1F5-OS-06b-3

スパースモデリングとデータ駆動科学

Sparse modeling and data-driven science

岡田 真人
Masato Okada

東京大学 大学院新領域創成科学研究科
Graduate School of Frontier Sciences, The University of Tokyo

独立行政法人 理化学研究所 脳科学総合研究センター
RIKEN Brain Science Institute

First, I propose *three-level description of machine learning*, that is, computational, algorithmic and hardware-implementational levels, which is strongly inspired by the David Marr's tri-level hypothesis. According to the three-level description of machine learning, I introduce Bayesian approach to spectral deconvolution, in which a multimodal spectrum is decomposed into a linear sum of a suitable number of unimodal basis functions, as an example of the sparse modeling. It is based on fundamental principle of sparseness: most of useful information is embedded in the low-dimensional subspace for high-dimensional observation data in the various fields of natural science. In my opinion, the sparse modeling enables us to promote high dimensional data-driven (HD^3) science.

1. はじめに

David Marr は脳の情報処理を理解するために，図 1 の三つのレベルで考えるのが効果的であると述べている [Marr 82, 乾 87, 川人 96]．Marr は彼の著書 "Vision" においてスーパーマーケットのキャッシュ・レジスタを例に挙げ，この三つのレベルは脳の情報処理だけでなく，"情報処理課題を実行する機械"を理解するために有用であると主張している．これに基づけば，Marr の三つのレベルの視点で機械学習を論じることは興味深い．§2. では，機械学習に関する三つのレベルを論じる．

次に，図 2 のような多峰性スペクトルを，ガウス関数のような単峰性の基底関数の線形和に分解するスペクトル分解 [Nagata 12] を図 1 の三つのレベルを元に解説し，それを用いて機械学習の三つのレベルの具体例とする．§3. では，スペクトル分解の計算理論を，分光学の知見を元に構築する．その中で，機械学習の計算理論を構築するには，機械学習の適用分野に対する深い知識が必要であることを述べる．§4. では，スペクトル分解の計算理論に関するベイズ推論を定式化する．この枠組にはベイズ事後確率の多峰性や，それと等価な局所解が存在することを指摘し，§5. では，それを解決するアルゴリズムとして交換モンテカルロ法を紹介する [Hukushima 96]，交換モンテカルロ法の結果を用いると，基底関数の個数を決めることがきる．

§6. では，スパースモデリングを紹介し，それにもとづくデータ駆動科学の創成を論じる [SpM-HD3 13]. スパースモデリングの基本的な考え方は，(1) データの説明変数が次元数よりも少ない (スパース) と仮定し，(2) 説明変数の個数が小さくなることと，データへの適合とを同時に要請することにより，(3) 人手に頼らない自動的な説明変数の選択を可能にする枠組みである．スペクトル分解も無限自由度の関数を，有限個の基底関数で和で表現するので，スパースモデリングの一種であると考えることができる．

連絡先: 岡田真人:okada@k.u-tokyo.ac.jp

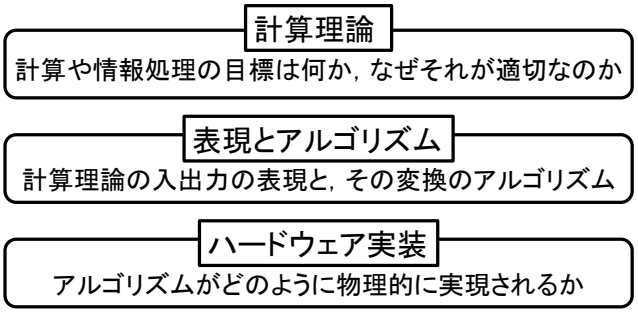

図 1: 機械学習と David Marr の三つのレベル [Marr 82, 乾 87, 川人 96].

