

データ駆動科学と計算物質科学の接点



福島孝治

東京大学 大学院総合文化研究科,
物質・材料研究機構 (NIMS)

2016 年 11 月 29 日

京都生まれ 京都市内で生まれるも、育ちは日本海側

1987-1991 筑波大学 第一学群 自然科学類

- 物理はカッコいいと思って、憧れだけで物理学を目指す
- 水が氷になることの「むずかしさ」を理解して、統計物理へ

1991-1996 筑波大学 物理学研究科 大学院生

- 本格的に計算機を使った物理学の研究をはじめる
- ランダムスピン系の統計力学的研究
- 拡張アンサンブル型のモンテカルロ法の提案

1996-2002 東京大学 物性研究所 助手 (高山研) (六本木から柏へ)

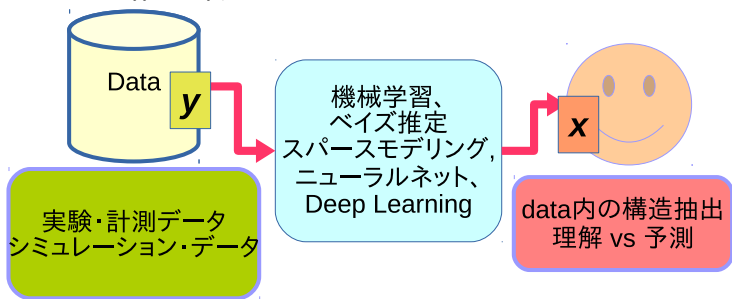
- スピングラスの相転移理論 (カイラル秩序, カオス ...)
- エージング現象, 非平衡ダイナミクス, 自由エネルギー計算
- 特定領域研究「情報統計力学」SMAPIP(代表 田中和先生 (東北大), 2001-2005)

2002-現在まで 東京大学 大学院総合文化研究科 准教授

- 相転移論一般・最適化問題の相転移など... ガラスにも興味を..., データ駆動科学の方法論
- 特定領域研究 DEX-SMI(代表 樺島先生 (東工大), 2006-2009)
- 新学術領域「スパースモデリング」(代表 岡田先生, 2013-)
- 国立研究法人 物質・材料研究機構 (NIMS) @つくば 兼任

データ駆動科学 + 計算物質科学 = データ駆動物質科学??

- ビッグデータ解析は広く一般社会で興味を持たれている
- 自然科学の問題でも、近年の高精度な実験・計測，さらに数値計算は大容量データを作り出す



- 大量のデータから隠れている (数理) 構造を抽出すること。
- 伝統的に前向きに理解の方向とは逆アプローチと言えるかもしれない
- 一つの有力な戦略が、機械学習の技法を用いること。

THE U.S. MATERIALS GENOME INITIATIVE

"...to discover, develop, and deploy new materials twice as fast, we're launching what we call the Materials Genome Initiative"
— President Obama, 2011

Meeting Societal Needs

Advanced materials are at the heart of innovation, economic opportunities, and global competitiveness. They are the foundation for new capabilities, tools, and technologies that meet urgent societal needs including clean energy, human welfare, and national security.



Clean Energy

Human Welfare

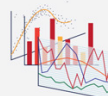
National Security

Accelerating Our Pace



Building Infrastructure for Success

The MGI is a multi-agency initiative to renew investments in infrastructure designed for performance, and to foster a more open, collaborative approach to developing advanced materials, helping U.S. institutions accelerate their time to market.



