# 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　五十嵐康彦　中西（大野）義典　永田賢二

Abstract

スパースモデリング（SpM）は，高次元データに普遍的に内在するスパース性を利用することで，最大限の情報を効率良く抽出できる技術として情報科学分野で大きな注目を集めている．本稿ではSpMの重要性と概要について述べる．まず，SpMの基本的な考え方がケプラー（Kepler）の法則まで遡ることができることを述べ，我々が提案したSpMのアルゴリズムであるSVM全状態探索法（ES-SVM）を紹介する．最後に，データを系統的に取り扱うデータ駆動科学の創成を目指し，その学理の原点としてのデータ駆動科学の三つのレベルを紹介する．

キーワード：スパースモデリング，データ駆動科学，Jim Grayの第4のパラダイム論，ES-SVM，交換モンテカルロ法

## 1. はじめに

「ビッグデータ」という言葉が盛んに使われるようになってきており，様々な場面で大量の高次元観測データを扱うことが増え続けている．本特集では，高次元データに普遍的に内在するスパース性を利用することで必要な情報を効率良く抽出できる「スパースモデリング」について取り上げる．

本稿は，本特集の皮切りとして，スパースモデリングの重要性と概要について述べる．我々は文部科学省科研費新学術領域「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）」において，データを系統的に取り扱うデータ駆動科学の創成を目指し，そのキーテクノロジーとして，スパースモデリングを用いた．表1のように，スパースモデリングの基本的な考え方は（ⅰ）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，（ⅱ）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，（ⅲ）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．

## 2. スパースモデリングとケプラー（Kepler）の法則

ここではビッグデータを，その時代の情報処理機器が辛うじて処理可能な規模のデータと定義する．この定義に従えば，現在使われているビッグデータの概念とは矛盾なく，それぞれの時代にビッグデータが存在することになる．この視点に立てば，人類の長い学問的営みの中で，どのような戦略を持って，ビッグデータを取り扱えばよいかが見えてくると，我々は考えた．その戦略がデータのスパース化である．

先の定義によれば，ケプラーを経由しニュートン力学の発見につながった，Tycho Braheの天体観測のデータは，その当時のビッグデータであった．1609年に発

表1　スパースモデリングの基本的な考え方

| |
|---|
| （ⅰ）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 |
| （ⅱ）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 |
| （ⅲ）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 |

岡田真人　東京大学大学院新領域創成科学研究科
E-mail　okada@k.u-tokyo.ac.jp
五十嵐康彦　東京大学大学院新領域創成科学研究科
E-mail　igayasu1219@mns.k.u-tokyo.ac.jp
中西（大野）義典　東京大学大学院新領域創成科学研究科
E-mail　nakanishi@mns.k.u-tokyo.ac.jp
永田賢二　正員　東京大学大学院新領域創成科学研究科
E-mail　nagata@mns.k.u-tokyo.ac.jp
Masato OKADA, Yasuhiko IGARASHI, Yoshinori NAKANISHI-OHNO, Non-members, and Kenji NAGATA, Member (Graduate School of Frontier Science, The University of Tokyo, Kashiwa-shi, 277-8561 Japan).
電子情報通信学会誌　Vol.99 No.5 pp.370-375 2016年5月

