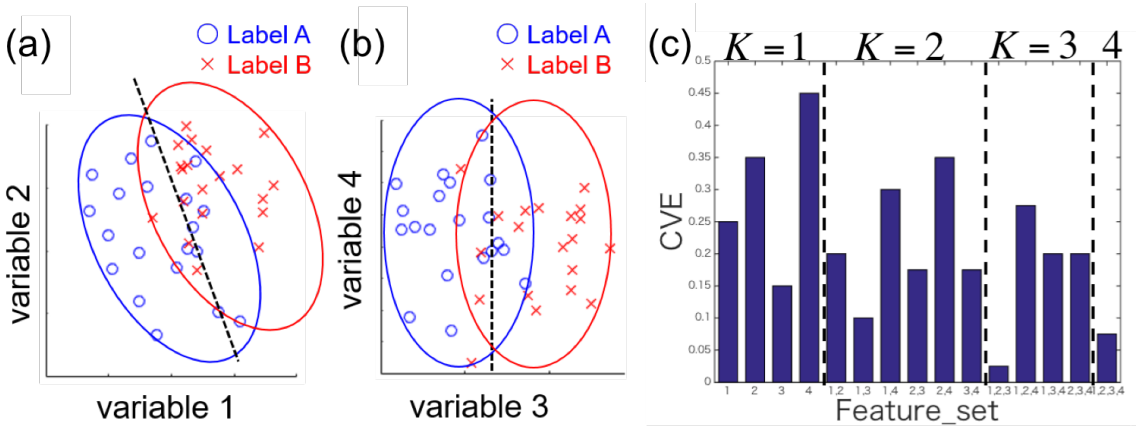


Figure 15: 人工データと ES-SVM を適用した結果．(a),(b)4 次元の人工データ．2 つラベルのデータをプロットしたものである．(c)ES-SVM を適用し，各説明変数の組み合わせとそれに対応する CVE をヒストグラムで表記した．

5.6 ES-DoS の実データへの適用

5.6.1 電気生理学データ - ES-SVM

我々は ES-SVM を，前節で照会した，Eifuku らが行ったサルを対象とした神経生理学実験 [46, 47] によるデータに適用した [48]．CVE が 0 となる変数組み合わせを最適解 c_{opt} とするとき，ES-SVM を用いることで，そのすべての最適解 $\mathbf{c}_{opt}$ を導出することができる．この最適解の解構造を可視化するため，Nagata らはすべての最適解 $\mathbf{c}_{opt}$ をウェイトダイアグラム (WD) を用いて可視化した [48]．図 16 は，個体 1-3 の識別問題（図 16(a)）と個体 1-4 の識別問題（図 16(b)）の識別面を決定するための重み係数 $\mathbf{w}(\mathbf{c}_{opt})$ を可視化した結果である．横軸は最適解のインデックスを表す，長さ $|\mathbf{c}_{opt}|$ のベクトルである．縦軸は変数のインデックスを表す，長さ $|\mathbf{x}|$ のベクトルである．各列では，着色されたセルが，任意の最適解 c_{opt} を構成する変数の重み係数 w_{opt} を示している．色の濃淡は，重みの大きさを，正の値を示す赤色から負の値を示す青色のグラデーションで表現している。黒色のセルは選択されなかった変数を示している．この WD を見ることによって，ある一つの最適な変数の組み合わせだけでは得られない，最適な変数の組み合わせの解構造全体を把握することが可能になる．WD を見ることで，識別に重要な変数は，個体 1-3 の識別問題では L1-LoR で見られたような傾向が改めて確認され，個体 1-4 の識別問題では特定の変数を選択することが難しいことが示唆される．図 16(c)(d) は，個体 1-3 と個体 1-4 に対して，式 (55) の状態密度を計算した結果である．個体 1-3 の方が個体 1-4 に比べて CVE が小さくなる密度が大きくなっていることがわかる．さらに個体 1-4 は CVE が 50%を中心に左右対称に分布していることがわかる．この左右対称な分布は，データに個体 1-4 の識別をする情報がないことを示唆し，L1-LoR や SLR で変数選択がうまくいかなった知見と一致する．

この予想を定量的に示すために，Nagata らはランダムゲスとの比較手法を提案している [48]．M-fold 交差検証法や Leave-one-out 交差検証法では，あたえられたデータそれぞれ一度ずつテストデータで用いられるため，識別の場合，全体での識別回数はデータ数と一致する．そうして得られた CVE を対立仮説とし，帰無仮説を各データを入力によらずランダムに識別することを考える．この場合，誤識別率は以下の二項分布で与えられる．

$$p(k) \quad = \quad \frac{p!}{k!(p-k)!}(0.5)^k(1-0.5)^{p-k}, \tag{60}$$

ここで，p は全体のデータ数，k は誤識別したデータ数を示す．これに，データ数 p をかけることで，ランダムゲスした場合の度数分布 $D_r(E)$ が得られ，DoS と直接比較可能になる．もし，CVE の度数分布 $g(E)$ がランダムゲスした場合の度数分布 $g_{bi}(E)$ と比べて十分に異なれば，与えられたデータが二値分類における情報を有していることを示し，そうでなければ，与えられたデータにそのような情報が組み込まれていないことを示す．

青の棒グラフが DoS を表し，赤の実線がランダムゲスした場合の度数分布 $D_r(E)$ を表す．図 16(c) では，低い誤差率のヒストグラムで大きな違いがあり，ランダムゲスした場合よりも，データを用いた識別が有効に機能していることを示唆する．一方で，図 16(d) では，DoS とランダムゲスした場合の度数分布に大きな違いはなく，1-4 の顔識別の場合，ランダムゲスした場合とほとんど変わらない識別能力しかデータ