2. 機械学習の三つのレベルによる記述

図 1 の第一のレベルの計算理論では，計算や情報処理の目標や，その目標の適切さなどを取り扱う．計算理論のレベルでは，何 (What) をするのか，なぜ (Why) そうしなければならないかを問うのである．例えば，バイオインフォマティクスのマイクロアレイに関する計算理論の一例は，マイクロアレイのデータから，医学的な観点に基づく識別の問題を解くことである．物質科学や材料科学を取り扱うマテリアルズインフォマティクスでは，実験計測のデータや，物質の電子状態に関する大規模計算のデータから，電気伝導度，誘電率，透磁率などを予測することが計算理論の一例になる．そこでは，これらのパラメータの物理的化学的な意味や機序等の物質科学的な学問背景なしには，計算理論を構築することは不可能である．これらの具体例から，機械学習の計算理論を構築するには，機械学習の適用分野に対する深い知識が必要であり，時として，その分野に直接たずさわる研究者が気づかない盲点をつく，鋭い洞察力が必要になるかもしれない．

第二のレベルは、計算理論をどのようにして実現することかを問うアルゴリズムのレベルである．またアルゴリズムの入力と出力の表現は何かという視点も，この第二のレベルで議論

される．バイオインフォマティクスのマイクロアレイの識別に対して，どのような識別器を用いて，どのような学習アルゴリズムを用いるかを議論するのが，第二のレベルの問題である．この例からも明らかなように，一つの計算理論に対して，複数のアルゴリズムや表現が存在することが通常である．これら複数のアルゴリズムの正否は通常，予測誤差やアルゴリズムの実行時間により評価される．しかし，いくら予測誤差が低くても，与えられた計算理論に合致しないアルゴリズムであれば，それはなんの意味も持たない．例えば，図 2 のようなノイジーな観測データが与えられたとしよう．この観測データからノイズを除去するためには，多項式フィッティングするアルゴリズムも考えられるし，スプライン補間も考えることができる．フーリエ基底で展開するということも考えられる．§3. において，図 2 の観測データの情報処理に関する計算理論を考察し，続く §4. において，それにもとづくアルゴリズムを構築する．

第三のレベルは，第二のレベルであるアルゴリズムが，どのようにして物理的に実現されるかを問うハードウエア実装のレベルである．我々は機械学習において，第三のレベルのハードウェア実装が今後の重要課題の一つになると考えている．行列分解やマルコフランダムフィールドモデルのハイパーパラメータ推定に関する研究から，ベイズ推論を行なう際の近似手法である変分法等の近似手法が，計算理論の観点から無視できない系統的誤差を含む可能性があることが示唆されている [Nakajima 12, Nakanishi 14]．例えば，変分ベイズ法では陽に変数選択の事前確率が導入されていなくても，変分ベイズ法の数理的な性質より変数選択が行なわれやすい性質を持つことが知られている [Nakajima 12]．このような性質から類推すると，計算理論の実現のために導入した変数選択の結果として変数が選択されたのか，近似アルゴリズムの副次的効果で変数選択が行なわれたかの区別がつかなくなる可能性がある．この可能性を避けるための最も確実な方法は，現実的な範囲で使用可能な高速な計算機を用いて，モンテカルロ法などのサンプリング手法を用いて愚直にベイズ推論することである．そのためには，並列計算機や GPGPU のハードウェア上の特性を良く理解して，モンテカルロ法をいかに実装するかのテクニックが必須である．この数値的な厳密計算のみが，計算理論を数理的に定式化した枠組の是非を直接的に問える手段である．また，これは変分ベイズ法等の近似手法が，どの程度の精度が保証されるかに関する指針にもなる．

以上から機械学習においても，計算論的神経科学の現状と同様に，これら三つのレベルを俯瞰しながら研究を進めることが今後，重要になると予想している．情報処理の目標として，与えられた時間内に結果を出す必要がある場合は存在する．例えば，惑星探査においては，カメラでとらえた 2 次元画像から小惑星の三次元形状を限られた時間と計算リソースの中で計算する必要がある．そのためには厳密なサンプリング手法の結果を基準として，種々の高速近似手法を，それを実行するハードウェアの制約も考慮に入れて，オフラインで評価しておく必要がある．それをもとに，近似精度と計算時間のトレードオフの観点で，最適な近似手法を予め設定しておくことが重要である．そのようなシステムの設計指針を得るためにも，機械学習の三つのレベルを貫いて近似手法を評価することが必要である．