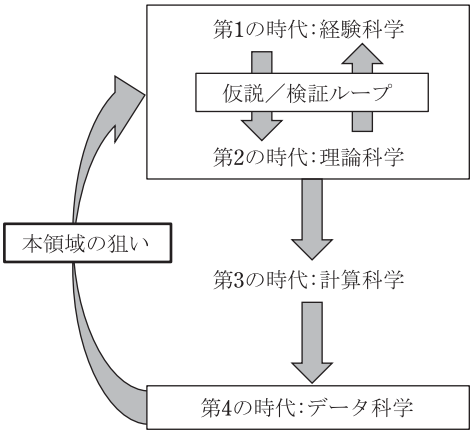



図1　疎性モデリングの立場と Jim Gray の第 4 のパラダイム論との関係

表された第 1 法則と第 2 法則にあるように，惑星は太陽を一つの焦点とするだ円軌道上を動くことをケプラーは示した．更にケプラーは Tycho Brahe の天体観測のデータから，だ円軌道上の面積速度を計算し，それが時間によらず一定であることを示した．つまり，天体観測のデータから面積速度という一つの説明変数を抽出して，それが従う定量的な現象論を抽出したわけである．引き続き 1619 年に発表された第 3 法則にあるように，ケプラーは天体観測のデータから公転周期 $T$ とだ円軌道の長半径 $R$ の二つの説明変数を抽出し，それらの間に，

$$T^2 \propto R^3, \qquad (1)$$

の関係があることを示した．つまり，ケプラーの法則は表 1 の（ⅰ）と（ⅱ）の条件を以下のように満たしている．ケプラーの法則を含む現象論に基づき，ニュートンの運動方程式と万有引力の法則は発見された．スパースモデリングはケプラーの洞察力を，現代の計算科学とデータ科学の力で，表 1 の（ⅲ）の形で援用するものと，我々は考えた．

この考察を，図 1 に示す Jim Gray の第 4 のパラダイム論と比較する．新学術領域「疎性モデリング」では，図 1 で示すように，第 3 のパラダイムの立脚点である計算機の力と，第 4 の科学であるデータ科学での最先端のデータ解析技術を融合する．そして第 1 と第 2 のパラダイムで行われた仮説／検証ループに基づくモデル化を現代的な立場で復興することを目的として，そのキーテクノロジーとしてスパースモデリングを取り上げた．

## 3. スパースモデリングの深化：ES-SVM

ここでは，表 1 の具体的な例として，特徴選択問題を

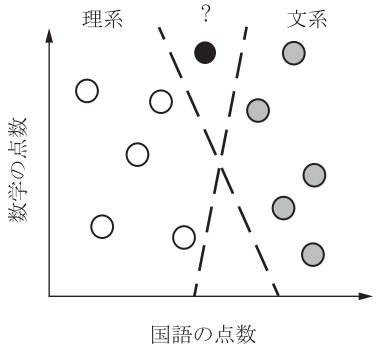



図2　理系と文系の判別

取り上げる．例えば，図 2 のように数学と国語の試験で，理系と文系を分けることにしよう．これは国語の点数を $x$ とし数学の点数を $y$ としたとき，直線，

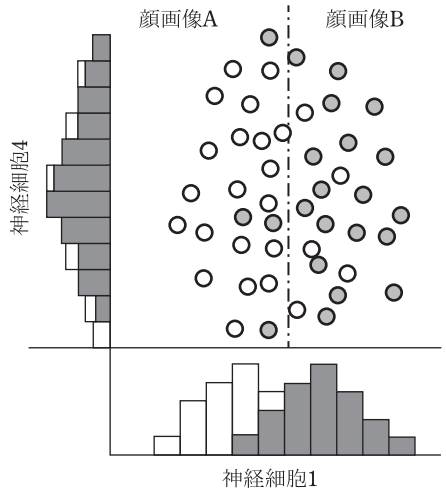
$$ax+by+c=0, \qquad (2)$$

を用いて，理系と文系を判別するわけである．これからの議論では，理系や文系などを識別する際の，数学や国語の点数を特徴と呼ぶことにする．ここで重要な点は，例えば，なぜそこに身長が出てこないかということである．身長は理系と文系を決める特徴であるはずはない．これはこのケースでは正しそうであるが，一般の場合はそうでない．一つの特徴に対して，理系と文系の分布をとり，二つの分布に差がなければ，その特徴は理系と文系の識別に関係がないと判断すればよいと考えられるが，一般にはそれは正しくない．脳科学を例にとり，これを図 3 で説明しよう[(1)]．サルに二人の顔画像を見せて，パターン認識に関係する脳の領野から，4 個の神経細胞を測定したとしよう．この場合は，神経細胞の活動度合いが特徴となる．図 3 のように二次元上で異なる分布（同時分布）を持つ場合でも，片方の次元を無視した一次元の分布（周辺分布）が同じになることがある．この場合，一つの特徴に対して，分布に差がなくても，図 3(b)のようにほかの特徴を組み合わせると，その特徴は識別に有効になることが分かる．つまり，その特徴が識別に有効かどうかは，残りの特徴と組み合わせてみないと分からない．これは，$N$ 個の特徴に対する全ての組合せ数である，

