

第 64 回物性若手夏の学校・講義 2019/8/6~8/8

# 物性科学研究のための データ駆動科学入門

熊本大学・パルスパワー科学研究所 赤井一郎

## 要旨

データ駆動科学は、ベイズ推定やスパースモデリング等の機械学習の手法を計測・解析に組み込み、従来の知見を超える情報を抽出する取り組みである。これまで科学計測は、現象の源に原因があり、因果律に従って結果が決まると考える一方、誤差論では、計測値は偶然誤差によって確率分布すると考えてきた。データ駆動科学で用いるベイズ推定ではその考え方を逆転させ、計測で得られたデータを「起点」として原因が確率的に分布すると考える。一方スパースモデリングは適切な物理モデルを用いて、データに含まれる要素を顕著な主成分から順に抽出し、適切な規範に沿って現象を理解する正則化の方法である。

データ駆動科学の基本的思想は、計測データを最善の方法(機械学習)で骨の髄まで解析して虚心坦懐にデータに対峙することである。その結果、一切の作為的な先入観を排除するとともに、例え院生であっても経験豊富な研究者と対等に戦うことが可能となる。特にベイズ推定では物性パラメータの統計的分布の評価が可能で、新しく発見した現象や物性の、誰もが納得できる統計的証拠を示すことができる。更に、一切の先入観を排除してデータだけから現象を説明する物理モデルの選択も可能である。講義では、従来の誤差論を復習し、ベイズ推定とスパースモデリングの方法論の紹介と、具体的に物性研究に適用した例を示し、若手研究者の参画を喚起したい。

## 1 復習:誤差論

### 1.1 真値、最確値、誤差

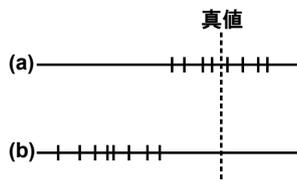
**真値**  
ある物理量を測定したとき、得られた値がどこまで正しいか、その評価が重要である。測定には必ず誤差(不確かさ)が伴うため、測定値は「真値」に一致するわけではない。

**誤差**  
測定値の誤差(error)  $\xi$  は、測定値  $x$  の真値  $X$  からのずれとして定義される。  
誤差 = 測定値 - 真値, ( $\xi = x - X$ ) (1)  
真値は分からないので、そのままでは誤差も分からない。

**最確値とその誤差**  
複数回の測定等を行なった後、後述する方法を用いて真値にもっとも近いであろう値を求めることが可能である。それを最確値(the most probable value)という。しかし最確値も誤差が伴う。よって最確値は最確値の誤差を伴って次の様に表す。  
最確値の表し方: 最確値  $\pm$  最確値の精度 (2)

**偶然誤差と系統誤差**  
誤差には偶然誤差(random error)と系統誤差(systematic error)がある。  
● 偶然誤差  $\xi_r$ : 実験において避けがたいもので、複数回の測定に対し真値のまわりに正負同等にばらついて値がふれる誤差(図1-(a)).  
● 系統誤差  $\xi_s$ : 測定において真値から系統的にずれた値を与える誤差(図1-(b)).  
誤差  $\xi$  はこれらの和になる。

$$\xi = \xi_r + \xi_s \quad (3)$$



第1図: 測定値のばらつきの様子。(a) 偶然誤差のみの場合。(b) 偶然誤差と系統誤差がある場合。

## 1.2 ガウス分布(正規分布)

測定回数を増やすと測定値のヒストグラムは図2の様になり、ある分布関数に漸近する。

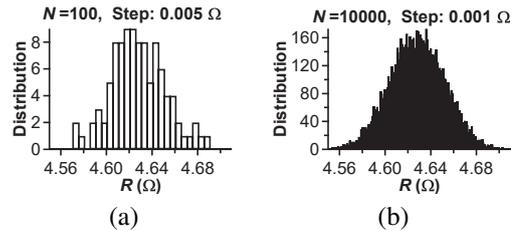


図2: ヒストグラム。(a)100回測定、(b)10000回測定。

その分布関数がガウス分布(正規分布)である。

**ガウス分布**  
真値  $X$  である物理量の測定値の分布は、 $x$  という測定値を得る確率が

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-X)^2}{2\sigma^2}\right] \quad (4)$$

となるガウス分布(Gaussian distribution)に従う。  
 $\sigma$  は分布の広がりを表わし標準偏差と呼ばれる。  
この関数は次の特徴を持つ。  
● 真値  $X$  に対して対称。  
● 真値  $X$  で最大値をとる。  
●  $|x - X| \gg \sigma$  で急激に 0 に近づく。

図3: ガウス分布

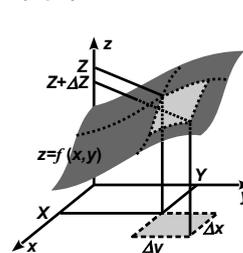
- この関数は分布確率なので、測定値  $x$  の取りえる範囲  $x = -\infty \sim +\infty$  の全積分で 1 となる。
$$1 = \int_{-\infty}^{\infty} f(x) dx \quad (5)$$
- 次の関係が成り立つ。
$$\sigma^2 = \int_{-\infty}^{\infty} (x - X)^2 f(x) dx \quad (6)$$

### 1.3 間接測定と誤差の伝播

直接測定(direct measurement)できる物理量  $x, y$  から、ある物理量  $z$  がある関数  $f(x, y)$  で、 $z = f(x, y)$  と求められるとする。これを物理量  $z$  の物理量  $x, y$  による間接測定(indirect measurement)という。

**平均値を用いた間接測定値の導出**  
直接測定可能な物理量  $a, b, \dots$  と  $z$  に  
$$z = f(a, b, \dots)$$
  
の関係があるとする。  
● 物理量  $a$ : 平均値  $\bar{a}$ , 真値  $A$ , 測定値  $a_1, a_2, \dots, a_n$   
● 物理量  $b$ : 平均値  $\bar{b}$ , 真値  $B$ , 測定値  $b_1, b_2, \dots, b_n$   
物理量  $z$  の最確値  $\bar{z}$  は、直接測定可能な物理量の平均値を用いて、  
最確値:  $\bar{z} = f(\bar{a}, \bar{b}, \dots)$  (7)  
で得ることが出来る。

では、得られた間接測定値の誤差をどう評価すればよいであろうか?