Also in Japan (2015.6)

JSTイノベーションハブ構築支援事業「情報統合型 物質・材料開発イニシアティブ」が採択 NIMSを拠点にデータ科学をフル活用した材料開発を推進

2015.06.11

[前の記事](#) [一覧に戻る](#) [次の記事](#)

国立研究開発法人 物質・材料研究機構

このたび、JST(科学技術振興機構)「イノベーションハブ構築支援事業」にNIMSが拠点実施機関として提案したイノベーションハブ「情報統合型物質・材料開発イニシアティブ」の採択が決定しました。

概要

情報統合型物質・材料開発とは、従来の物質・材料科学とデータ科学とを融合させたまったく新しい材料開発手法です。膨大なデータ群の蓄積と、ビッグデータ解析の一種である機械学習など、最先端の情報科学を駆使した解析を組み合わせ、新規物質・材料を探索します。本拠点では、産学官の密接な連携によりいち早く革新的な磁性材料や蓄電池材料などを開発し、社会実装することを目指しております。

当機構としては、既に昨年10月より機構内組織として「マテリアルズ・インフォマティクス・プラットフォーム」を設置し、本分野における体制を強化してきたところですが、今回の採択を受けて、クロスアポイントメント制度等を活用しながら産学官の人材糾合を図る等一層の体制充実を図り、我が国における情報統合型の物質・材料研究のイノベーションハブとなるための取組を進めてまいります。



関連ファイル・リンク

[プレスリリース詳細\(PDF\)](#) pdf: 325KB

本件に関するお問い合わせ先

(事業内容に関すること)

国立研究開発法人 物質・材料研究機構
企画部門企画調整室
河西 純一
TEL: 029-859-2000
E-Mail: KASAI.Junichi=nims.go.jp
([=] を [@] にしてください)

(報道担当)

国立研究開発法人 物質・材料研究機構
企画部門 広報室
〒305-0047 茨城県つくば市千現1-2-1
TEL: 029-859-2026
FAX: 029-859-2017
E-Mail: pressrelease=ml.nims.go.jp
([=] を [@] にしてください)

似たキーワードを含むニュース

2007.12.14

Also in Japan (2015.6)

MI²I @ NIMS

“Materials Research by Information Integration” Initiative (MI²I) of the Support Program for Starting Up Innovation Hub, Japan Science and Technology Agency.

- ① Basic Materials measurements
- ② First-principle calc.
- ③ Information techniques

⇒ New materials design

Also in Japan (2015.6)

MI²I @ NIMS

“Materials Research by Information Integration” Initiative (MI²I) of the Support Program for Starting Up Innovation Hub, Japan Science and Technology Agency.

MI² =

- ① Basic Materials measurements
- ② First-principle calc.
- ③ Information techniques

⇒ New materials design

Also in Japan (2015.6)

MI²I @ NIMS

“Materials Research by Information Integration” Initiative (MI²I) of the Support Program for Starting Up Innovation Hub, Japan Science and Technology Agency.

MI² = Mission Impossible 2

- ① Basic Materials measurements
- ② First-principle calc.
- ③ Information techniques

⇒ New materials design

Outline

- ① 「スパースモデリング」の考え方
- ② 今後の展開にむけて
PCA を例に

Outline

① 「スパースモデリング」の考え方

② 今後の展開にむけて PCA を例に

新学術領域研究「スパースモデリング」岡田代表 (東大新領域)

2013-, <http://sparse-modeling.jp/>

Sparse Modeling

English お問い合わせ サイトマップ

文部科学省科学研究費補助金「新学術領域研究」平成25年度～29年度
 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
 Initiative for High-Dimensional Data-Driven Science
 through Deepening of Sparse Modeling

ホーム

領域代表のあいさつ

領域概要

総括班

計画研究

公募研究

活動記録

領域代表のあいさつ

より深く自然を知りたい。未知への飽くなき探究心が、とどまることを知らない計測技術の向上をうみ、我々は大量の高次元データを手に入れることができるようになりました。さらに、生命情報科学の誕生のように、データからの効率的な情報抽出を目指して情報科学技術の知見を集学的に活用することが新たな研究領域を次々に生み出し、「データ科学へのパラダイムシフト」論を生んでいます。