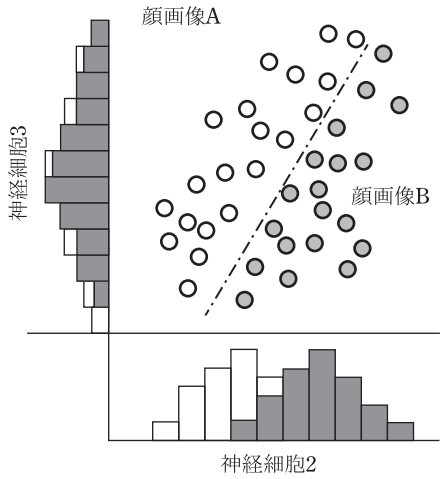
$${}_NC_1+{}_NC_2+\cdots+{}_NC_N=2^N-1 \qquad (3)$$

を試す網羅的探索法が必要であることを意味する．この直感は数学的にも証明されており，Cover と Van Campenhout は，特徴選択問題において，最適な組合せを導出する唯一のアルゴリズムは，考えられる特徴の組合せ全てを探索する網羅的探索法であり，その計算量がデー

タの次元 $N$ に対して $O(e^N)$ となることを示した[2]．仮説やモデルの提案で最も難しいのは，ここでいう特徴，すなわち系を記述する説明変数を決定することであり，




(a) 二つの神経細胞の活動度合いに相関がない場合




(b) 相関がある場合

**図 3　同じ周辺分布を持つ場合でも，同時分布が違う例**

通常の科学では，その分野のエキスパートが，それまでの知見に基づく思索で決定した．生物学，地球惑星科学などの理科第 2 分野に属する，第一原理からの演えきが難しい科学の歴史は，その現象を支配する説明変数を決定する歴史だと言って過言ではない．更に，理科第 1 分野の物理学や化学でも，強相関電子系など複雑な物質の性質の解明において，人手による説明変数の抽出が難しくなっている．

その計算量困難を回避するために，計算量が $O(N^3)$ の Least Absolute Shrinkage and Selection Operator (LASSO)[3] などの高速近似アルゴリズムが提案された．また LASSO の $L_1$ 最適化と網羅的探索法である $L_0$ 最適化の結果が，ある条件では一致するなどの数理的な保証がされ，スパースモデリングが多くの分野で使用されるきっかけとなった．しかしながら，この等価性は特殊な条件下で保証されただけであるため，実データにスパースモデリングを適用する場合は，十分注意する必要がある．

ここでは，その一例として，Nagata らによって提案された，網羅的探索法のアルゴリズムである ES-SVM (Exhaustive Search with Support Vector Machine) を紹介しよう[1]．ES-SVM は識別に関する特徴選択問題という幅広い分野に適用可能なアルゴリズムであるので，本特集の津波識別問題，NIRS のチャネル選択の問題や，脳科学に既に適用されている[1),4),5]．また ES-SVM は SVM などの識別の問題だけでなく，回帰の問題にも容易に拡張できる．ES-SVM では，図 4 の交差検証（CV : Cross Validation）を用いて識別器の予測誤差を推定して，その予測誤差により特徴抽出を行う．CV では計測データを二つに分割して，図 4 の左側の学習データで学習した識別器で，右側のテストデータの誤差である CV 誤差を計算し，これを予測誤差に用いる．図 3(a) では，神経細胞 1 を用いると識別面は垂直になるが，二つの神経細胞を使うと識別面がぶれるので，CV 誤差が大きくなることが分かる．一方，図 3(b) の場合は，二つの神経細胞を使う方が，CV 誤差が小さくな