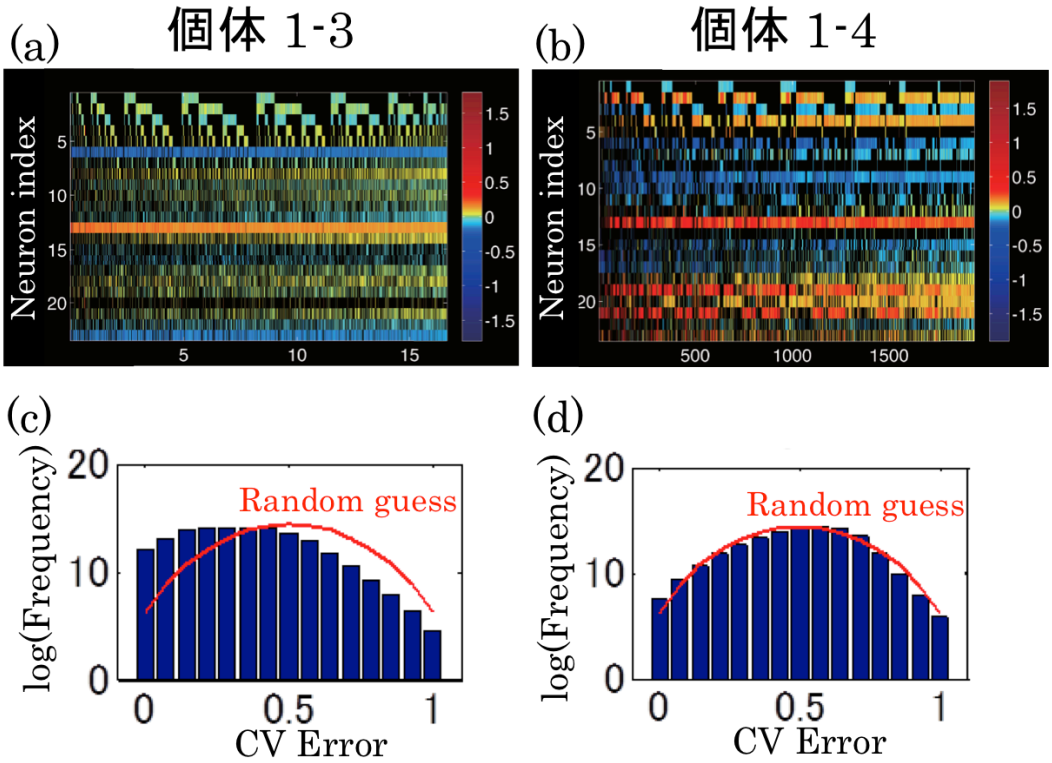


Figure 16: 個体 1-3 のウェイトダイアグラム (a) と状態密度分布 (DoS)(c)．個体 1-4 のウェイトダイアグラム (b) と状態密度分布 (DoS)(d)．ウェイトダイアグラムにおいて，縦軸にニューロン番号，横方向に組み合わせのインデックスを表す．各列で，色のついた部分が CVE=0 となる $\mathbf{c}_{opt}$ を表し，色のグラデーションは $\mathbf{w}(\mathbf{c}_{opt})$ を表す．

が有していないことを示唆する．このように，データに所望の目的に対応する識別能力を有するか否かを DoS を利用すると判別できる．

図 17 は，図 16(b) で検証した Eifuku らの個体 1vs4 のデータに対して AES を適用した結果である．全状態が 8,388,607 通りであるのに対し，サンプリング計算回数を 828,000 と 82,800 にした場合，すなわち，ES に比べて計算量がおよそ 10 分の 1，100 分の 1 にした場合の結果を示す．図 17(a)(c) が推定された状態密度分布を示しているが，どちらの場合でもおおよそ全体の形状を再現できていること，特に，高い誤識別率よりも低い誤識別率のほうが精度がいいことが確認できる．また，図 17(b)(d) に EMC 法によりサンプリングされた CVE=0 を示す変数の組み合わせに対するウェイトダイアグラムを示す．ES に対する結果である図 16(b) と比較して，色のついてる変数などの全体の特徴を捉えられていることがわかり，全状態探索のよい近似を行えていることがうかがえる．

5.6.2 ASD, ADHD の識別のためのヒトの脳活動計測 - ES-SVM

ES-SVM をヒトの脳活動計測に応用した例として，Ichikawa らが行った NIRS（近赤外分光法）を用いた研究を紹介する [51, 61]．NIRS とは，近赤外光が血中ヘモグロビンに吸収される性質を利用し，近赤外光を頭皮上から光ファイバー（照射器）を用いて照射し，その反射光を 2〜3cm 離れた光ファイバー（検出器）計測することによって局所脳血流の相対的変化を計測する手法である．照射器と検出器の間の脳領域をチャネルと呼び，複数の光ファイバを用いることで，活動が期待される部位を含めた広範な範囲を多チャンネル同時計測することが可能である．このとき，計測したチャネルのうち，どのチャネルで血流量が変化したかを同定するというチャネル選択の問題は，変数選択の問題として取り扱われる．

Ichikawa らは，脳血流量の変化から発達障害の鑑別が可能であるかを検討するため，異なる発達障害のある児の脳血流量を測定し，二群で脳血流量が異なったチャネルを選択するために ES-SVM を用いた [51]．脳血流の測定では，自閉症スペクトラム障害（ASD）のある児 8 名と注意欠陥多動性障害（ADHD）のある児 9 名を対象に，母親の顔写真を観察させ，そのときの両側側頭領域の脳血流量の変化を NIRS をもちいて計測した．血流変化は 24 チャネルで計測され，このうちどのチャネルで，ASD と ADHD とで異なる脳血流が見られたかを検討した．従来の NIRS データの解析法として，チャネルごとの t 検定を行ったが，