我々は、普遍的な視点でデータ科学を議論することには、以下のような利点が存在すると考えています。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の異なる生命科学でも有効に働くような状況に遭遇します(Science, Feb. 11, 2011)。こうした多様な視点の導入は様々な場面で革新的展開を生み出す原動力となっています。さらに普遍的な視点は、天文学と生命科学のような分野を越えたアナロジー／普遍性への探究心を深め、結果、そうした原理にもとづく新しい解析法の発展につながります。



これらの利点を活かすためには、実験・計測・情報科学の分野融合が必要で、我々は、その方法論構築のため

モデリングのための三層構造

計画研究の構成

実験・計測グループA01,A02

医学



生命科学



脳科学



地球科学



惑星科学



天文学



モデリンググループB01

B01-1
計測モデリング



B01-2
スパースモデリング

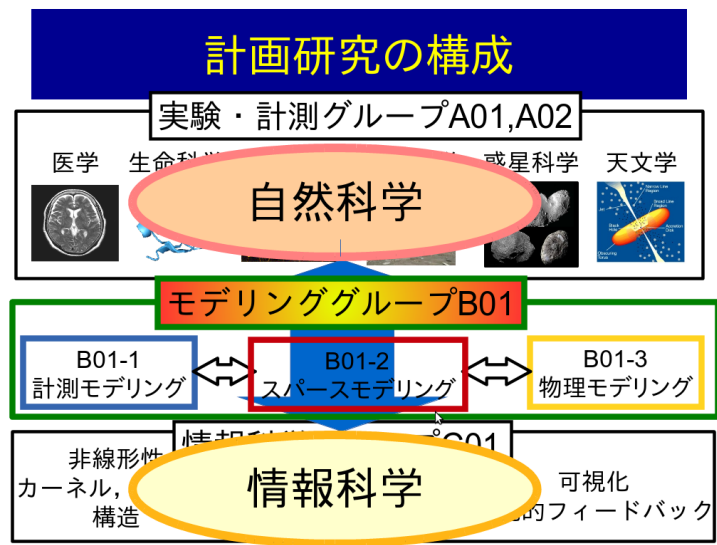


B01-3
物理モデリング

情報科学グループC01

非線形性
カーネル, 階層性, セミパラメトリック推論, 可視化
構造 大自由度性 統計力学 視覚的フィードバック

モデリングのための三層構造



チュートリアル・公開シンポジウムのご案内

📌 ホーム
📌 領域代表のあいさつ
📌 領域概要
📌 総括班
📌 計画研究 >
📌 公募研究 >
📌 活動記録 >
📌 技術支援
📌 研究業績 >
📌 ニュースレター

❖ 2016年度チュートリアル講演会・公開シンポジウム

以下のようにチュートリアル講演会ならびに公開シンポジウムを開催予定です。参加費は無料ですが、会場の収容人数の都合上、こちらの登録フォームから参加登録の手続きをお願いいたします。定員に達し次第、参加登録は締め切らせていただきます。あらかじめご了承ください。

チュートリアル講演会と公開シンポジウムで開催場所が異なります。ご注意ください

日時：2016年12月18日（日）『チュートリアル講演会』

「スパースモデリングの深化と展開」

場所：東京工業大学 すすかけ台キャンパス すすかけホール（アクセス）（会場）

日時：2016年12月19日（月），20日（火）『公開シンポジウム』

場所：慶應義塾大学三田キャンパス 北館ホール（アクセス、会場）

注)12月18日(日)のチュートリアル講演会の会場周辺には昼食を取れるレストラン、食堂はほとんどありません。また、休日開催のため、大学生協、カフェテリアも閉まっています。昼食は必ずご持参いただけますようよろしくお願いいたします。