測定可能な物理量  $x, y$  から、関数  $f(x, y)$  で物理量  $z$  が次式で得られるとする。

$$z = f(x, y)$$

$x, y$  の真値を  $X, Y$  とし、それらの値から求まる  $z$  の真値  $Z$  は

$$Z = f(X, Y)$$

とする。

$x, y$  が、それらの真値からそれぞれ  $\Delta x, \Delta y$  だけ微小変化した時、物理量  $z$  の微小変化  $\Delta Z$  は、偏微分を用いて、

$$\Delta Z = \left. \frac{\partial f(x, y)}{\partial x} \right|_Z \Delta x + \left. \frac{\partial f(x, y)}{\partial y} \right|_Z \Delta y \quad (8)$$

である。ここで、 $\left. \frac{\partial f(x,y)}{\partial x} \right|_Z, \left. \frac{\partial f(x,y)}{\partial y} \right|_Z$  は、 $f(x,y)$  の  $x = X, y = Y$  における偏微分の微係数である。

平均値を用いた間接測定値の誤差 (誤差の伝播)

直接測定可能な物理量  $a, b, \dots$  と、得たい物理量  $z$  に、 $z = f(a, b, \dots)$  の関係があり、物理量  $a, b, \dots$  が互いに独立であるとする。

- 物理量  $a$  :  
平均値  $\bar{a}$ , 真値  $A$ , 平均値の分布の標準偏差  $\sigma_{am}$
  - 物理量  $b$  :  
平均値  $\bar{b}$ , 真値  $B$ , 平均値の分布の標準偏差  $\sigma_{bm}$
- 物理量  $z$  の最確値  $\bar{z}$  [(7)式] の分布の標準偏差  $\sigma_{zm}$  は (9)式で、確率誤差  $\varepsilon_{zm}$  は (10)式で求められる。これを誤差の伝播 (Propagation of error) という。

$$\sigma_{zm} = \sqrt{\left(\left. \frac{\partial f}{\partial a} \right|_Z\right)^2 \sigma_{am}^2 + \left(\left. \frac{\partial f}{\partial b} \right|_Z\right)^2 \sigma_{bm}^2 + \dots} \quad (9)$$

$$\varepsilon_{zm} = 0.6745 \sigma_{zm}$$

$$= \sqrt{\left(\left. \frac{\partial f}{\partial a} \right|_Z\right)^2 \varepsilon_{am}^2 + \left(\left. \frac{\partial f}{\partial b} \right|_Z\right)^2 \varepsilon_{bm}^2 + \dots} \quad (10)$$

### 1.4 最小二乗法

最小二乗法

測定値の確率分布  $g(X)$  は  $X$  が本当の真値であるときに最大となり、 $f(\xi)$  は正しいガウス分布となる。 $g(X)$  が最大になるには、誤差の2乗の和 (指数関数の中)

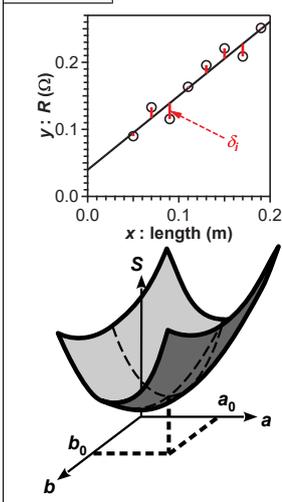
$$S = \sum_{i=1}^n \xi_i^2 = \sum_{i=1}^n (x_i - X)^2 \quad (11)$$

が最小となる必要がある。この様に誤差の2乗の和を最小にして、最確値を推定する方法が最小二乗法である。

最小二乗法は直線回帰でよく使われる。 $n$  個の測定値の対  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  が得られたとし、これらの値において以下が満たされるとする。

- $y_i$  の各値の確率誤差が等しく、 $x_i$  には誤差が無い。これらの測定値の対に次の直線関係が期待されるとする。  
 $y = ax + b$

直線回帰



$\sigma_x = 0$  なので、 $y_i$  の測定値と理論値 ( $y = ax_i + b$ ) の残差  $\delta_i$  を取り扱う。

$$\delta_i = y_i - (ax_i + b) \quad (12)$$

この残差の2乗  $\delta_i^2$  の和  $S$

$$S(a, b) = \sum_{i=1}^n \delta_i^2 \quad (13)$$

が最小値となる条件から得られる  $a, b$  が最確値である。最小となるには、(14)式の連立条件が満たされればよい。

$$\frac{\partial S}{\partial a} = 0, \quad \frac{\partial S}{\partial b} = 0 \quad (14)$$

この方程式を正規方程式 (normal equation) という。

任意の関数  $f(x; a, b, \dots)$  への回帰は同様に以下のように行うが、推定したいパラメータが関数に非線形で含まれる場合 (一般的にそうだが)、図4に示した様に (16)式の解がユニークに決まらない。

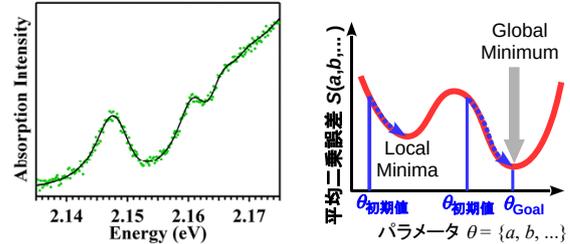
関数回帰

$\sigma_x = 0, y$  値の誤差は同じ ( $\sigma_y = \sigma_{y_i} \neq 0$ ) とした場合、任意の関数  $f(x; a, b, \dots)$  への回帰は次式の  $S$

$$S = \sum_{i=1}^n [y_i - f(x_i; a, b, \dots)]^2 \quad (15)$$

に対して (16)式の正規方程式を解いて行われる。

$$\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = \dots = 0 \quad (16)$$



(a) (b)