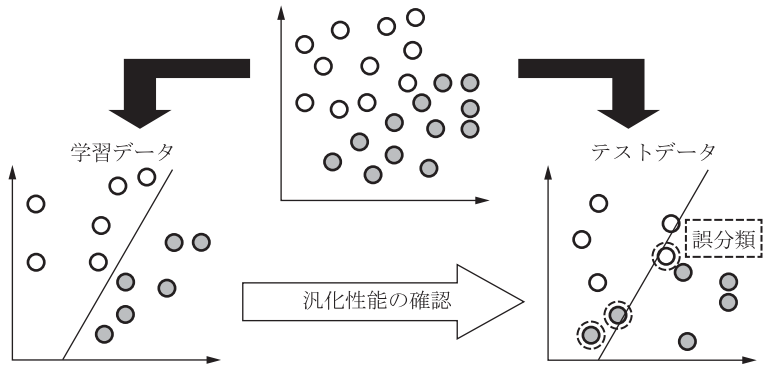
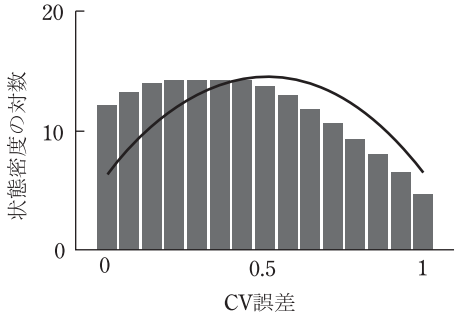
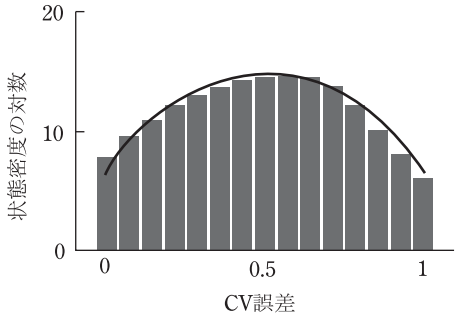



**図 4　交差検証**




(a)近似的スパースモデリングで2値識別が適切に行うことができる場合




(b)行うことができない場合

**図5　CV誤差の状態密度**

ることが分かる．ここで識別器としてSVM（サポートベクトルマシン）を用いた．予想どおり，神経細胞1，2，3を用いた場合が，一番CV誤差が低くなる．これがES-SVMのアルゴリズムである．

NagataらはES-SVMをEifukuらによって測定された側頭葉の神経細胞に適用した[1),6)]．ここでは，4枚の顔画像をサルに提示して，側頭葉の23個の神経細胞の出力を測定している．4枚の顔画像から2枚の顔画像を選んでその2値識別を考え，合計${}_4C_2=6$通りの2値識別に関して，ES-SVMを行った．神経細胞が23個であるので，特徴の組合せの数は$2^{23}-1=8388607$となり，その全てについてCV誤差を求めた．その結果を図5に示す．図5(a)，(b)は，提示した顔画像の種類の違いによる結果を表している．図5の横軸はCV誤差を示し，縦軸がそのCV誤差を与える，特徴の組合せの数である状態数の対数である．この状態数を合計すると$2^{23}-1$となる．図5はCV誤差を与える特徴の組合せの数を与えるので，CV誤差の状態密度と考えることができる．図5(a)と(b)の状態密度は定性的に異なる形をしていることが分かる．図5(a)の二値識別の場合は，LASSOなどの近似的なスパースモデリングは適切に特徴選択を行えるが，図5(b)の場合は，そうでないことが分かっている[1)]．更に，図5の曲線で示される，ランダム識別器を帰無仮説と考えることで，図5(b)のケースでは，23個の神経細胞に識別課題を遂行する情報がないことを示した．更に図5のCV誤差の状態密度が求まれば，現存若しくは今後提案されるはずの全てのスパースモデリングの近似アルゴリズムの性能を，このCV誤差の状態密度上で評価することが可能になるので，ES-SVMはアルゴリズムの性能を評価するメタアルゴリズムとも考えることができる[7)]．ES-SVMの全状態探索が計算量爆発する状況に対して，Nagataらは，交換モンテカルロ法による解のサンプリングと，交換モンテカルロ法とマルチヒストグラム法を用いた状態密度の推定法であるAES-SVM（Approximated Exhaustive Search with Support Vector Machine）法を提案した[1)]．また我々は，ES-SVMから特徴選択問題での誤差の状態密度の推定が重要であることに気付き，情報統計力学的手法で，その数理的な研究を開始している[8)]．