プログラム

12月18日（日）

チュートリアル講演会

「スパースモデリングの深化と展開」

09:45-10:00

福島孝治（東京大学大学院総合文化研究科）

「オープニング」

10:00-11:30

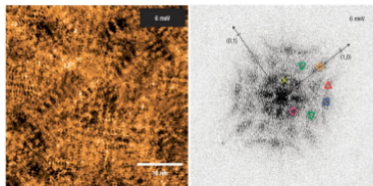
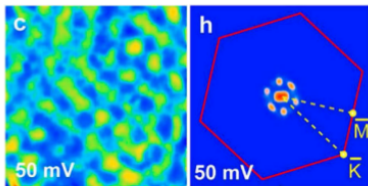
福水健次（統計数理研究所）

「カーネル法と非凸最適化の基礎と応用」

走査トンネル分光測定の高速度・高精度化

A

走査トンネル分光

電子状態密度(dI/dV)像とそのフーリエ変換像超伝導体 $\text{Ca}_{2-x}\text{Na}_x\text{CuO}_2\text{Cl}_2$ トポロジカル絶縁体 Bi_2Te_3 [Hanaguri et al., *Nat. Phys.* 07] [Zhang et al., *Phys. Rev. Lett.* 09]

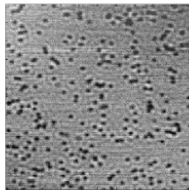
- 準粒子干渉パターンと呼ばれる波模様
- 印加電圧に対して, 波数を求めるとバンド構造が分かる($eV = \frac{\hbar^2 k^2}{2m^*}$)

スパースモデリングによる STS 解析

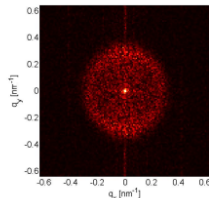
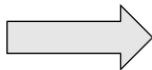
A

- Ag(111)表面の準粒子干渉(電子定在波)パターンを考える
 - 「円形」の散乱ベクトルパターンが見えるはず

Ag(111)表面
Topography像
100nm×100nm
256×256 pts
Bias: 50mV



フーリエ変換



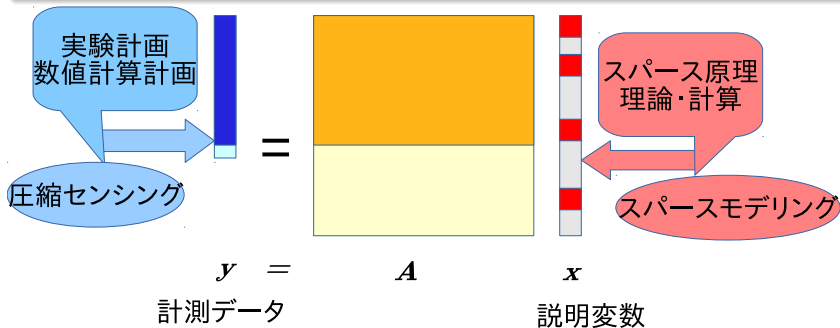
- フーリエ変換先での**スパース性**を活用する.
 - スパース性とは非零成分が少ないことをいう
 - 散乱ベクトルパターンが「真っ黒」
 - cf.) MRI, NMR, 電波望遠鏡

共通の数理：スパースモデリングの考え方

B

スパースモデリング

- 説明変数がスパース (ゼロが多い) である



圧縮センシング (実験計画的発想)