図4: 最小二乗法の困難さ

## 2 ベイズ推定

データ駆動科学は、これまでのデータ解析の価値観を一変させるもので、以下のことが可能となる。

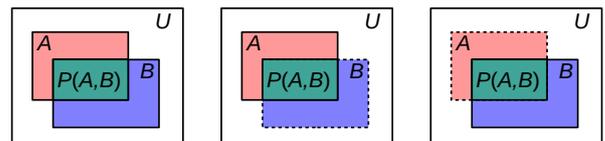
- データを骨の髄まで解析する
    - 間接測定される物理量の統計的精度を評価
    - 物理モデルの妥当性の評価
    - データ解析における後ろめたさを一掃する
    - 既存の解析法を凌駕する新解析法
  - 実験の最適化
    - 極限測定で、必要な精度を得るための実験条件の推定
    - 実験を行う前の事前シミュレーションによる実験の効率化
  - 理論研究へのフィードバック
    - 解析に用いる物理モデルの良し悪しを判定
- 本節ではデータ駆動科学で用いられるベイズ推定の概要をまとめる。

### 2.1 ベイズの定理

まず、ベイズの定理 [1] とベイズ推定の表現に用いられる、同時確率 (joint probability) と条件付き確率 (conditional probability) について述べる。

同時確率とは、 $A$  という事象と  $B$  という事象が同時に起こる確率 [図5-(a)] で、 $P(A, B) \equiv P(A \cap B)$  と表す。

一方、 $A$  という事象と  $B$  という事象が同時に起こる確率は、まず  $A$  という事象が起こる確率  $P(A)$  と、 $A$  という事象が起こった条件下で  $B$  という事象が起こる条件付き確率の積  $P(A, B) = P(B|A)P(A)$  で書くことができる。ここで  $P(B|A)$  が、 $A$  という事象が起こった条件下で  $B$  という事象が起こる条件付き確率で、次の様に書ける。



(a) (b) (c)

図5: 同時確率 (a) と条件付き確率 (b), (c)

- $A$  条件下の  $B$  の確率 [図5-(b)],  $P(B|A) = \frac{P(A, B)}{P(A)}$
- $B$  条件下の  $A$  の確率 [図5-(c)],  $P(A|B) = \frac{P(A, B)}{P(B)}$

**ベイズの定理**

200年以上前に T. Bayes は、[1]で、同時確率について、以下のベイズの定理を提唱した。

$$P(A, B) = P(B|A)P(A) = P(A|B)P(B) \quad (17)$$

$$P(A|B) = \frac{P(B|A)}{P(B)}P(A), \quad P(B|A) = \frac{P(A|B)}{P(A)}P(B) \quad (18)$$

図6: ベイズの定理と同時確率

(18)式で表したベイズの定理について再認識しておく。ベイズの定理の言うことは、(19)式に示すように、事前条件がない状況では A 事象が P(A) の確率で起こるのに対し、その確率 P(A) が、B が起こった条件下では A の起こる確率が、尤度 P(B|A)/P(B) によって、P(A) から P(A|B) へと更新されることを意味する。

**事前確率、事後確率と尤度**

$$\underbrace{P(A|B)}_{\text{事後確率}} = \frac{\underbrace{P(B|A)}_{\text{尤度}}}{\underbrace{P(B)}_{\text{事前確率}}} \underbrace{P(A)}_{\text{事前確率}} \quad (19)$$

- 事前確率 (prior probability)  
P(A) は、何の条件もないときに A の事象が起こる確率で、事象 A の事前確率 (prior probability) と言う。
- 事後確率 (posterior probability)  
一方 P(A|B) は、事象 B が起こった後の確率を意味し、事後確率 (posterior probability) と呼ばれる。
- 尤度 (ゆうど: likelihood)  
この事前確率と事後確率を結ぶ係数にあたる P(B|A)/P(B) を尤度 (ゆうど: likelihood) と呼ぶ。

## 2.2 因果律と物理モデル

因果律 (causality) は、結果には何らかの原因があり、その間に一定の法則性があることを言う。実験の測定行為における因果律についてまとめておく。

**因果律と P(θ, D)**

ある物理現象を測定することを考えよう。物理現象にはそれが従う基本方程式があり、基本方程式に従って測定量を与える物理モデルが存在する。

図7: 測定行為の因果律

物理モデルには、現象や測定量の特徴づける複数のパラメータ (パラメータセット θ) が内在する。

因果律では、基本方程式、物理モデルとパラメータセット θ が現象の原因であり、測定することによって得られるデータセット D が結果となる。

物理モデルのパラメータセット θ に基づき、あるデータセット D が得られる同時確率を考えよう。因果律に従えば、原因であるパラメータセット θ が確率分布 P(θ) で決まってい、測定することにより、P(D|θ) が得られ、次式のように同時確率が得られることになる。

因果律:  $P(\theta, D) = \underbrace{P(D|\theta)}_{\text{測定行為}} P(\theta) \quad (20)$

## 2.3 データ駆動科学におけるベイズ推定

[2]を参考文献として、話をベイズ推定に戻そう。データセット D は (21)式の様に、

$$y_i = \underbrace{f(x_i; \theta)}_{\text{物理モデル}} + \underbrace{n_i}_{\text{ノイズ}} \quad (21)$$

物理モデル  $f(x_i; \theta)$  に偶然誤差  $n_i$  が重畳して与えられる。以降このノイズが標準偏差  $\sigma_{\text{data}}$  で特徴づけられる正規分

布 (ガウス分布) に従う場合について考える。

この場合、i 番目の測定点で正規分布ノイズ  $n_i$  が発生する確率  $P(n_i)$  は、正規分布により、

$$P(n_i) = \frac{1}{\sigma_{\text{data}}\sqrt{2\pi}} \exp\left(-\frac{n_i^2}{2\sigma_{\text{data}}^2}\right) \quad (22)$$

である。ベイズ推定ではこの仮定に基づき因果律を遡る。

データセットの横軸  $x_i$  は確定して与えられているとしよう。そうすると、(21)式より、ノイズ  $n_i$  は、

$$n_i = y_i - f(x_i; \theta)$$

であるから、(22)式のノイズ  $n_i$  を与える確率  $P(n_i)$  は、(20)式の因果律に従って、パラメータセット θ が決まった条件下でデータ  $D_i$  が与えられる条件付き確率

$$P(n_i) = P(D_i|\theta)$$

と等価であることが分かる。よって (21)式を用いて  $n_i$  を残差  $y_i - f(x_i; \theta)$  で書き換えると、(23)式と書ける。

$$P(n_i) = P(D_i|\theta)$$

$$= \frac{1}{\sigma_{\text{data}}\sqrt{2\pi}} \exp\left(-\frac{[y_i - f(x_i; \theta)]^2}{2\sigma_{\text{data}}^2}\right) \quad (23)$$