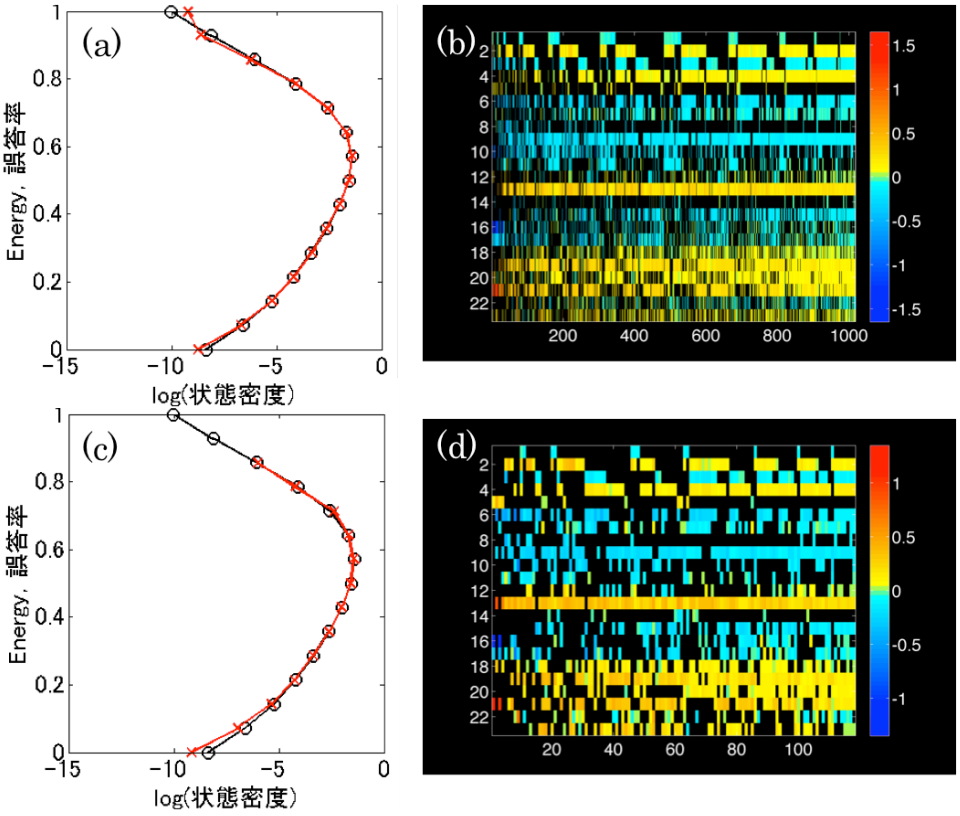


Figure 17: EMC 法とマルチヒストグラム法により推定された CVE の状態密度分布 (DoS)．(a)(b) は，サンプリング計算回数を 828,000 とした結果で，(c)(d) は 82,800 とした結果である．(a)(c) が推定された状態密度分布で，(b)(d) がサンプリングにより得られた CVE=0 となる最適解の組み合わせを表す．

安静時の脳血流に比べて有意に血流が上昇したチャネルは検出されなかった．ES-SVM によって，24 チャネル，すなわち 24 変数のすべての組み合わせ，16,777,215 通りについて，学習機械を SVM，識別性能の評価関数を CVE として，識別面を決定した．その結果，識別率 84%の変数組み合わせを 2 通り選択することができた．さらにこのとき，3.1 節で言及したような，1 変数では CVE が相対的に高い変数であっても，2 変数での識別面では，CVE が小さくなる場合もあった．

5.6.3 津波堆積物判別 - ES-SVM

ES-SVM 法を実データ解析に応用した例として，津波堆積物の地球化学判別について紹介する [52,59]．津波堆積物は津波到達範囲の直接的な証拠になるため，その客観的な認定は地球科学のみならず，社会的にも重要な研究テーマである．津波堆積物の判別は現在に至るまで，堆積学的な観点から行われることがほとんどである．しかし，最も一般的に用いられている粒径（粒径の大きい砂層の存在）を基にした判別方法では，粒径の大きい構成粒子ほど浸水の初期に堆積する傾向にあるため，津波到達範囲の下限しか推定できないという問題点がある．そこで，砂層に加えて粒径の小さな泥層に関しても客観的な津波堆積物の認定を可能にするために，Kuwatani ら [52] では，津波堆積物の地球化学データに注目した．地球化学データは，十から数十の元素からなる多次元の元素含有量からなり，これらの高次元データを有効に活用することが望ましい．そこで，ES-SVM を用いて，津波堆積物と非津波堆積物の判別分析を実施した．

解析に用いた試料は，岩手県から宮城県を経て福島県までの津波浸水域から採取した 129 個の 2011 年東北沖津波堆積物と，東北地方沿岸域に分布する海成堆積層から採取した 75 個の岩石・土壌からなる非津波堆積物である．この研究では，エネルギー分散蛍光 X 線分析装置により得られた，地殻主要元素と重金属元素からなる計 18 元素 (Na, Mg, Al, Si, K, Ca, Ti, Mn, Fe, V, Cr, Ni, Sb, Cu, Zn, As, Cd, Pb) の地球化学データを用いた．試料および化学分析方法の詳細は，土屋ほか [60] などを参考にしていただきたい．解析においては，識別器として線形 SVM，評価基準として 10-fold CV をそれぞれ用いて，18 元素からなる全元素組み合わせ $2^{18} = 262,144$ 通りについて変数選択の全探索を行った．