3. スペクトル分解の計算理論

図 2 のような多峰性スペクトルを，ガウス関数に代表される単峰性の基底関数の線形和で表現するスペクトル分解について議論する [Nagata 12]，

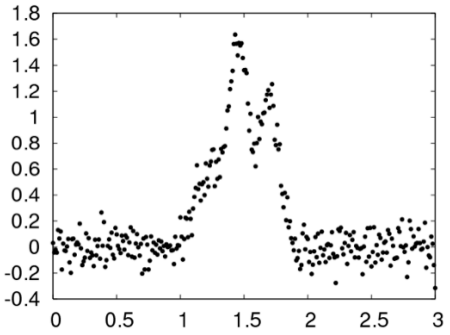

図 2: スペクトル分解

$$G(x;\theta,K) \quad = \quad \sum_{k=1}^{K} a_k \exp\left(-\frac{(x-\mu_k))^2}{2\sigma_k^2}\right). \qquad (1)$$

ここで K はガウス関数の個数であり，μ_k は k 番目のガウス関数の平均値であり，a_k は k 番目のガウス関数の大きさである．σ_k^2 は k 番目のガウス関数の分散に対応する．ここでパラメータセット θ を $\theta = \{a_k, \mu_k, \sigma_k\}_{k=1}^{K}$ と定義する．

分光学においては，このような多峰性スペクトルのスペクトルを構成する個々の単峰性の基底関数は，物質を構成する電子のエネルギー準位等に対応する．このようなスペクトルを観測する意図 (Why) は，測定対象である物質の未知の電子状態を知ることである．この情報処理でやるべきこと (What) は，多峰性スペクトルを単峰性の基底関数に分解することである．その結果，その物質の電子状態が何個のエネルギー準位から構成され，それぞれのエネルギー準位に含まれる電子の状態密度を知ることで，対象の電子状態を知るという情報処理の意図が達成できる．ここでガウス関数の個数 K は多峰性スペクトルのピーク数であり，それは測定対象のエネルギー準位の数に対応する．ガウス関数の平均値 μ_k は k 番目のエネルギー準位の値であり，ガウス関数の大きさ a_k は k 番目のエネルギー準位の電子の状態密度に対応する．ガウス関数の幅に相当する σ_k は，装置の観測誤差を表すとともに，X 線等の入力で励起され空になった k 番目のエネルギー準位が，平衡状態へ緩和していくときの緩和時間を表現している．

このように，第一のレベルの計算理論を構築するためには，情報処理課題が取り扱う対象に対する知識に基づき，データが獲得された意図や，その学問的背景を理解するとともに，それらを数理的に定式化する必要がある．以上の考察から，スペクトル分解の計算理論の数理的な定式化は，図 2 のような多峰性スペクトルを構成する N 個のデータセット $D = \{x_i, y_i\}_{i=1}^{N}$ から，多峰性スペクトルのピーク数に対応する基底関数の個数 K と，それにともなうパラメータセット θ をデータから適切に決めることである．

4. スパクトル分解のアルゴリズム: ベイズ推論の導入

ここでは §2. で述べたスペクトル分解の計算理論を実行するために，まず最小二乗法を考える．N 個のデータセット $D =$

$\{x_i, y_i\}_{i=1}^N$ に対して，以下の平均二乗誤差 $E(\theta, K)$ を定義する，

$$E(\theta, K) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - G(x_i;\theta, K))^2 . \quad (2)$$

この誤差関数 $E(\theta, K)$ を最小化する θ と K を求めるアルゴリズムは，単純に誤差関数 $E(\theta, K)$ を最小にしたければ，ピーク数 K を多くすればよいが，ノイズにまでフィットしてしまい，真のピーク構造と誤った結果を抽出してしまう．これはスペクトル分解の計算理論にそぐわない．さらにもう一つの問題が存在する．仮にピーク数を決めたとしても，誤差関数 $E(\theta, K)$ は一般に局所解を持つことが知られている．