データに重畳するノイズ  $n_i$  は i に関して独立であるから、パラメータセット θ が与えられた条件下でデータセット D が得られる条件付き確率  $P(D|\theta)$  は、(24)式の平均二乗誤差  $E(\theta)$  を用いて、(25)式となる。

$$E(\theta) = \frac{1}{2N} \sum_{i=1}^N [y_i - f(x_i; \theta)]^2 \quad (24)$$

$$P(D|\theta) = \left(\frac{1}{\sigma_{\text{data}}\sqrt{2\pi}}\right)^N \exp\left(-\frac{N}{\sigma_{\text{data}}^2} E(\theta)\right) \quad (25)$$

$$P(\theta, D) = P(\theta|D)P(D)$$

測定行為は因果律に従い、(20)式が示した様に、パラメータセット θ が確率分布 P(θ) で決まった上で P(D|θ) を得ることである。その結果、P(θ, D) の同時確率が得られることに対応する。

図8: ベイズの定理の適用

一方ベイズの定理によれば、P(θ, D) は (26)式と書ける。

因果律を遡る:  $P(\theta, D) = P(\theta|D)P(D) \quad (26)$

**ベイズの定理の適用**

我々人類が得たいものはデータセット D が得られた条件下でのパラメータセット θ の確率分布であって、その P(θ|D) はベイズの定理を用いて、(27)式と書ける。

$$\underbrace{P(\theta|D)}_{\text{事後確率}} = \frac{\underbrace{P(D|\theta)}_{\text{尤度}}}{\underbrace{P(D)}_{\text{事前確率}}} \underbrace{P(\theta)}_{\text{事前確率}} \quad (27)$$

尤度の P(D) はデータが取得された時点で値が確定するため、P(θ|D) は (25)式より、(28)式となる。

$$\underbrace{P(\theta|D)}_{\text{事後確率}} \propto \exp\left(-\frac{N}{\sigma_{\text{data}}^2} E(\theta)\right) \underbrace{P(\theta)}_{\text{事前確率}} \quad (28)$$

## 2.4 データ駆動科学の最初の一步

表1にまとめた様に、誤差論とベイズ推定は根本的なデータに対する認識が異なる。

表1: これまでの誤差論とベイズ推定の考え方の違い

$f(x_i; \theta)$	$\theta$	$y_i$
誤差論	確定的	確率的
ベイズ推定	確率的	確定的

- これまでの誤差論: 「パラメータ θ は真値で確定されていて、データがノイズで確率的に与えられる」と考える。

- **ベイズ推定**: 「データが確定されていて、 $\theta$  が  $P(\theta|D)$  によって確率的に決まる」と考える。この様な「データを出発点と考える」考え方は、データ駆動科学の最初の一步である。

### 2.5 マルコフ連鎖モンテカルロ法

(28)式の  $P(\theta|D)$  をどの様にして得るか? ベイズの定理は200年以上前に提唱されていたにもかかわらず、その実質的な計算には近年のコンピュータの進化を必要とした。

一番単純に  $P(\theta|D)$  を評価するアルゴリズムは、パラメータ空間を必要な精度でメッシュに切って総当たりするグリッドサーチである。まずその例として、表2に示した様にパラメータは5つで、指定した範囲内を必要な有効桁の精度で、 $P(\theta|D)$  を評価することを考えよう。この場合  $P(\theta|D)$

物理量	真値	範囲	有効桁	単位
$f$	3.0	2.5 ~ 3.5	3桁	THz
$\phi$	0.0	$-\pi/2 \sim +\pi/2$	3桁	rad.
$\tau_r$	0.4	0.1 ~ 1.0	2桁	ps
$\tau_d$	6.0	1.0 ~ 10.0	2桁	ps
$a$	1.0	0.1 ~ 2.0	2桁	

表2: 解析により得たい物理量とその有効桁

の計算は  $4.835 \times 10^{13}$  回必要で、例えば  $P(\theta|D)$  の1回の計算に16.4 ms(実測例)かかったとすると、全ての計算を終えるのに、25,000年の時間を費やす。

この計算時間爆発を解消するアルゴリズムとして様々なマルコフ連鎖モンテカルロ法<sup>[1][2]</sup> (MCMC: Markov chain Monte Carlo methods)[3]が開発されている。ここではその基本的な考え方をまとめておく。

1. まず、各パラメータに「適当」な初期値  $\theta_{init}$  を与える。真値(不明)らしき値でも構わないし、乱数で適当にパラケさせても良い。
2. 次に、現在のパラメータの値  $\theta_t$  を元にして、何らかの乱数  $\epsilon$  によって、パラメータの次の候補  $\theta'$  を生成する。

$$\theta' = \theta_t + \epsilon$$

3. 現在の値  $\theta_t$  と、候補の値  $\theta'$  の事後確率の比  $r$  を評価する。

$$r \equiv \frac{P(\theta'|D)}{P(\theta_t|D)} \quad (29)$$

ここで比を評価する意味は、(27)式に現れる  $P(D)$  を評価する必要がなくなり、事前確率  $P(\theta)$  が一様で、重畳するノイズが正規分布ノイズならば

$$r = \exp \left[ -\frac{N}{\sigma_{data}^2} \{E(\theta') - E(\theta_t)\} \right] \quad (30)$$

となることである。

4. 比  $r$  の大小で、候補の値  $\theta'$  の採択・非採択を決定し、次のパラメータ値(2.から繰り返す)とする。

マルコフ連鎖モンテカルロ法では、2.~4. の繰り返し過程において、マルコフ連鎖的に

$$\theta_{init} \rightarrow \dots \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \dots$$

$P(\theta|D)$  を与える度数分布を、繰り返しながら蓄積する。

よってマルコフ連鎖モンテカルロ法では、 $P(\theta|D)$  を与える度数分布が得られた時点で繰り返すを終了し、得られた度数分布の度数  $N(\theta)$  から、 $P(\theta|D)$  を得る。

[1] マルコフ連鎖は、A. A. Markov の「マルコフ過程」名に由来したものである。マルコフ過程は、物理現象を理解する確率過程で、未来の挙動が現在の挙動のみによって決定され、それ以前には依存しない確率過程のことである。