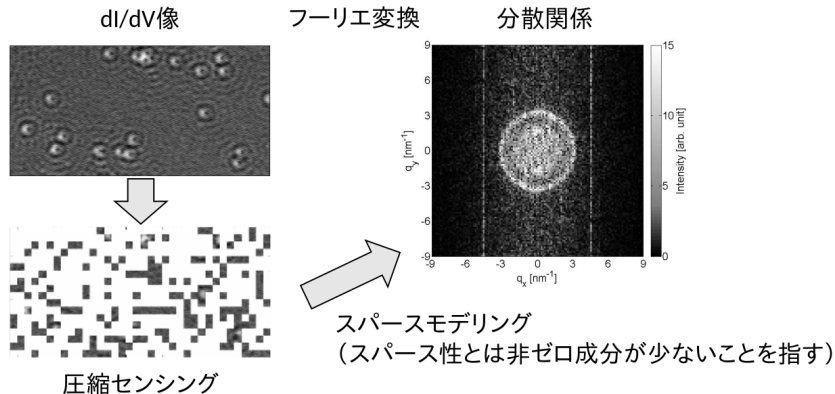
- スパース性を仮定して、データ取得・生成の圧縮方法

圧縮センシング

—Donoho, IEEE Inf. Theory (2006)—

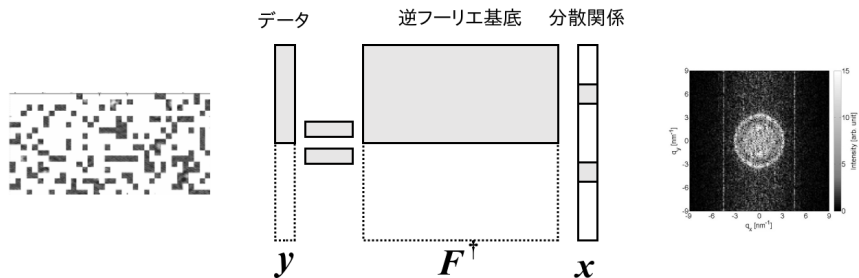
B

- 計測点数を少なくして計測時間を減らすこと
- スパースモデリングを活用する



圧縮センシングの数理

B-C



従来法 (フーリエ変換)

$$\hat{x} = Fy$$

スパースモデリング (LASSO)–Tibshirani, J. Royal Stat. Soc. Ser. B 58 (1996)

スパースモデリングの方法

B-C

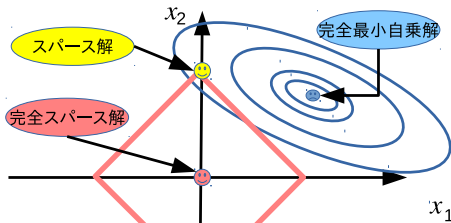
LASSO(Least Absolute Shrinkage and Selection Operator)

あるスパースなベクトル x を見つける方法

$$\operatorname{argmin}_x \left(\frac{1}{2} \|y - F^\dagger x\|_2^2 + \lambda \|x\|_1 \right)$$

- 観測行列 $F : N \times M$ ($N > M$) は条件不足
- $y = F^\dagger x$ を満たす解 x^* は複数あることになる。
どれを選ぶか？

- スパースな解を選ぼうという原理
- $\lambda \|x\|_1$ を正則化とする
- 最適な λ は CV で決める



解法：FISTA(Fast Iterative Shrinkage-Thresholding Algorithm)

C

Beck-Teboulle(2009)

$$F(x) := f(x) + g(x) \quad f(x) = \frac{1}{2} \|y - Ax\|_2^2 \quad g(x) = \lambda \|x\|_1$$

① 二次関数メジャライザーを用いた減少

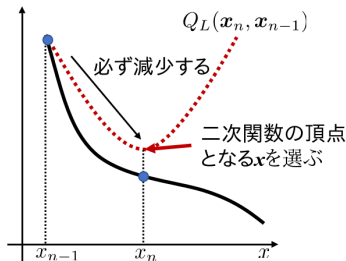
② 軟判定しきい値関数の適用

メジャライザー

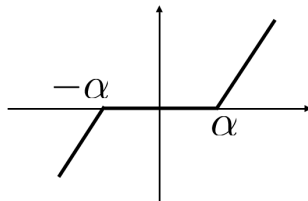
$$q_L(x, y) = f(y) + \langle x - y, \nabla f(y) \rangle + \frac{L}{2} \|x - y\|^2$$

$$x_n = S_{\lambda/L} \left(x_{n-1} - \frac{1}{L} \nabla f(x_{n-1}) \right)$$

メジャライザーの満たすべき性質

$$f(x_n) \leq q_L(x_n, x_{n-1}) \leq f(x_{n-1})$$
軟判定しきい値関数

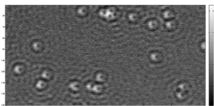
$$S_\alpha(x_i) = \text{sgn}(x_i)(|x_i| - \alpha)_+$$

