3I3-OS-02a-2

# マッシブデータフローとスパースモデリング

Massive data flow and sparse modeling

岡田真人 *1
Masato Okada

*1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

Data has become highly dimensional due to recent developments in measurement technology, resulting in an explosive increase in computational load. This makes it very difficult to perform modeling based on a hypothesis proposal/verification loop. Sparse modeling (SpM) is a generic term for modeling that has been proposed to resolve such difficulties. Its basic idea is that of a framework in which (1) explanatory variables in high-dimensional data are assumed to be sparse (i.e., fewer than the number of dimensions), and (2) the number of explanatory variables is requested to be made as small as possible while at the same time requesting them to be consistent with the data, thereby facilitating (3) the automatic selection of explanatory variables without manual effort. In this presentation, I introduce an analysis of population dynamics of face-responsive neurons in the inferior temporal cortex as an example of the SpM, and discuss roles of the SpM in the framework of massive data flow.

## 1. はじめに

本講演では，池上と岡により提唱されたマッシブデータフロー (MDF: Massive Data Flow)[池上 12] と，スパースモデリング (SpM: Sparse Modeling) やデータ駆動科学 [SpM-HD3 13, 岡田 14, 永田 15] との関係を述べる．池上は，MDF を基礎科学として膨大なデータをどう扱うかのパラダイムであると述べている．池上は更に，莫大なデータに対して，どうやってデータを「理解」するかといった方法論や認識論が欠如していると述べている [池上 HP]．また池上と岡は，ウェッブの登場が加速した膨大なデータの出現の前後で，トイモデル (toy model: 簡単な仕掛けで本質をつく数理モデル) の説得力が，莫大なデータの前で減弱していると述べている [池上 12]．彼らは，地震のバネブロックモデルや地磁気反転のモデルをトイモデルの例としてあげて，膨大なデータからトイモデルの還元の仕方が難しい場合を指摘している．本講演では，前述の膨大なデータを「理解」するための方法論や認識論と，膨大なデータとトイモデルの接地の問題をデータ駆動科学で取り扱う．

## 2. データ駆動科学と Marr の三つのレベル

我々はデータ駆動科学を，機械学習に代表される情報科学と，そのデータ解析の対象となる学問分野が，図 1 のように緊密に融合した新しい枠組みと方法論であると定義している [永田 15]．ここで重要な概念が，David Marr が提唱した図 2 の三つのレベルである [Marr 82, 乾 87, 川人 96]．図 2 の第一のレベルの計算理論では，情報処理であるデータ解析の目標や，その目標が適切に実行できる科学的根拠を議論する．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．第二のレベルは、計算理論のレベルで議論した方略を，数式として表現し，それを具体的に計算するアルゴリズムを考察するものである．

この計算理論のレベルが，自然科学等のデータ解析の対象の科学と，データ駆動科学を結ぶ要となる．前述の膨大なデータを「理解」するための方法論には二つのステップがある．一つ目はデータ解析の目的と，データ解析が可能な方略を対象の科学の知見から明らかにすることである．次の二つ目のステップは，その目的／方略をベイズ推論や SpM を用いて定式化して表現することである．計算理論と表現・アルゴリズムはそれぞれ，対象の科学と機械学習に代表される情報科学によって担われているので，この目的／方略から表現への変換がトランスディシプリナリーな能力を必要とする．

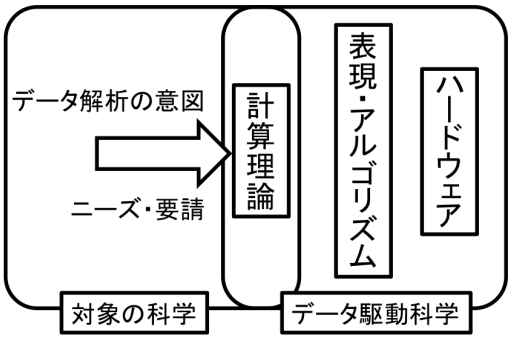


図 1: データ駆動科学と対象の科学 [岡田 14, 永田 15].

## 3. トイモデルとしてのニューラルネット

ここでトイモデルとして，ニューラルネットワークの一種である連想記憶モデルを議論する [Hopfield 82, Okada 96]．連想記憶モデルでは，記憶パターンと呼ばれる高次元ベクトルを，系のアトラクターとして記憶する．ここで，図 3(a) の $\boldsymbol{\xi}_1$，$\boldsymbol{\xi}_2$，$\boldsymbol{\xi}_3$ の三つは，アトラクターとして連想記憶モデルに記憶されているとする．これら三つのアトラクターが比較的近い場所にあるとき，図 3(a) の $\boldsymbol{\eta}$ で示される，三つのアトラクターの中点も自発的にアトラクターになる場合があることが知られている．この中点の $\boldsymbol{\eta}$ を混合状態と呼ぶ．そこで Amari は，この現象を概念形成と名付けた [Amari 77]．この系において記憶パターンと混合状態の力学構造は，図 3(b) のようになる．○で示される安定平衡点である記憶パターン $\boldsymbol{\xi}$ と混合状態 $\boldsymbol{\eta}$ の間に，×で示される不安定平衡点が存在する．この不安定平

連絡先: 岡田真人，okada@k.u-tokyo.ac.jp

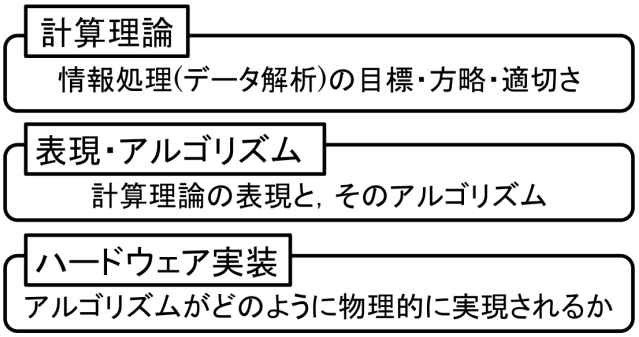


図 2: 機械学習と David Marr の三つのレベル [Marr 82, 乾 87, 川人 96, 岡田 14, 永田 15].