[2] モンテカルロ法は、カジノで有名な国家モナコ公国にあるモンテカルロの地名に由来する。カジノでサイコロを振るように、乱数を用いる方法であるためである。

### 2.6 メトロポリス法

マルコフ連鎖モンテカルロ法の代表的なアルゴリズムにメトロポリス法(Metropolis algorithm)[4]がある。実質はコンピュータの高速化により近年になって具現化[5]された。

メトロポリス法は、以下の過程で  $P(\theta|D)$  の確率分布を得るアルゴリズムである。

1. 乱数  $\epsilon$  を用いて、確率的に遷移候補を生成する。

$$\theta' = \theta_t + \epsilon, \quad (31)$$

2. (28)式から求められる採択確率  $\alpha$

$$\alpha = \min\{1, r\}, \quad r \equiv \frac{P(\theta'|D)}{P(\theta_t|D)} \quad (32)$$

で、 $\theta_{t+1} \leftarrow \theta'$  の状態更新を行う。採択されなかった場合は、 $\theta_{t+1} \leftarrow \theta_t$  更新を行う。

特に候補値生成用の乱数  $\epsilon$  に、左右対称な一様乱数、正規乱数なども用いるものをメトロポリス法と呼ぶ。

その制限を拡張したものは、メトロポリス・ヘイスティングス法(M-H法: Metropolis-Hastings algorithm)と呼ばれる。メトロポリス法の概要を以下に示す。

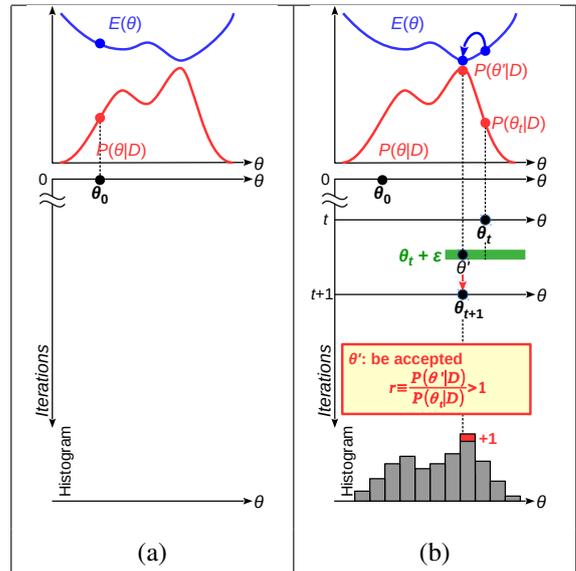


図9: メトロポリス法の状態更新の例-1

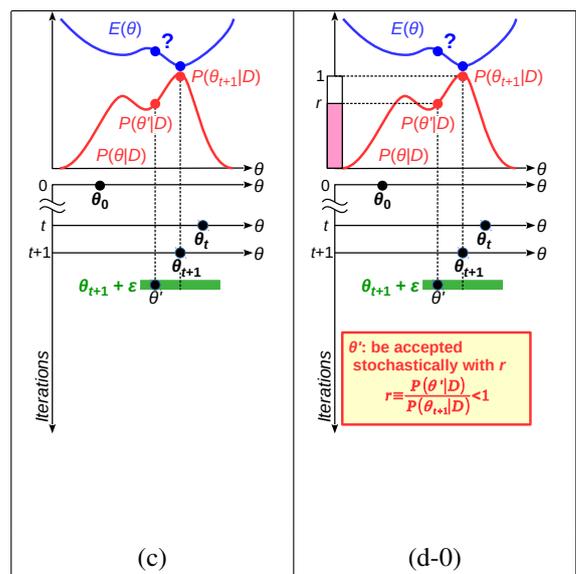


図10: メトロポリス法の状態更新の例-2

- 図9-(a):  $\theta$  の初期値  $\theta_0$  を適当に設定する。
- 図9-(b):
  1. 現在の状態  $\theta_t$  から、乱数  $\epsilon$  で候補  $\theta'$  ( $= \theta_t + \epsilon$ ) を生成する。 $\epsilon$  は、平均値 = 0 の一様乱数や正規乱数等。

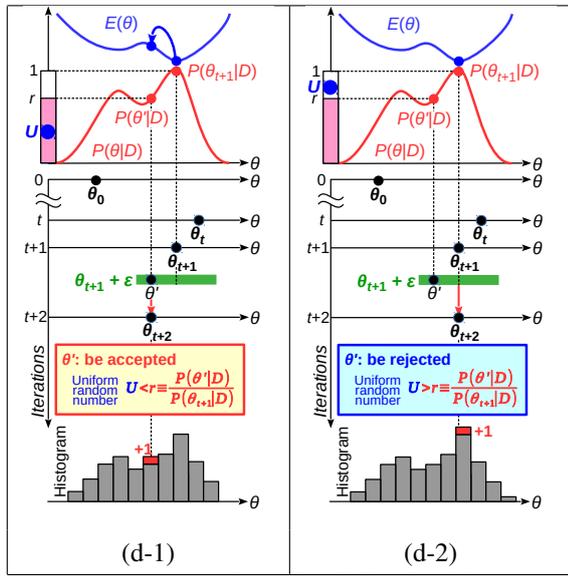


図11: メトロポリス法の状態更新の例-3

2. 事後確率を比較する。

$$r = \frac{P(\theta'|D)}{P(\theta_t|D)} > 1$$

図9-(b)では、候補  $\theta'$  の方が確率が高い ( $r > 1$ ) ので、無条件で候補を採用 ( $\theta_{t+1} \leftarrow \theta'$ ) して記録。

- 図10-(c): 現在の状態  $\theta_{t+1}$  から、乱数  $\varepsilon$  で候補  $\theta'$  ( $\theta' = \theta_{t+1} + \varepsilon$ ) を生成する。
- 図10-(d-0): 事後確率を比較する。

$$r = \frac{P(\theta'|D)}{P(\theta_{t+1}|D)} < 1$$

図10-(d-0)では、 $r < 1$  で、候補  $\theta'$  の方が確率が低い。

- 図11-(d-1):  $0 \sim 1$  の一様乱数  $U(0,1)$  を発生させ、 $r$  と比較する。
- 図11-(d-1)では、 $U < r$  であるので、候補  $\theta'$  を採用 ( $\theta_{t+2} \leftarrow \theta'$ )。
- 図11-(d-2):  $0 \sim 1$  で発生させた一様乱数  $U(0,1)$  が、 $U > r$  であるので、候補  $\theta'$  を棄却し、 $\theta_{t+1}$  を採用 ( $\theta_{t+2} \leftarrow \theta_{t+1}$ ) する。
- 以上を繰り返し、 $\theta_t$  の履歴のヒストグラムとして  $P(\theta|D)$  を得る。