解析の結果を図 18 に示す．識別率は 1-CVE(%) により定義されるが，92–94%付近で最頻値を持ち，ほ

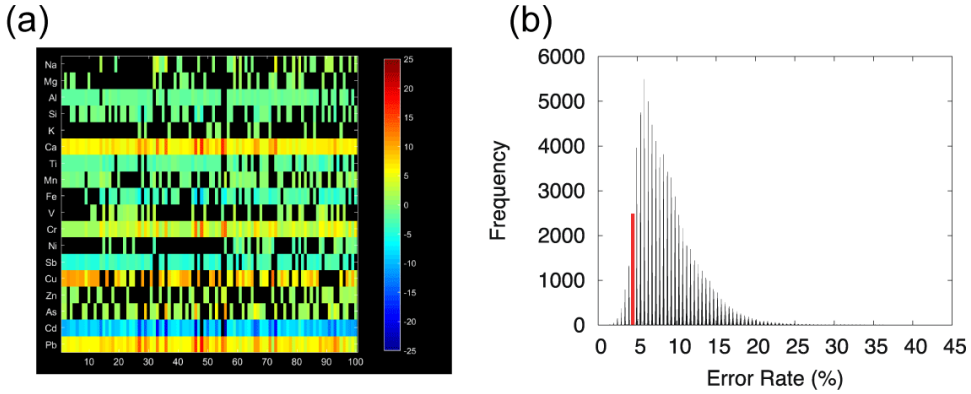


Figure 18: ウェイトダイアグラムと状態密度分布 (DoS)．(a) 識別率の高い元素組み合わせの上位 100 位を示したウェイトダイアグラム．縦軸に元素記号，横方向に組み合わせの順位を示している．各列で，色のついた部分が $\mathbf{c}_{opt}$ を表し，色のグラデーションは $\mathbf{w}(\mathbf{c}_{opt})$ を表す．(b) 全ての元素組み合わせに対する CVE の状態密度分布 (DoS)．赤で示した部分は全 18 元素を用いたときの CVE を示す．

とんどの組み合わせが識別率 80%以上の高い値を示すことが状態密度分布図から見てとれる（図 18b）．最高の識別率は 11 元素 [Al, Ca, Ti, Mn, Cr, Sb, Cu, Zn, As, Cd, Pb] および 12 元素 [Mg, Al, Ca, Ti, Mn, Cr, Sb, Cu, Zn, As, Cd, Pb] の元素組み合わせを用いた際の 100.0%である．識別率が 99%以上のものが 44 通りあり，全体の 0.017%を占めていた．一方，最低の識別率は 57.3%であった．図 18(a) は，識別率が高い上位 100 位までの元素組み合わせの重みベクトルを図示したウェイトダイアグラムである．全体的な特徴から，Al，Ca，Ti，Cr，Cd，Pb などの元素が多くの組み合わせに頻出しており，特に識別に重要であることがわかる．多数の元素を利用している組み合わせが高い識別率を示す一方，18 元素全てを用いた場合の識別率は 95.6%(12,250 位) であり，最高精度を示した組み合わせの場合よりも 4%程度も識別率が劣っていた．これは訓練データに対して過学習していることを示し，未知データに対する汎化性能の悪化に由来する．つまり，高い識別率を実現させるためには，測定した全元素を使用するのではなく，適切な元素の組み合わせを探索することが重要であることを如実に示している．

図 18(a) からは，津波堆積物の地球化学的特徴も読み取ることも可能である．地殻主要元素の中では，Ca が高い正のウェイトを示しているが，これは Ca が海水中や生物由来の海底堆積物に多く含まれる元素であることと調和的である．また，重金属元素の中では，Cu や Pb などが津波堆積物の特徴を示す元素であることがみてとれる．これらの元素は，もともとは鉱山開発により暴露された後に海底堆積物中に蓄積され，再び津波により陸上に戻ってきたものとも考えられており [60]，元素の濃集を引き起こしたメカニズムの詳細な解析が待たれている．以上のように，ES-SVM は単に津波堆積物の高精度判別を可能にするだけでなく，津波堆積物の起源や物理化学プロセスの解明にも貢献するものと期待されている．

5.6.4 天文データへの適用 - ES-LiR

線形回帰を用いた全状態探索 (ES-LiR) 法によるスパース解の網羅的探索法を，Berkeley Supernova Database [62] から得た Ia 型超新星の極大等級の説明変数抽出に適用した [63]．このデータ解析の目的は，線形回帰により Ia 型超新星の極大時の絶対等級に関する推定公式を得ることにある．こうした推定は宇宙における距離を正確に推定するために必要となることから，天文学において重要な推定となっている．我々は前処理を施したデータサンプルとして，78 個の天体について考える [64]．説明変数として，光度曲線の幅 (x_1)，色 (c)，スペクトルから計算された説明変数 (274 個) を用いて，法によるスパース解の網羅的探索法を適用した．

この研究では説明変数が 276 個となるため，全説明変数の組み合わせを評価するのは困難である．そこで，データ解析の目的として，従来から説明変数として用いられているのは光度曲線の幅 (x_1)，色 (c) の二つのみであるため，スペクトルデータを用いることで推定精度向上を行うことであることを考慮して，少ない説明変数の個数 K をもつ組み合わせを全て評価し，説明変数の個数 K を徐々に増加させることで K-スパース全状態探索法を行った．