これら二つの問題を解決する方法として，ベイズ推論に基づくスペクトル分解のアルゴリズムを提案する．ベイズ推論では，観測データ y が生成された因果律を確率的に定式化する，まず確率 $p(K)$ に従って，その対象の物質のエネルギー準位の数 K が生成されると考える．つぎにその K に従い，パラメータセット θ が条件付き確率 $p(\theta|K)$ で与えられると考える．ここでは簡単のために $p(K)$ と $p(\theta|K)$ は一様であるとする．観測データ y は，式 (1) に従う真の値に平均 0 分散 1 のガウスノイズ ϵ が重畳されて観測されるとする，

$$y = G(x;\theta, K) + \epsilon. \quad (3)$$

これらの仮定より観測データ y は，ピーク数 K とパラメータセット θ が与えられた元で，入力 x の条件付き確率で生成される，

$$p(y|x,\theta, K) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(y - G(x;\theta, K))^2}{2}\right). \quad (4)$$

それぞれのデータ (x_i, y_i) が独立に得られたとすると，パラメータ θ が与えられた元での，N 個のデータ D の条件つき確率は，

$$p(D|\theta, K) = \prod_{i=1}^{n} p(y_i|x_i,\theta, K) \quad (5)$$

$$\propto \exp(-NE(\theta, K)) \quad (6)$$

となり，$P(D|\theta)$ は誤差関数 $E(\theta, K)$ をエネルギーと見なし，N を逆温度と見なした場合のボルツマン分布に従う．ここで，これまで登場してきた全ての変数の同時確率 $p(D,\theta, K)$ を，以下のように形式的に書き下すことが可能となる，

$$p(D,\theta, K) = p(D|\theta, K)p(\theta|K)p(K) \quad (7)$$

$$\propto \exp(-NE(\theta, K)) \quad (8)$$

この同時確率が存在すれば，ベイズの定理と周辺化の手続きを用いて，データ D と基底関数の個数 K が与えられた場合の，パラメータセットの事後確率 $p(\theta|D, K)$ を形式的に書き下すことは可能である，

$$p(\theta|D, K) = \frac{p(D,\theta, K)}{p(D, K)} \propto \exp(-NE(\theta, K)) \quad (9)$$

$$-\log p(\theta|D, K) = NE(\theta, K) + \text{定数} \quad (10)$$

ここで $p(K)$ と $p(\theta|K)$ は一様であるとことを用いた．式 (9) と (10) より，事後確率 $p(\theta|D, K)$ を最大化する θ を θ の推定値とする最大事後確率推定は，式 (2) の最小二乗法から得られる θ と一致することがわかる．

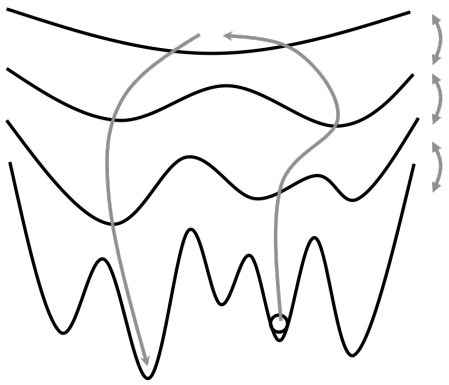

図 3: 交換モンテカルロ法の概念図

データ D が与えられた場合の，基底関数の個数 K の事後確率 $p(K|D)$ は，

$$p(D, K) = \int d\theta p(D,\theta, K)$$

$$\propto \int d\theta \exp(-NE(\theta, K)) \quad (11)$$

$$p(K|D) = \frac{p(D, K)}{p(D)} \propto p(D|\theta, K)$$

$$\propto \int d\theta \exp(-NE(\theta, K)), \quad (12)$$

で与えられる．先ほどと同様に，D が与えたれたもとでのピーク数 K の推定は，最大事後確率推定を用いて，式 (12) の事後確率 $p(K|D)$ を最大化する K を推定値とする．また式 (12) の右辺は統計力学の分配関数に相当する．分配関数を用いて自由エネルギー $F(K)$ を，