ここで実データ解析が，アルゴリズム解析に与えるインパクトについて述べたい．我々はまず，図5(b)に対して，LASSOなどの近似アルゴリズムが，うまく動かないことを発見して，その数理的なメカニズムを探る過程で，ES-SVMを提案した．自然科学の実データ解析をする過程で，必然的にスパースモデリングを改良しなければならない状況が生じたのである．そこで我々がとった戦略は，現存の近似アルゴリズムのパラダイムにのり，アルゴリズムを複雑化することではなく，CoverとVan Campenhoutに代表される特徴選択問題の原理的な性質から，それまでとは違う考え方でアルゴリズムを構築することであった．これは，神経細胞のデータという実データが存在し，そこからパターン認識をつかさどる神経細胞を抽出したいという，データ解析の目的が明確であったからである．更に重要なことは，神経細胞の数が23個であったことにより，厳密なアルゴリズムである網羅的探索法で，その目的が達成されたからである．これは，どのアルゴリズムを用いるかが，どのような目的と状況で，データ解析をするかに，強く依存することを意味している．これを，データ解析の目的から切り離して，ある数理的なパラダイムの視点だけから達成することは難しい．なぜなら，$N=23$で網羅的探索を行うことは，アルゴリズムのレベルではトリビアルだからである．このトリビアルなES-SVMがなければ，AES-SVMが生まれず，誤差の状態密度に注目した情報統計力学的研究につながらなかったことを強調したい．

## 4. データ駆動科学の三つのレベル

3.で述べたような，実データに基づく実証的な研究と，それに基づく数理的な研究のループ構造を繰り返すことこそ，データ駆動科学の本質であると我々は考える．我々は，その本質を具象化するために，David Marrが指摘した三つのレベルが，重要な着眼点となる

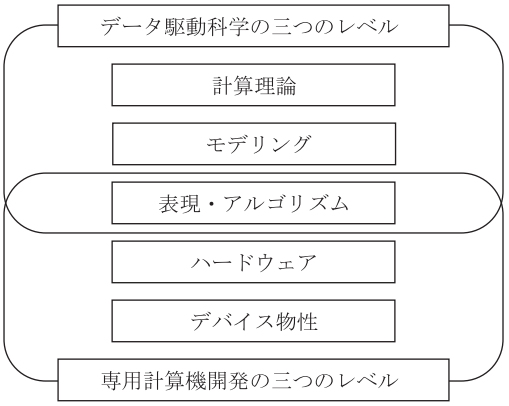



図 6　データ駆動科学と専用計算機開発の三つのレベル

ことを確信した[9]．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付いた．図 6 に示す，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”と名付け，データ駆動科学の学理の原点に位置付けた．第 1 のレベルの計算理論では，データ解析の目的，その目的を達成するための戦略の論理である方略，これら目的と方略の適切さなどを，データが獲得された対象の科学の知見と計測機器に関する計測科学の知見から議論する．表現・アルゴリズムのレベルは，機械学習などのデータ解析手法に対応する．計算理論と表現・アルゴリズムの間に，モデリングのレベルが存在する．そこでは，自然言語で表現される計算理論の目的，方略，適切さを，対象をモデル化することで数学的に表現する．我々は，データを解析する際に，現在どのレベルの議論をしているかを認識し，今の議論は残り二つのレベルでの議論と矛盾しないかを考えることが，重要であると考えた．その具象化がデータ駆動科学の三つのレベルである．