交差検証誤差やベイズ自由エネルギー (BFE) の網羅的評価によって，K-スパース全状態探索法によって導出された最適な説明変数の組み合わせを導出するだけでなく，網羅的な評価によって LASSO や上記で述べた従来の説明変数の組み合わせを評価した．また，実データ解析では，真の正解は分からないことから，得られたモデルから人工データを作成し，データ解析を行う VMA(Virtual Measurement and Analysis, 仮想計測解析) を行った．その結果，少数のサンプル数により情報の抽出が難しく，が選択されなかった可能

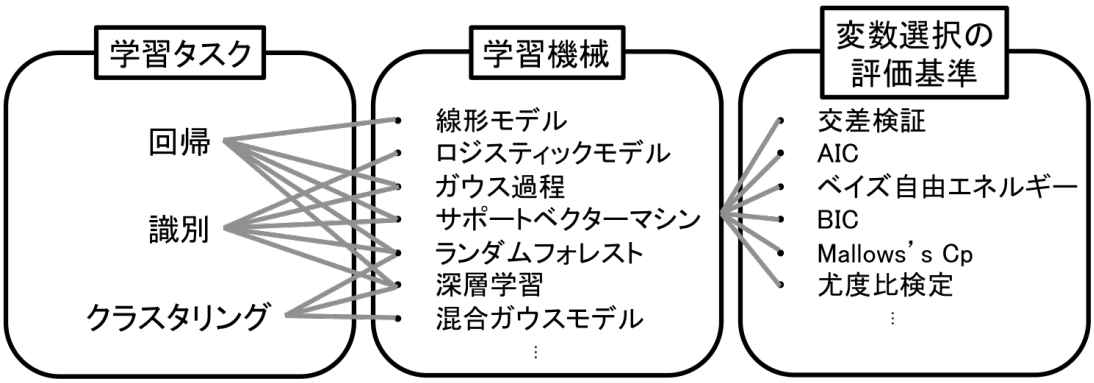


Figure 19: 全状態探索の構造．全状態探索は学習タスク，学習機械，そして変数選択に用いる評価基準の三つ組みに対して定式化される．

性を指摘した．データ数が少ない場合にはどの方法でも不確定性が増す．そのため，それぞれの方法にどういう傾向があるかを理解した上でデータ解析の結果をみる必要がある．その方略のひとつが VMA である．

5.6.5 歩行動作からの自閉性特性の推定 - ES-LiR

線形回帰を用いた全状態探索 (ES-LiR) 法を実データに適用した研究例を紹介する．人の歩き方には個人差があり，歩行動作中の関節の動きを表すバイオロジカルモーションには年齢や性別，性格特性などが表れることが知られている。特に，日常生活で頻繁にみられる歩行場面として，前方から向かってくる他者とすれ違う場面があるが，このときスムーズに身をかわせない行動は，他者がよける方向がわからない場合に顕著にみられるという [65]．重田ら [66] は，他者とすれ違う際の歩行動作には，他者の意図を汲み取ることが苦手な自閉性症状の強さが反映されると考えすれ違い歩行時の身体の動きから，個人の自閉性特性を推定する試みを行った．大学生 28 名を対象とし，まず自閉性特性の強さを質問紙検査によって調べた．次に，身体の 4 部位（腰，右足，左足，頭部）に 3 次元モーションキャプチャを取り付けた状態で 2 人一組となり，互いに向かい合って歩いてきた後にすれ違うという歩行動作を行っている最中の歩行を計測した．計測したパラメータのうち，体の動きの不安定さを表す指標として，歩行ピッチ，腰/右足/左足/頭それぞれについての加速度のノルムの標準偏差（$\omega_{S.D.}$）および角速度のノルムの標準偏差 ($a_{S.D.}$)，という 9 変数を取り出し，自閉性特性の強さを回帰分析によって検討した．このときステップワイズ法による変数選択では，腰の（$\omega_{S.D.}$）だけが説明変数として選ばれた．さらに ES-LiR 法によって，9 変数すべての変数組み合わせについて回帰分析を行い，CVE を評価基準とした変数選択を行ったところ，CVE が低く汎化性能の高い上位 50 組の変数組み合わせすべてに，腰の（$\omega_{S.D.}$）が変数として選択された．さらに，全状態である 511 の回帰式を概観すると，腰の（$\omega_{S.D.}$）を含む回帰式が上位に集中していることがわかった．全状態を探索することによって，ステップワイズ法で選択された唯一の変数が，ほかの変数と比べて回帰に必要な情報を相対的に大きいもつことを確認できた．今後，ES-K 法の適用，BFE の導出によってさらに検討していく．

6 まとめと今後の展開

本チュートリアルをまとめるにあたり，スパースモデリング（スパース変数選択）の歴史を概観する [7]．まず最初に強調すべきことは,「スパースモデリングといえば L1 最適化」と考えるのは早計であるということである．スパースモデリングの歴史は以下のように，説明変数集合の部分集合全てについて調べ上げ，最適な説明変数の組合せを選択する全状態探索法（ES) から始まった．

あらゆる機械学習タスクにとって有効な説明変数を見出すことは重要であり，このプロセスは変数選択や subset selection と呼ばれる [67–69]．変数選択で本質的に必要なことは，どのような基準を用いるかということに加えて，その基準の下で説明変数の組み合わせの全てを評価比較する ES を行うことである [55]．その初期には，stepwise regression 法の提案もあり [70]，ES は不要と言及する文献も見られ [71]，変数選択における ES 法の歴史は 1960 年代初頭，商用計算機の素子として真空管に代わりトランジスタが採用され始めた頃である半世紀以上前にまで遡ることができる [55,72]．変数選択は，長く深い歴史を持つのである．当時から，変数選択の問題ではそれぞれの説明変数を選ぶか選ばないかを考慮すると，どの変数も選ばれない null state を含めて全状態数が 2^N であることは知られていた．二項定理から $2^N = {}_N\mathrm{C}_0 + {}_N\mathrm{C}_1 + \cdots {}_N\mathrm{C}_N$ である．これは，素朴に説明変数の数が少ない状態から調べたり，逆に多い状態から調べたりするだけでは計算量を減らすことはできないことを意味する．しかし，線形回帰のタスクでは，隣接状態に対する計算プ