必ず $F(x)$ が減少する方向に逐次的に更新する

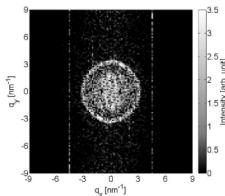
スパースモデリングによる STS 解析

- スパースモデリングにより分散関係を鮮明に抽出.

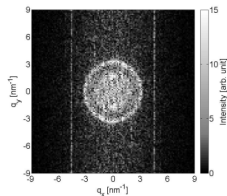
dl/dV 像



スパースモデリング



従来法

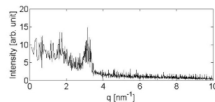
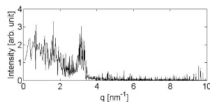


計測試料: Ag(111)

印加電圧: 200 meV

計測面積: $70 \times 35 \text{ nm}^2$

計測点数: $360 \times 180 = 64800 \text{ pts}$

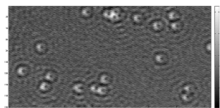


Nakanishi et al, JPSJ **85**, 093702 (2016).

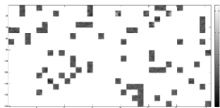
圧縮センシングによる STS 解析

- ・ スパースモデリングにより圧縮センシングが可能.

dI/dV 像

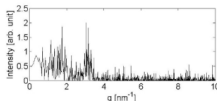
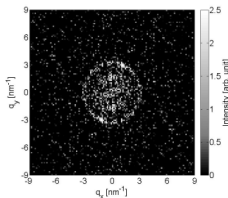


64800 pts

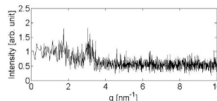
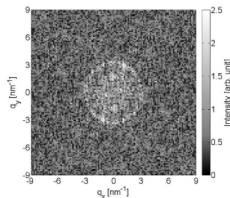


7200 pts (9分の1)

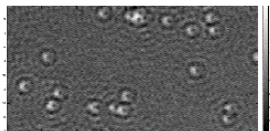
スパースモデリング



従来法

Nakanishi et al, JPSJ **85**, 093702 (2016).

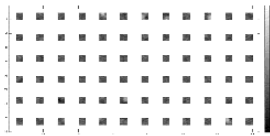
実験計画的考察



random



periodic

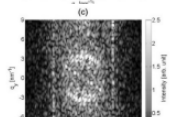
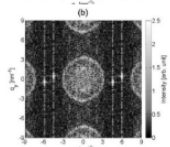
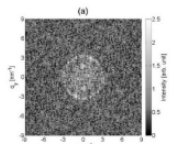
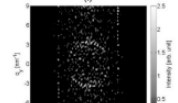
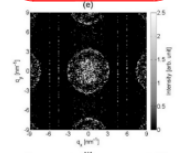
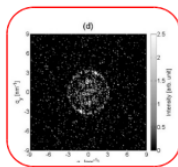


partial



スパースモデリング

従来法



Outline

① 「スパースモデリング」の考え方

② 今後の展開にむけて
PCA を例に

多変量解析としてのシミュレーションデータ

M 個のサンプルのそれぞれに、 N 個の変数の値が観測されているとする。

(サンプル) $\times$ (変数) のデータセットを多変量データと呼ぶ。
多変量解析とは、多変量データの様々な解析法の総称。

例：10 人の生徒の 4 教科の試験の成績

生徒 No.	国語 x_1	英語 x_2	数学 x_3	理科 x_4
1	86	79	67	68
2	71	75	78	84
3	42	43	39	44
4	62	58	98	95
5	96	97	61	63
6	39	33	45	50
7	50	53	64	72
8	78	66	52	47
9	51	44	76	72
10	89	92	93	91