$$F(K) = -\log \int d\theta \exp(-NE(\theta, K)), \quad (13)$$

定義する．対数の単調性から，事後確率 $p(K|D)$ を最大化する K と自由エネルギー $F(K)$ を最小化する K は一致する．

5. スパクトル分解のアルゴリズム: 交換モンテカルロ法の導入

§4. のスペクトル分解のベイズ推論を実行する際には二つの困難が存在する．一つは，式 (2) や (10) の誤差関数 $E(\theta, K)$ が局所解を持つために，誤差関数を最小にする θ を求めることが難しいことである．もう一つは，式 (12) や (13) の θ に関する数値積分である．一つのガウス関数に対して，平均値，分散，強度の三つのパラメータを持つので，θ の次元は簡単に 10 を越える．このような状況では，通常の数値積分では不十分である．

これらの問題を，解決するため交換モンテカルロ (EMC) 法を用いる [Hukushima 96]．交換モンテカルロ法はマルコフ連鎖モンテカルロ (MCMC) 法の一つであり，物性物理学のスピングラスと呼ばれる多峰性を持つエネルギーを系を統計力学を用いて数値的に研究する際に提案された手法である．ボルツマン分布の対応から，式 (9) の事後確率 $p(\theta|D, K)$ に逆温度 $\beta = 1/T$ を導入し，

$$p_\beta(\theta|D, K) \propto \exp(-N\beta E(\theta, K)) \quad (14)$$

温度Tが違う系を複数用意し，図3のように同時並列にMCMC法を行う．これら複数の系をレプリカと呼ぶ．シミュレーティッドアニーリングでは，温度Tを高温から低温に徐々に下げて(アニールして)いく．一方，EMC法では各逆温度でMCMC法の途中に，隣り合った温度間で確率的にレプリカを交換することで，アニーリングだけではなく，高温化の効果を取り入れ，局所解から脱却し，効率的に最小解を探索することができる．

以下に示すように，複数温度でモンテカルロ法を行うEMC法の利点は，効率的に最小解が探索できた時点で，ピーク数の決定に必要な自由エネルギー$F(K)$も同時に計算できる点である．ここで逆温度βでの自由エネルギー$f(K;\beta)$を定義すると，

$$f(K;\beta) = -\log \int d\theta \exp\left(-N\beta E(\theta,K)\right), \quad (15)$$

$$F(K) = f(K;1) = \int_0^1 d\beta \frac{\partial f(K;\beta)}{\partial \beta} \quad (16)$$

$$\frac{\partial f(K;\beta)}{\partial \beta} = \frac{\int d\theta N E(\theta,K) \exp\left(-N\beta E(\theta,K)\right)}{\int d\theta \exp\left(-N\beta E(\theta,K)\right)}, \quad (17)$$

となる．式(17)は，自由エネルギーの逆温度βによる微分が誤差関数$E(\theta,K)$の期待値になることを示している．EMC法では，複数温度温度でMCMC法を行っているので，複数温度での誤差関数$E(\theta,K)$の期待値はすでに求まっている．これらを用いて，式(16)の積分を数値的に行うことより，自由エネルギー$F(K)$を数値的に計算できる．

図4は，これまで説明したスペクトル分解のベイズ推論の枠組を，図2のデータに関して適用した結果である．図2は人工データであり，$K=3$を用いてデータは生成されている．図4(b), (c), (d)のそれぞれは，ピーク数を$K=2$，$K=3$，4個と変化させた時の，θの推定値を用いて計算した結果である．図4(a)は，EMC法により数値的に求めた，自由エネルギー$F(K)$のピーク数K依存性である．図に示すようにピーク数が$K=3$のときに自由エネルギーが最小になり，$K=3$が正しく推定された．

この様子を図4(b), (c), (d)を用いて説明する．自由エネルギーは，誤差関数$E(\theta,K)$の期待値に対応するエネルギーと用いたモデルの複雑さをあらわすエントロピーの二つの項から成る．図4(b)の$K=2$では，データをフィットできずに誤差が大きく，エネルギーが高い．一方，図4(d)の$K=4$では，差は$K=3$と同等であるが，パラメータの次元数が高くなったことにより，エントロピーが増大している．その結果，図2(c)の$K=3$のときが，エネルギーとエントロピーの釣り合いがとれたモデルとして選択されている．