ロセスを共有したり，探索木の技術を組み合わせたりすることにより，ある程度効率的に ES を行う手法は提案された [73]．しかしながら，Cover と Van Campenhout が示したように，線形であっても計算量は指数オーダーなのである [45] この計算量困難を解決するために，マルコフ連鎖モンテカルロ（MCMC，[74]）法に代表される確率的探索技術が ES に適用されてきた [75–78]．特に Kim らの手法 [78] は，数千にも及ぶ遺伝子のうち，特殊な白血病を識別するのに重要な少数の遺伝子を抽出している．この論文を始めとした論文情報を取得することで，IBM の Watson が女性患者の遺伝子情報から，特殊な白血病であることを識別し，このがん患者の治療の成功に導いた [79]．つまり，ES はスパースモデリングの原点にして state of the art なのである．

このように ES の歴史を振り返ると，§5 で紹介した我々の成果は ES を単に自然科学に応用しただけのように思えるかもしれない．しかしながら，我々はその指摘は当たらないと反論したい．なぜなら，§5.3 で述べたように，我々は ES 法をデータ駆動科学に用いただけでなく，その過程で ES-DoS(状態密度付き全状態探索法) を提案することによりスパースモデリングを深化させたと考えているからである [7, 48]．

科学とは説明変数の探究であり，注目する現象を司る少数の説明変数が見つかるとその問題が解決されたと感じる．対象が複雑になればなるほど，説明変数の特定を巡って激しい議論が繰り返される．ここで改めて考えたいのは，同じスパースモデリングを標榜するといえども，異なるアルゴリズムを用いて得られた異なる結果を突き合わせるだけで議論を収束させることはできるだろうか，ということである．たとえば，日常的に L1 最適化を用いて変数選択をしている研究者が，別の研究者による貪欲法の解析結果を見て異議を唱える場合がある．もしくは，これはアルゴリズムのレベルというよりは計算理論のレベルの問題かもしれないが，対象固有の先見知識により，ある変数はそもそも説明変数にはなりえないとの主張がある場合もある．たしかに，科学はこうした百家争鳴的な議論が花開くことにより前進してきたという側面もあるが，一方では，喧々囂々とした議論の末に，狭隘なセクショナリズムに陥り，科学的な交流が阻害されていることも少なくない．さらに悪いことには，自らが選好する説明変数を導出するアルゴリズムを探し回るという，科学者としては本末転倒なあるまじき行為に走っている可能性があることを自戒の念も込めて言及する．

我々は，ES-DoS を用いることにより，こうした状況を打開することができると考えている．既に説明してきたように ES-DoS は，説明変数としてどの変数を選択するかに関する個々の研究者の仮説全てをインディケータベクトル $\boldsymbol{c}$ に変換し，評価基準 $H(\boldsymbol{c})$ の値に従って DoS を求める．また，LASSO などの近似アルゴリズムをインディケータベクトル生成器とみなすことで，それ以外も含めてすべての近似アルゴリズムも DoS にマップできる [63]．DoS を見ることであらゆる仮説の良し悪しが順位づけされることになる．また，§5.4 で述べた AES を用いることにより，計算量爆発が生じる状況でも DoS の推定をすることは可能である．仮に全人類がそれぞれ異なる仮説を立てたとしても，たかだか世界人口に過ぎない数の仮説 $\boldsymbol{c}$ に対して $H(\boldsymbol{c})$ を計算し DoS を構成することは，仮説の良し悪しを比較することは原理的に可能である．これまでのデータ解析の進め方が，分野に固着したアルゴリズムに支配されていたことにより相反する論が乱立していたとすれば，ES-DoS はそれらを統一的に発展させるアウフヘーベンの試みである．このように，単に最適な説明変数を求める ES と，DoS を伴う ES-DoS は質的に異なる枠組みであるといえる．

本チュートリアルでは線形回帰や線形識別を主な題材に ES-DoS を議論してきた．しかしながら，ES-DoS は図 19 中央に示すように，およそ考えられる全ての学習機械に適用できる [7]．学習機械の説明変数を固定して学習を行う過程を内部ループとすれば，その外部ループとしてインディケータベクトル $\boldsymbol{c}$ を変えながら調べるだけで ES-DoS の手続きは終わりである．また，学習機械によってインディケータベクトル $\boldsymbol{c}$ の次元が莫大になる場合もあるが，その場合も AES-DoS により対応することができる．

通常のアプローチでは，どのような学習機械を用いるかに加えて，L1 最適化の導入や劣モジュラ最適化，平均場近似など多岐にわたるアルゴリズム研究が必要になる．その開発研究のために数年の月日が流れる可能性があるかもしれない．これに対して，同等の時間をかけて ES-DoS を行うことにより，解きたい問題の答えは出るかもしれない．しかもそれは数値的ではあるが厳密解である．科学データの解析に学習機械を用いるのが現代としては常識であり，科学の本質として，説明変数の選択が必須ならば，その普遍的手法として ES-DoS が存在すること，そして ES-DoS が説明変数選択の問題についてアウフヘーベン的側面を持つことを図 19 は示している．