テストのデータもシミュレーションデータも同じようなもの。

主成分分析 (Principal Component Analysis)

PCA の手続き

- ① 相関係数行列 (4×4) の計算
- ② 対角化し、その第一固有値に対応する固有ベクトルから第一主成分を求め、第二固有値の固有ベクトルから第二主成分を求める。

- ③ 固有値 λ はその主成分の寄与率

$$p = \lambda / (\sum_M \lambda_M)$$

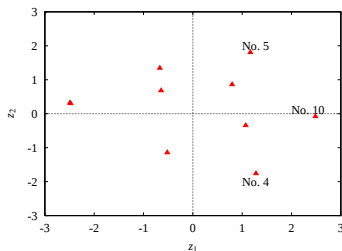
- ④ 主成分とサンプルとの内積から傾向

例題の場合、第一主成分は総合能力、第二成分は理系文系の違いを表している。

$$\lambda_1 = 2.721 \quad z_1 = 0.487u_1 + 0.511u_2 + 0.508u_3 + 0.493u_4$$

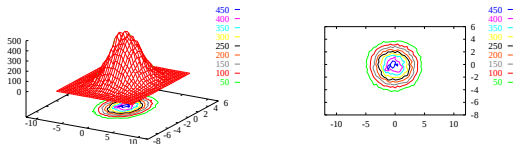
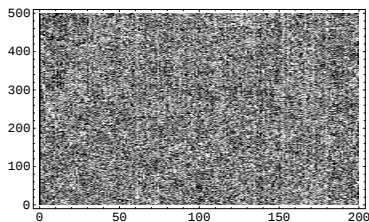
$$\lambda_2 = 1.222 \quad z_2 = 0.527u_1 + 0.474u_2 - 0.481u_3 - 0.516u_4$$

主成分 z_1 と z_2 とサンプル (生徒) との内積の散布図



PCA for simulation data in physics

Monte Carlo simulation data



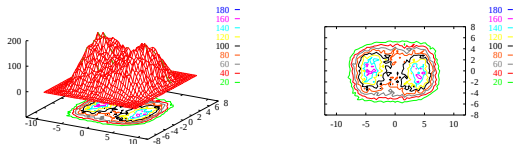
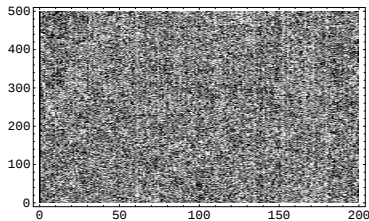
M.Inoue, KH and M.Okada(2006)

Eigen-mode analysis of susceptibility matrix: $N \times N$ matrix

$$\chi_{ij} = \frac{\partial^2}{\partial h_i \partial h_j} F(\{h_i\}) \Big|_{h=0} = \frac{\partial}{\partial h_i} \langle S_j \rangle \Big|_{h=0} = \beta (\langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle)$$

PCA for simulation data in physics

Monte Carlo simulation data



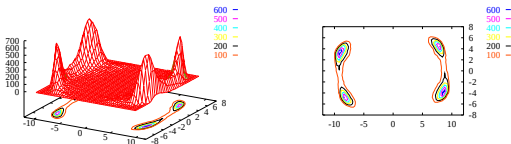
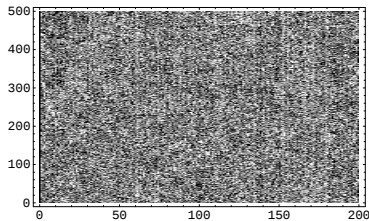
M.Inoue, KH and M.Okada(2006)