6. スパースモデリングとデータ駆動科学

機械学習の計算理論は，対象とするデータを獲得者の目的や意図等のデータの出自に強く依存する．一方，これまでの機械学習のアルゴリズムの応用範囲の広さ等の汎用性の高さを考慮すると，一見多様に見えるデータの背後に，アルゴリズムを普遍的に適用できる計算理論的な背景の存在が示唆される．近年のデータ科学の勃興を背景とし，Scienceでもデータ科学が特集として取り上げられている．その中に，天文学における高次元データ解析手法が，全く対象とスケールが異なる生命科学でも有効に働くことが報告されている[Reed 11]．

この事例を受け流すこと無く，分野を越えたアナロジー／普遍性への探究心とともに，多様な視点を貫く普遍的な原理にもとづく，新たな学問体系を提案する好機ととらえることもできる．我々は，その普遍的な計算理論を高次元データのスパース性に求めた．データの出自は問わずに，与えられたデータが疎に表現できることを指導原理 t としてデータ解析を行ったり，そのデータをスパースに表現できる基底を探ることを計算の目標とするのである．その計算の目標が適切かどうかは，データのスパース化を行った結果を，データ獲得者の学問的背景で判断するしかない．自然科学を応用範囲とした機械学習の計算理論の検証には，計算理論とアルゴリズムのレベルをループする，実験・計測・情報科学の分野融合が必須である．

我々は，そこで用いられるアルゴリズムの総称を，データの背後にある機序のスパース原理による発見を目指して，スパース解析ではなく，スパースモデリングと名付けた．このキーテクノロジーにより，高次元データ駆動科学と称すべき新学術領域の創成を目指している[SpM-HD3 13]．

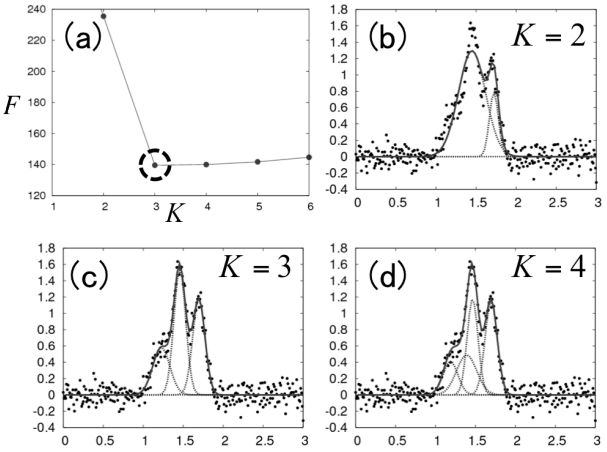

図4: 図2のベイズ推論の結果．(a) 自由エネルギー$F(K)$の値であり，(b),(c),(d)はそれぞれ，ピーク数Kが$K=2$，$K=3$，$K=4$の場合のフィッティング結果である．

参考文献

[Marr 82] David Marr, Vision, MIT press (1982).

[乾 87] 乾 敏郎，安藤広志 訳，ビジョン –視覚の計算理論と脳内表現–，産業図書 (1987).

[川人 96] 川人光男，脳の計算理論，産業図書 (1996).

[Nagata 12] Nagata, Sugita and Okada: *Neural Networks*, Vol. 28, 82-89 (2012).

[Hukushima 96] Hukushima and Nemoto: *J. phys. Soc. Jpn*, Vol. 65, 980- 988 (1996).

[Nakajima 12] Nakajima, Tomioka, Sugiyama, and Babacan: Perfect Dimensionality Recovery by Variational Bayesian PCA, *Neural Information Processing Systems* Vol.25, (2012)

[Nakanishi 14] Nakanishi-Ohno, Nagata, Shouno and Okada: *J. Phys. A*, Vol. 47, 045001 (2014).

[SpM-HD3 13] 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」http://sparse-modeling.jp/

[Reed 11] Reed: *Science*, vol. 331, 696 (2011).