このような牛刀のような議論を用いずとも，説明変数の候補の個数が 20 程度までで，科学的もしくは技術的に解決したい問題は山ほどあるに違いない．歩留まり管理，プロセス管理，予測精度が 1%あがるだけで，その部署の生き死にが決まるというような側面は多々あるはずである．その場合は，まずは既存の方法のアウターループに ES-DoS を適用すれば良いだけである．

本チュートリアルでは，本新学術領域の目的であるデータ駆動科学の創成と，スパースモデリングの深化について，著者の一人の岡田が代表を務める計画研究 B01-2 スパースモデリング班の研究成果を中心として述べた．データ駆動科学を推進する一つの切り口は，筋の良い統計学/機械学習の枠組みをターゲットにし，そこに幅広い分野の問題を射影することである．その枠組みがスパースモデリングであることを述べた．そのような構造を明確にし，スパースモデリングの深化を引き起こすことにより，一網打尽的に科学全体を進めてきた．当該領域申請時に予想した通りに，データ駆動科学が進んでいくことを，4 年半の領域

活動で実証することができた．

さらに重要なことは，このような構造を見出すことにより，情報科学の枠組みを，情報科学の外である自然科学から加速できることを実証できたことである．ES-DoS は統計学・機械学習の本からの演繹的に生まれたわけではない．実データを解析し，その困難の本質を見極めることで帰納的に生まれた枠組みが ES-DoS である．DoS は自然科学者の営みがあるからこそ，その存在価値が高まるのである．自然はいつも人知を超えるものを提示しつづけ，それにより科学は進む．これが物理学の学理の根本の一つである．ES-DoS はその具現化の一つであり，スパースモデリングの深化と高次元データ駆動科学の創成の実例の一つである．

7 謝辞

本研究は JSPS 科研費 新学術領域研究 (No.2512005, 25120009, 16H01552, 16H01555), 若手研究 (B)(17K12735, 17K12749), 特別研究員奨励費 (15J07765) 及び，JST CREST, さきがけ (JPMJPR15E8, JPMJPR1676, JPMJPR17N2) の支援を受けた．また，総合科学技術・イノベーション会議の SIP (戦略的イノベーション創造プログラム)「革新的構造材料」（管理法人：JST）によっても実施された．

References

[1] Kabashima Y, Wadayama T and Tanaka T 2009 *J. Stat. Mech.*, L09003.

[2] Marr D 1982 *Vision* (Cambridge MA: MIT press)

[3] Igarashi Y, Nagata K, Kuwatani T, Omori T, Nakanishi-Ohno Y, and Okada M 2016 *J. Phys. Conf. Ser.* **699** 012001

[4] 大関 2014 新学術領域「疎性モデリング」*2014* 年度チュートリアル講演概要集, 3

[5] 植村 2014 新学術領域「疎性モデリング」*2016* 年度チュートリアル講演概要集, 21

[6] 日野 *in press* 新学術領域「疎性モデリング」*2017* 年度チュートリアル講演概要集

[7] Igarashi Y, Ichikawa H, Nakanishi-Ohno Y, Takenaka H, Kawabata D, Eifuku S, Tamura R, Nagata K and Okada M *under review J. Phys. Conf. Ser.*

[8] 西森 1999 スピングラス理論と情報統計力学，岩波書店.

Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing: An Introduction.* Oxford University Press.

[9] Edwards S and Anderson PW 1975 *Journal of Physics F: Metal*, **5**, 965

[10] Sherrington D and Kirkpatrick S 1975 *Physical Review Letter*, **35**, 1792

[11] 岡田 2009 物性研究, **91**, 427

[12] Hopfield JJ 1982 *Proceeding National Academy of Sciences*, **79**, 2554

[13] Amari S 1977 *Biol. Cybern.*, **26**, 175.

[14] 堀口, 佐野 2000 情報数理物理，講談社.

[15] Okada M 1996 *Neural Networks*, **9**, 1429

[16] Okada M, 2006 *New Generation Computing*, **24**, 185

[17] Amit DJ, Gutfreund H, and Sompolinsky H 1985 *Phys. Rev. A*, **32**, 1007

[18] Amit DJ, Gutfreund, H, and Sompolinsky H 1985 *Phys. Rev. Lett.*, **55**, 1530

[19] Sourlas N, 1989 *Nature*, **339**, 693

[20] Kabashima Y and Saad D 1999 *Europhys. Lett.*, **45**, 97

[21] Kabashima Y, Murayama T and & Saad D 2000 *Phys. Rev. Lett.*, **84**, 1355

[22] Murayama T and Okada M 2003 *J. Phys. A: Math. Gen.*, **36**, 11123

[23] Gardner E 1988 *J. Phys. A: Math. Gen.*, **21**, 257

[24] Cover TM 1965 *IEEE T. elec. comp.* **3**, 326-334

[25] Hosaka T, Kabashima Y, and Nishimori H 2002 *Physical Review E*, **66**, 066126

[26] Mimura K and Okada M 2006 *Phys. Rev. E*, **74**, 026108

[27] Sompolinsky H, Tishby N, and Seung S 1990 *Phys. Rev. Lett.*, **65**, 1683

[28] Tanaka T 2001 *Europhys. Lett.* **54** 540

[29] Tanaka T 2002 *IEEE T. Inform. Theory*, **48**, 2888

[30] Okada M 1995 *Neural Networks*, **8**, 833

[31] Tanaka T and Okada M 2005 *IEEE T Inform. Theory*, **51**, 700

[32] Ishikawa M 1989 *International Joint Conference on Neural Networks*, 1989

[33] 樺島 1996 日本物理学会講演概要集 *(*分科会*)***1996**, 616

[34] Tibshirani R 1996 *J.Royal Stat. Soc. B* **58**, 267

[35] Donoho DL 2006 *IEEE T. Inform. Theory***52**, 1289

[36] 竹田 2014 新学術領域「疎性モデリング」*2014* 年度チュートリアル講演概要集, 47

[37] Yamins D & DiCarlo JJ 2016 *Nat. Neuroscience* **19**, 356

[38] Hassabis D, Kumaran D, Summerfield C, & Botvinick M 2017 *Neuron* **95**, 245