3. の例では，パターン認識をつかさどる神経細胞を抽出することが，データ解析の目的であった．それを実現する表現は特徴選択問題であり，それを最も忠実に実行するアルゴリズムが網羅的探索である．この計算理論から表現・アルゴリズムへの変換がモデリングのステップである．表現・アルゴリズムの議論では，網羅的探索が計算量困難であるので，通常はその近似アルゴリズムを使うが，実際の観測データが近似アルゴリズムの適用可能性を満たしていない場合が考えられるので，それは一般に適切ではない．そこで $N=23$ であれば，近似を使わない網羅的探索が可能であることに気付くという流れである．その中で，近似アルゴリズムの限界を CV 誤差の状態密度上で議論することが可能であり，更に情報統計力学的な解析の提案など，情報科学の観点でのスパースモデリングの深化が可能となった．データ駆動科学の三つのレベルは，計算理論を構成する対象の科学と計測科学，モデリングを担う理論物理学や数理科学，データ解析を担う情報科学を有機的につなげるので，多くの学問の融合する際の戦略のロールモデルの一つである．

## 5. まとめと今後の発展

本稿は，本特集の皮切りとして，スパースモデリングの重要性と概要について述べた．2. では，スパースモデリングの基本的な考え方が，ケプラーの法則まで遡ることができることを述べ，これらの考察から，Jim Gray の第 4 のパラダイム論とデータ駆動科学との関係を述べた．また，4. ではデータ駆動科学の三つのレベルを解説した．これら 2. と 4. の議論は，本特集での，脳生命科学，地球惑星科学，医療，情報通信工学分野などの応用に深く関係している．3. では，ES-SVM を具体例として，スパースモデリングがどのような形で深化するかのロールモデルを説明した．これは本特集でのスパースモデリングの基本理論と基盤技術，情報通信工学分野などの応用に関係している．

本稿の結びとして，本特集 4-1（藤澤克樹氏著）でも議論されているスパースモデリングのハードウェア実装についての私見を述べる．ハードウェア実装と，ハードウェアのデバイス特性は，図 6 のような階層構造を持つ．近年，網羅的検索を含む計算困難な問題を，レーザ，CMOS，量子デバイスなどのアナログ専用計算機で取り扱う機運が急速に高まっている[10]～[12]．その際にも，図 6 のようなレベル記述を行い，ほかのレベルとの親和性を持ちながら研究していくことが必要であると，我々は信じている．

**謝辞**　本稿は，科学研究費補助金新学術領域研究（25120001，25120009），JST-ERATO 岡ノ谷情動情報プロジェクト及び，JST さきがけ領域「社会的課題の解決に向けた数学と諸分野の協働」の支援を受けたものである．

### 文　　献

（1） K. Nagata, J. Kitazono, S. Nakajima, S. Eifuku, R. Tamura, and M. Okada, “An exhaustive search and stability of sparse estimation for feature selection problem,” IPSJ Trans. Math. Modeling Appl., vol. 8, no. 2, pp. 23-30, 2015.
（2） T.M. Cover and J.M. Van Campenhout, “On the possible orderings in the measurement selection problem,” IEEE Trans. Syst., Man Cybern., vol. 7, no. 9, pp. 657-661, 1977.
（3） R. Tibshirani, “Regression shrinkage and selection via the lasso,” Journal of the Royal Statistical Society : Series B, vol. 58, no. 1, pp. 267-288, 1996.
（4） T. Kuwatani, K. Nagata, M. Okada, T. Watanabe, Y. Ogawa, T. Komai, and N. Tsuchiya, “Machine-learning techniques for geochemical discrimination of 2011 Tohoku tsunami deposits,” Scientific reports, vol. 4, no. 7077, pp. 1-6, 2014.
（5） H. Ichikawa, J. Kitazono, K. Nagata, A. Manda, K. Shimamura, R.