2.7 メトロポリス法を用いたベイズ推定の例

[6] に示したベイズ推定結果の概要をまとめる。

図12は、ビスマスのコヒーレントフォノン信号の例である。

これは、フェムト秒のレーザーパルスを試料に照射させると、そのレーザー光によって、試料内の格子振動や分子振動が一斉に励振され、その振動に応じて物質の反射率や透過率が周期的に変化する現象を時間差のつけた弱いプローブ光パルスで計測して得られる信号である。

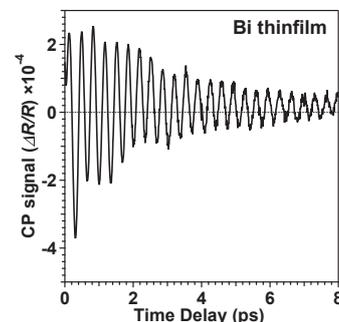


図12: ビスマス薄膜のコヒーレントフォノン信号

このコヒーレントフォノン計測は、試料内の格子振動や分子振動のダイナミクスを時系列変化として得られる方法である。このデータに

$$Y(t) = \frac{a}{\tau_d - \tau_r} \left[ \exp\left(-\frac{t}{\tau_d}\right) - \exp\left(-\frac{t}{\tau_r}\right) \right] \sin(2\pi ft + \phi)$$

の物理モデルを適用し、メトロポリス法でベイズ推定した。パラメータは表3に示した5個である。データに重畳す

物理量	記号	$\bar{\theta} \pm \epsilon_{\theta}$
振動数	$f$	$2930.04 \pm 0.93$ GHz
位相	$\phi$	$5.219 \pm 0.012$ rad.
立上り時定数	$\tau_r$	$0.0809 \pm 0.0048$ ps
減衰時定数	$\tau_d$	$2.811 \pm 0.051$ ps
積分強度	$a$	$(8.63 \pm 0.10) \times 10^{-16}$

表3: パラメータセットとその平均値/Table I in [6].

る正規分布ノイズの標準偏差  $\sigma_{\text{data}}$  は、信号強度の次元で、 $\sigma_{\text{data}} = 3.1 \times 10^{-5}$  と設定した。

メトロポリス法の繰り返し回数は100万回である。その内、初期の10万回は、適当に設定した初期値から適切なパラメータ値付近に収束する焼きなまし区間 (**Burn-In phase**) として破棄し、それ以降の度数分布を事後確率分布  $P(\theta|D)$  として取得した。表3の  $\pm \epsilon_{\theta}$  は  $P(\theta|D)$  の分布幅を公差誤差 (確率誤差) で示したものである。

このベイズ推定で注目すべき点は、周波数  $f$  の推定精度である。従来、減衰振動波形の振動モードの抽出にはフーリエ変換等が用いられるが、振動振幅が減衰特性を持つため、フーリエ変換スペクトルで不確実性幅は免れない。それに対しこのベイズ推定の結果では、振動振幅の変化を適切にモデル化することによって、フーリエ幅と比較して2桁程度高精度な周波数の推定に成功 [6] している。

2.8 候補値を生成する乱数の大きさ  $\sigma_{\theta}$

MCMC法では、パラメータ候補値の生成で用いる乱数の大きさ  $\sigma_{\theta}$  を適切に選択する必要がある。

原理的には得られるパラメータの  $P(\theta|D)$  の広がり (幅) は、実験データに重畳するノイズ  $n_i$  の標準偏差  $\sigma_{\text{data}}$  によって決定づけられるが、不適切な  $\sigma_{\theta}$ 、例えば小さすぎる  $\sigma_{\theta}$  を用いると値の遷移幅が小さく、広いパラメータ空間内で十分な履歴を得る効率が低下する。一方大きすぎる  $\sigma_{\theta}$  を用いると、パラメータ空間の全域探索には有効であるが、確率分布比  $r$  による採択効率が落ち、やはり、十分な履歴を得る効率が低下する。よって、メトロポリス法では、確率分布比  $r$  をモニターして、適切な  $\sigma_{\theta}$  を選択するチューニングが必要である。

2.9 最小2乗法とメトロポリス法の比較

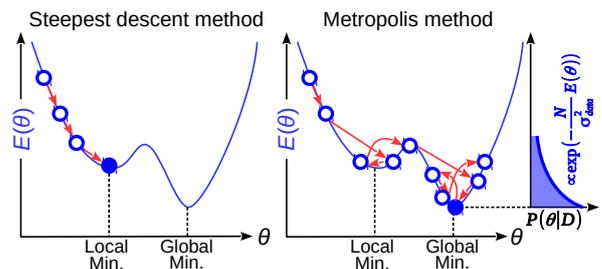


図13: 最急降下法等による最小2乗法とメトロポリス法の比較

通常の最小二乗法は、(33)式に示した、測定データ  $y_i$  とフィッティング関数  $g(x_i; \theta)$  の平均二乗誤差  $E(\theta)$

$$E(\theta) = \frac{1}{n} \sum_{i=1}^n [y_i - g(x_i; \theta)]^2 \quad (33)$$

を最小化して、パラメータセット  $\theta$  を決定する。また、最小化させる方法として最急降下法が用いられる場合が多い。しかし、 $E(\theta)$  がパラメータセットの  $\theta$  空間で図13に示したように、多谷構造を取る場合、パラメータセットの初期値によっては、多谷構造の内、大域的極小値 (global minimum) に到達できず、局所的極小値 (local minimum) にトラップされることがある。

それに対しメトロポリス法では、以下のメリットがある。

- 局所的極小値から解放される。  
パラメータの候補を乱数生成するため、局所的極小値に局在せず、原理的には大域的極小値の探索が可能である。
- パラメータセット全空間の事後確率が得られる。  
詳細つりあいが保証されるため、無限に時間をかければ、パラメータセットの全空間にわたってパラメータセットの事後確率  $P(\theta|D)$  を得ることができる。
- パラメータセットの確率分布を得ることができる