[39] Toya K, Fukushima K, Kabashima Y, & Okada M 2000 *J. Phys. A: Math. Gen.* **33**, 2725

[40] Matsumoto N, Okada M, Sugase-Miyamoto Y, Yamane S 2005 *J. Comput. Neurosci.* **18**, 85

[41] Sugase Y, Yamane S, Ueno S, & Kawano K 1999 *Nature* **400**, 869

[42] Matsumoto N, Okada M, Sugase-Miyamoto Y, Yamane S, & Kawano K 2005 *Cereb. Cortex* **15**, 1103

[43] Reed S 2011 *Science* **331**, 696

[44] 五十嵐, 竹中, 永田, 岡田 2016 応用統計学 **45**, 75

[45] Cover TM & Van Campenhout JM 1977 *IEEE T. Syst. Man. Cyb.* **7**, 657

[46] Eifuku S, De Souza W C, Tamura R, Nisijo H, & Ono T 2004 *J. Neurophysiol.* **91**, 358

[47] Eifuku S, De Souza W C, Nakata R, Ono T, & Tamura R 2011 *PLoS One* **6**, e18913

[48] Nagata K, Kitazono J, Nakajima S, Eifuku S, Tamura R, & Okada M 2015 *IPSJ Online Trans.* **8**, 25

[49] Tomioka R and Muller KR 2010 *NeuroImage* **49**, 415

[50] Yamashita O, Sato M, Yoshioka T, Tong F, & Kamitani Y 2008 *Neuroimage* **42**, 1414

[51] Ichikawa H, Kitazono J, Nagata K, Manda A, Shimamura K, Sakuta R, Okada M, Yamaguchi MK, Kanazawa S, & Kakigi R 2014 *Front. Hum. Neurosci.* **8**, 480

[52] Kuwatani T, Nagata K, Okada M, Watanabe T, Ogawa Y, Komai T & Tsuchiya N 2014 *Sci. Rep.* **4**, 7077

[53] Akaike H 1973, *Proc. 2nd Int. Symp. Inform. Theory*, 267

[54] Stone M 1977 *J. Roy. Stat. Soc. B (Methodological)* **39**, 44

[55] Garside MJ 1965 *J. R. Stat. Soc. Ser. C* **14** 196

[56] Hukushima K & Nemoto K 1996 *J. Phys. Soc. Jpn.* **65**, 1604

[57] Geyer CJ 1991 *Proc. 23th Symp. Interface Comput. Sci. Stat*, 156

[58] Ferrenberg AM & Swendsen RH 1989 *Phys. Rev. Lett.* **63**, 1195

[59] 駒井, 桑谷, 中村, 土屋 2016 電子情報通信学会誌 **99**, 418

[60] 土屋, 井上, 山田, 山崎, 平野, 岡本, 小川, 渡邊, 奈良, 渡邉, 東北地方津波堆積物研究グループ 2012 地質学雑誌 **118**, 419

[61] 市川, 仲渡, 島村, 金沢, 山口, 作田, 柿木 2016 電子情報通信学会誌 **99**, 428

[62] Silverman JM, Ganeshalingam M, Li W, Filippenko AV 2012 *Mon. Not. R. Astron. Soc.* **425**, 1889

[63] Igarashi Y, Takenaka H, Nakanishi-Ohno Y, Uemura M, Ikeda S, & Okada M 2017 *arXiv preprint* arXiv:1707.02050

[64] Uemura M, Kawabata KS, Ikeda S, & Maeda K 2015 *Publ. Astron. Soc. Jpn.* **67**, 55

[65] Honma M, Koyama S, & Kawamura M 2015 *Front. Psychol.* **6**, 1013

[66] Shigeta M, Sawatome A, Ichikawa H, & Takemura H 2017 生体医工学シンポジウム *2017* 2P-24

[67] Beale E M L, Kendall M G, and Mann D W 1967 *Biometrika* **54** 357–366

[68] Miller A J 1990 *Subset selection in regression* Chapman and Hall

[69] George E I 2000 *J. Am. Stat. Assoc.* **95** 1304–1308

[70] Efroymson M A 1960 *Mathematical Methods for Digital Computers* John Wiley and Sons 191–203

[71] Healy M J R 1963 *Comput. J.* **6** 57–61

[72] Kudô A 1963 *Memoirs of the Faculty of Science, Kyushu University. Series A, Mathematics* **17** 63–75

[73] Furnival G M and Wilson R W 1974 *Technometrics* **16** 499–511

[74] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H, and Teller E 1953 *J. Chem. Phys.* **21** 1087–1092

[75] George E I and McCulloch R E 1993 *J. Am. Stat. Assoc.* **88** 881–889

[76] Green P J 1995 *Biometrika* **82** 711-732

[77] Kuo L and Mallick B 1998 *Sankhyā Ser. B* **60** 65–81

[78] Kim S, Tadesse M G, and Vannucci M 2006 *Biometrika* **93** 877–893

[79] ＡＩ、がん治療法助言　白血病のタイプ見抜く，日本経済新聞, 2016 年 8 月 4 日