Eigen-mode analysis of susceptibility matrix: $N \times N$ matrix

$$\chi_{ij} = \left. \frac{\partial^2}{\partial h_i \partial h_j} F(\{h_i\}) \right|_{h=0} = \left. \frac{\partial}{\partial h_i} \langle S_j \rangle \right|_{h=0} = \beta (\langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle)$$

PCA for simulation data in physics

Monte Carlo simulation data



M.Inoue, KH and M.Okada(2006)

Eigen-mode analysis of susceptibility matrix: $N \times N$ matrix

$$\chi_{ij} = \left. \frac{\partial^2}{\partial h_i \partial h_j} F(\{h_i\}) \right|_{h=0} = \left. \frac{\partial}{\partial h_i} \langle S_j \rangle \right|_{h=0} = \beta (\langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle)$$

PCA 再考

PCA の大規模化

「すべてをスパコンへ」へむけて

- PCA には $M \times M$ サンプル行列と $N \times N$ データ行列には双対関係が存在する .
 $\text{rank} = \min(N, M)$
 - データサイズ N を大きくすると, サンプル数 M も大きくする必要がある?
- そもそも $N \times N$ 行列をディスクに出せない. 保存できない場合は難しい. つまり, バッチ処理的な PCA の限界
- online PCA: 注目するモード数 $p \times N$ 程度で可能 .
 - online 版にすることに意義があったかもしれないが, 大規模計算では必然的 .
 - 本質的には変分法 + 特異値分解

PCA 再考 (cont.)

PCA からの特徴抽出

- PCA はデータ構造を表現するある種のモード分解
 - 定常系に限定されているために、POD の劣化版とも言える？
 - ダイナミクス版は Dynamic Mode Decomposition(DMD)?
 - 主成分の解釈は依然として難しい
 - 非線形 PCA も開発されているが、解釈はより難解
- **よいモード**は分類学に使えるわけで、**記述子**とも呼ばれる
- マテリアルズ・インフォマティクスの現状はよい記述子探索問題となっているようである

PRL 114, 105503 (2015)

PHYSICAL REVIEW LETTERS

week ending
13 MARCH 2015

Big Data of Materials Science: Critical Role of the Descriptor

Luca M. Ghiringhelli,^{1,*} Jan Vybiral,² Sergey V. Levchenko,¹ Claudia Draxl,³ and Matthias Scheffler¹¹Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin-Dahlem, Germany²Department of Mathematical Analysis, Charles University, 18675 Prague, Czech Republic³Humboldt-Universität zu Berlin, Institut für Physik and IRIS Adlershof, 12489 Berlin, Germany

(Received 14 April 2014; revised manuscript received 20 October 2014; published 10 March 2015)

Statistical learning of materials properties or functions so far starts with a largely silent, nonchallenged step: the choice of the set of descriptive parameters (termed *descriptor*). However, when the scientific

まとめ

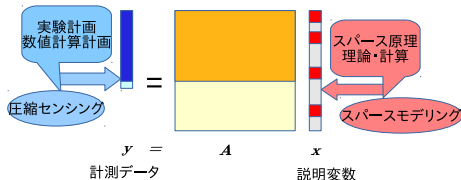
データ駆動物質科学へ

• スパースモデリングを概観

- 走査型トンネル分光の解析
- 将来的には装置に埋め込んで、実時間解析をしたい

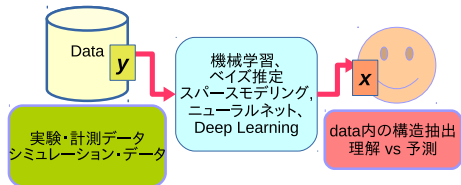
- 512×512 ピクセルをベクトルとする行列演算

- y, A, x に何を?



• PCA を例に大規模計算

- 並列計算を活用したデータ生成としてのスパコン利用
- オンライン化
- 特異値分解



• 第一原理計算と絡み機械学習

- 古典ポテンシャル学習：ある種の基底展開・回帰問題
- 構造最適化問題：ベイズ最適化