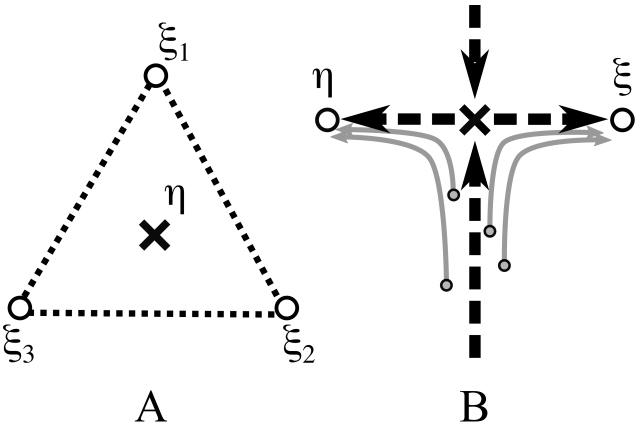


図 3: (a) 記憶パターン $\boldsymbol{\xi}$ と混合状態 $\boldsymbol{\eta}$. (b) 記憶パターン $\boldsymbol{\xi}$ と混合状態 $\boldsymbol{\eta}$ への収束のダイナミクス

衡点はサドルであり，$\boldsymbol{\xi}$ と $\boldsymbol{\eta}$ を結ぶ，不安定多様体が存在する．図 3(b) に示す力学構造には，その他に安定多様体が存在する．系の初期状態が，$\boldsymbol{\xi}$ と $\boldsymbol{\eta}$ から十分はなれていると，系はまず安定多様体にそって，不安定多様体に急速に収束する．次に，不安定多様体にそって，それぞれのアトラクターに収束している．もし脳における記憶の想起やパターン認識が，連想記憶モデルと同様のアトラクターダイナミクスで担われているなら，多数のニューロンの活動の時間変化の背後に，図 3(b) のような力学構造が存在するはずである．

## 4. スパースモデリングによるデータ駆動科学

我々は，トイモデルである連想記憶モデルと，電気生理学のデータとの接地を以下のように SpM を用いて行なった．SpM とは (1) 高次元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．

Sugase らは階層的に分類可能な画像をサルに提示し，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 [Sugase 99]．我々は，この画像の階層的な構造と図 3 の連想記憶モデルのアトラクターの力学的な階層構造に着目し，この Sugase らのデータを SpM を用いて解析している [Matsumoto 05, Katahira 10].

ここでのデータ解析の目標は，Sugase らのデータの背後に，図 3(b) のような力学構造が存在するか否かを確かめる事である．また，その方略として，不安定多様体を抽出するために次元圧縮とクラスタリングを用いた．それぞれを主成分分析と，ベイズ推論を用いた混合正規分布によるクラスター解析で表現した．その結果，Matsumoto らは主成分分析の結果から図 3(b) に示される，不安定平衡点への過渡的な接近を示唆する結果を得るとともに，画像セットに感じる階層的な構造が神経細胞集団のダイナミクスを用いて表現されることを示唆する結果を得た [Matsumoto 05]．Katahira らは，次元圧縮と変数選択を同時に行なう教師なし学習の枠組を提案し，画像を表現するニューロンを自動選択する枠組を提案している．

以上から，SpM にもとづくデータ駆動型アプローチにより，膨大なデータの理解のための方法論や認識論とトイモデルのデータへの接地のアプローチが確立されたことがわかる．

## 謝辞

図 1, 2, 3 の作成に対して，永田賢二氏，村田 伸氏に感謝する．

## 参考文献

[池上 12] 池上高志，岡 瑞起，人工知能学会誌, Vol. 27, No. 4，389-395, (2012).

[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/

[岡田 14] 岡田真人，人工知能学会全国大会 2014，1F5-OS-06b-3, (2014).

[永田 15] 永田賢二，岡田真人，人工知能学会誌，Vol.30, No.2, 209-216, (2015).

[池上 HP] http://massivedataflow.tumblr.com /post/61545851786

[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).

[Hopfield 82] J.J. Hopfield: *Proc. Natl. Acad. Sci. U. S. A.*, Vol. 79(8), pp. 2554-2558 (1982).

[Okada 96] M. Okada: *Neural Networks*, Vol. 9(8), pp. 1429-1458 (1996).

[Amari 77] S.-I. Amari: *Biol. Cybern.*, Vol. 26(3), pp. 175-185 (1977).

[Sugase 99] Y. Sugase et al.: *Nature*, Vol. 400, pp. 869-873 (1999).

[Matsumoto 05] N. Matsumoto et al.: *Cereb. Cortex*, Vol. 15, pp. 1103-1112 (2005).

[Katahira 10] K. Katahira et al.: *J. Phys. Conf. Ser.*, Vol. 233, pp. 012021-1—012021-10 (2010).