確率的にパラメータ候補を生成し、その度数分布から事後確率  $P(\theta|D)$  を得るため、大域的極小値近傍で、図13に示したように、パラメータセットの分布を生成することが可能で、事前確率  $P(\theta)$  が一様ならば、 $P(\theta|D)$  は

$$P(\theta|D) \propto \exp\left(-\frac{N}{\sigma_{\text{data}}^2} E(\theta)\right)$$

となる。ここで  $\sigma_{\text{data}}$  は、データに重畳するノイズ (正規分布ノイズ) の標準偏差で、

$$\sigma_{\text{data}}^2 \sim k_B T_{\text{noise}} \quad (34)$$

と考えれば、ボルツマン分布になっていることが理解できる。

### 2.10 レプリカ交換モンテカルロ法概論

メトロポリス法では原理的に一意の確率分布を得ることが可能であるが、 $E(\theta)$  の局所的極小値を脱して大域的極小値に達する (図14) には時間を必要とする。

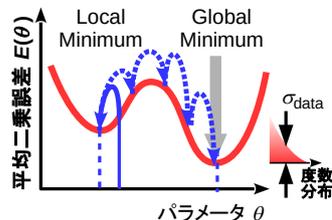


図14: メトロポリス法の問題点

その問題点を回避するアルゴリズムとして、レプリカ交換モンテカルロ法 [7, 8, 2] (RXMC: Replica eXchange Monte Carlo method) が開発された。

この拡張は、(34)式に示した様に、 $E(\theta)$  をエネルギーと捉え、 $\sigma_{\text{data}}^2$  と  $T_{\text{noise}}$  が対応しボルツマン分布と捉えること、イジングスピンにおける熱アニール効果をうまく組み合わせた着想 [8] を起点とする。

**レプリカ交換モンテカルロ法**

この考え方を発展させ、人為的に逆温度  $\beta$  を導入する。

$$P(\theta|D, \beta) \propto \exp\left(-\frac{N}{\sigma_{\text{data}}^2} \beta E(\theta)\right) P(\theta) \quad (35)$$

- $\beta = 1$ : (28)式と等価。
- $0 < \beta < 1$ : 有効温度を上げたことに対応する。

であるが、レプリカ交換モンテカルロ法では、 $0 < \beta \leq 1$  で複数の  $\beta$  のレプリカ (図15) を用意して大域的極小値を高速に探索する。その探索は、

- 各レプリカで、メトロポリス法で  $P(\theta|D, \beta)$  を取得。
- 隣り合うレプリカ間で状態交換を確率的に行う。を繰り返すことで実行する。

図15: 異なる  $\beta$  のレプリカ

ここで  $\beta < 1$  では有効温度を上げたことに意味する。そのような高温のレプリカ内でメトロポリス法を用いて探索する際、適切な採択率でサンプリングの効率化を行うため、候補値生成で用いられる乱数幅 (ステップ幅) を広げて探索する。よって高温では、図15に示した様に、 $E(\theta)$  の局所構造を無視して、パラメータ空間を大域的に探索が行われる。

この様に  $\beta < 1$  の高温で大域的に探索がなされるため、隣り合うレプリカ間で状態交換を確率的に行うことで、例えば低温  $\beta = 1$  で局所的極小値にトラップされても、一旦高温に状態交換がなされ、局所的極小値からの脱出が効率化される。

### 2.11 ベイズ自由エネルギーに基づくモデル選択

モデル選択とは、実験データを説明する物理モデルが複数存在する際に、どの物理モデルが統計的に妥当であるかを判断することである。以下の例がそれに当たる。



図16: 直線回帰?

- 直線回帰  
単純な例として、図16のようなデータに対し、

$$y = b, y = ax, y = ax + b, \dots$$

- どの物理モデルを選択すべきかの判断は難しい。
- スペクトル分解 [2]  
計測された様々なスペクトルを成分分解する場合を考えよう。その計測データにピーク形状のスペクトル要素がいくつ含まれているか? 多分このような問いかけをしたら、要素数の数は既知であると答えるかもしれない。しかし既知と思った要素数は正しいであろうか? 新しいピーク構造の発見を見落としていないか?
- ハミルトニアン選択 [9]  
計測されたデータの理解は様々な論争を引き起こす。根底にあるものはハミルトニアンで、データを説明するためには、ハミルトニアンの中のどの項が重要なのか? 複数のモデルハミルトニアンでどれが、データをよく説明するか?

ベイズ自由エネルギー最小化を情報量規準としたモデル選択 [2, 9] は、これらの問題や問いかけに対し、データを起点としたデータ駆動科学的な解決策を提供する。

このモデル選択を念頭に、因果律を再考しよう。これまででは物理モデルが決まっているとして、それに内包するパラメータセット  $\theta$  が因果律の原因と考えてきた。

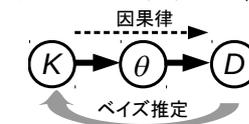


図17: モデル選択

ベイズ推定のモデル選択は、データセットが与えられた条件下で物理モデルである  $K$  の確率分布  $P(K|D)$  を評価することである。

図17の因果律に沿って  $K, \theta, D$  の同時確率  $P(K, \theta, D)$  を表すと、(36)式となる。

$$P(K, \theta, D) = P(D|\theta, K)P(K) \quad (36)$$

ここで、 $P(D|\theta, K)$  は  $K$  と  $\theta$  が決まった条件下でデータセット  $D$  が与えられる条件付き確率であるから (28)式と等価で、(37)式である。

$$P(D|\theta, K) = \prod_{i=1}^N P(D_i|\theta, K) \propto \exp\left(-\frac{N}{\sigma_{\text{data}}^2} E(\theta, K)\right) \quad (37)$$

モデル選択では  $P(K|D)$  を評価したいわけだから、 $P(K, D)$  の同時確率について考えよう。この  $P(K, D)$  は、 $P(K, \theta, D)$  で  $\theta$  を消去すれば良いので、 $P(K, \theta, D)$  を全パラメータ空間  $\theta$  に渡って積分すれば得られる。この様に特定の変数を消去することを周辺化 (marginalization) という。

$$P(K, D) = \int P(K, \theta, D) d\theta \quad (38)$$

その結果ベイズの定理を適用し、 $P(K|D)$  は (39)式となる。

$$P(K|D) = \frac{P(D|K)}{P(D)} P(K) \propto P(K) \int \exp\left(-\frac{N}{\sigma_{\text{data}}^2} E(\theta, K)\right) P(\theta) d\theta \quad (39)$$