Sakuta, M. Okada, M.K. Yamaguchi, S. Kanazawa, and R. Kakigi, “Novel method to classify hemodynamic response obtained using multi-channel fNIRS measurements into two groups : Exploring the combinations of channels,” Frontiers in human neuroscience, vol. 8, no. 480, pp. 1-10, 2014.
（6） S. Eifuku, W.C. De Souza, R. Tamura, H. Nishijo, and T. Ono, “Neuronal correlates of face identification in the monkey anterior temporal cortical areas,” Journal of neurophysiology, vol. 91, no. 1, pp. 358-371, 2004.
（7） Y. Igarashi, K. Nagata, T. Kuwatani, T. Omori, Y. Nakanishi-Ohno, and M. Okada, “Three levels of data-driven science,” to be published in J. Phys. : Conference Series.
（8） Y. Nakanishi-Ohno, T. Obuchi, M. Okada, and Y. Kabashima, “Sparse approximation based on a random overcomplete basis,” to be published in J. Stat. Mech.
（9） D. Marr, Vision, MIT press, Cambridge, MA, 1982.
（10） S. Utsunomiya, K. Takata, and Y. Yamamoto. “Mapping of Ising models onto injection-locked laser systems,” Opt. Express, vol. 19, no. 19, pp. 18091-18108, 2011.
（11） M. Yamaoka, C. Yoshimura, M. Hayashi, T. Okuyama, H. Aoki, and H. Mizuno, “24.3 20k-spin Ising chip for combinational optimization problem with CMOS annealing,” 2015 IEEE International Solid-State Circuits Conference-（ISSCC）, pp. 1-3, San Francisco, USA, Feb. 2015.
（12） M.W. Johnson, M.H.S. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A.J. Berkley, J. Johansson, P. Bunyk, E.M. Chapple, C. Enderud, J.P. Hilton, K. Karimi, E. Ladizinsky, N. Ladizinsky, T. Oh, I. Perminov, C. Rich, M.C. Thom, E. Tolkacheva, C.J.S. Truncik, S. Uchaikin, J. Wang, B. Wilson, and G. Rose, “Quantum annealing with manufactured spins,” Nature, vol. 473, no. 7346, pp. 194-198, 2011.

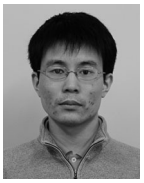
（平成 27 年 12 月 2 日受付　平成 28 年 1 月 15 日最終受付）

**岡田　真人**

昭 60 阪市大・理・物理卒．平 9 阪大大学院博士課程了．同年同大学・基礎工・助手．スパースモデリングによるデータ駆動科学の研究に従事．現在，東大大学院新領域創成科学研究科教授．博士（理学）．平 25 年度から文部科学省科研費・新学術領域研究「スパースモデリング」領域代表．

**五十嵐　康彦**

平 21 東大・工・計数卒．平 23 東大大学院新領域創成科学研究科複雑理工学修士課程了．平 26 同博士課程了．現在，東大大学院新領域創成科学研究科特任研究員．スパースモデリングを用いて，神経科学，地球科学などの自科学を対象にしたデータ駆動科学に関する研究に従事．博士（科学）．



**中西（大野）　義典**

平 23 東大・理・物理卒．平 25 東大大学院新領域創成科学研究科複雑理工学修士課程了．現在，同大学院博士課程在学中並びに学振特別研究員（DC1）．統計力学の手法によるデータ駆動科学の研究に従事．修士（科学）．

**永田　賢二**（正員）

平 16 東工大・工・情報卒．平 18 東工大大学院総合理工学研究科知能システム科学修士課程了．平 21 同博士後期課程了．現在，東大大学院新領域創成科学研究科助教並びに JST さきがけ研究員．MCMC 法によるデータ駆動科学の研究に従事．博士（工学）．