熱統計力学において分配関数  $Z$  は、逆温度  $\beta$  と状態を規定するパラメータ  $\theta$  を用いて、 $Z = \sum_{\theta} \exp(-\beta E(\theta))$  と定義される。これと (39)式との対応を見ると、平均二乗誤差

$E(\theta, K)$  をエネルギーと捉え、 $\beta \leftarrow N/\sigma_{\text{data}}^2$  と捉えると、情報科学の分配関数  $Z$

$$Z \equiv \int \exp\left(-\frac{N}{\sigma_{\text{data}}^2} E(\theta, K)\right) P(\theta) d\theta \quad (40)$$

を定義することは極めて有用である。

またこの分配関数  $Z$  を用いてヘルムホルツの自由エネルギー  $F$  に対応するベイズ自由エネルギー [2] が、次式で定義される。

$$F \propto -\ln \left[ \int \exp\left(-\frac{N}{\sigma_{\text{data}}^2} E(\theta, K)\right) P(\theta) d\theta \right] \quad (41)$$

モデル選択は熱統計力学の考え方に則り、ベイズ自由エネルギー最小化の情報量規準で行うことができる。

更にベイズ推定では、モデルが選択される事後確率  $P(K|D)$  が評価 [(39)式] できることにも意味がある。単純にモデルの妥当性を 0/1 のデジタルで判別するのではなく、そのモデル選択の妥当性が定量的に評価できるのである。

### 3 スパースモデリング

この節では、ベイズ推定と異なる発想に基づくスパースモデリング (SpM)[10] を紹介する。

何かのグループで取った記念写真を見た時を想像してほしい。その時、背景に何が写っているかを気にせずに誰が写っているか? 何人写っているか? 誰かの影に隠れているのは誰だ? という様なデータ処理が頭の中でなされている。SpM は、人間が物体や現象を認知・識別する際に頭の中で行われる処理行為を、実験データの解析に適用するものである。つまり、実験データを与える物理現象は、より簡潔な物理モデルで説明可能であるという前提に立ち、簡潔に説明できる基底関数を選択して、その基底関数で表させる少数の主要成分 (スパース解) の抽出を可能とする。

#### 3.1 数学的準備

##### 3.1.1 $p$ 次ノルム

ノルム (norm) は、平面や空間における幾何学的ベクトルの「長さ」の概念の一般化で、 $\vec{x}$  ベクトルの  $p$  次ノルム ( $1 \leq p < \infty$ ) は以下で定義される。

$$\|\vec{x}\|_p \equiv \left( \sum_{i=1}^n |x_i|^p \right)^{1/p} = \sqrt[p]{|x_1|^p + \dots + |x_n|^p}$$

SpM では、次の 1 次、2 次ノルムと 0 次ノルムが問題となる。

- 1 次ノルム

$$\|\vec{x}\|_1 \equiv |x_1| + \dots + |x_n|$$

- 2 次ノルム

$$\|\vec{x}\|_2 \equiv \sqrt{|x_1|^2 + \dots + |x_n|^2}$$

- 0 次ノルムは注意が必要である。0 次ノルムは  $\vec{x}$  の非ゼロ要素の数を意味し、次式で書かれる。

$$\|\vec{x}\|_0 = \sum_{i=1}^n |x_i|^0 = |x_1|^0 + \dots + |x_n|^0$$

しかし、非ゼロ値  $a (\neq 0)$  では、 $a^0 = 1$  であるのに対し、本来数学的には  $0^0$  は不定である。厳密にはそうであるが、 $0^0 = 0$  として定義されたものが 0 次ノルムである。つまり 0 次ノルムは、 $\vec{x}$  の非ゼロ要素数を意味する。

##### 3.1.2 arg max (min)

arg max (arg min) は、引数  $x$  を必要とする関数  $f(x)$  が最大値 (最小値) となる時の  $x$  を返す関数で、

$$\arg \max_{x \in S} f(x) \equiv \left\{ x \in S : f(x) = \max_{y \in S} f(y) \right\}$$

$$\arg \min_{x \in S} f(x) \equiv \left\{ x \in S : f(x) = \min_{y \in S} f(y) \right\}$$

で定義される。

$\vec{x}$  の横軸に対し  $\vec{y}$  の測定データ・セットが得られた場合、そのデータ・セットを説明する物理モデルを  $f(\vec{x}; \vec{\theta})$  とすると、最小二乗法は測定値  $\vec{y}$  と物理モデル  $f(\vec{x}; \vec{\theta})$  の平均二乗

誤差を最小化した  $\vec{\theta}$  を求める行為である。よって、次式と書ける。

$$\text{最小二乗法 } \hat{\vec{\theta}} = \arg \min_{\vec{\theta}} \left( \|\vec{y} - f(\vec{x}; \vec{\theta})\|_2^2 \right)$$

#### 3.2 連立 1 次方程式を解く

植村誠先生の「天文屋のための HOW TO スパースモデリング」の第 1 章 [11] を参考に進める。

計測データから主成分を抽出する行為は、観測値から推定パラメータを求める逆問題である。SpM では計測データを  $M$  個の要素を持つ  $\vec{y}$  とし、抽出したいパラメータ  $\vec{\omega}$  が  $N$  個の要素を「持ちうる」とし、その間を

$$\vec{y} = \vec{X} \vec{\omega} \quad (42)$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1N} \\ X_{21} & X_{22} & \dots & X_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ X_{M1} & X_{M2} & \dots & X_{MN} \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_N \end{pmatrix} \quad (43)$$

の線形写像で書き表す必要がある。ここで  $\vec{X}$  は物理モデル  $f(\vec{x}; \vec{\theta})$  を行列で表したものであり、既知であれば、この線形写像は  $\vec{\omega}$  に関する連立 1 次方程式であることが分かる。

- 最小二乗法:

観測点数がパラメータ数よりも多い場合 ( $M > N$ ) ならば、次式のように最小二乗法を使えば良い。

$$\text{最小二乗法 } \hat{\vec{\omega}} = \arg \min_{\vec{\omega}} \left( \|\vec{y} - \vec{X} \vec{\omega}\|_2^2 \right)$$

- 不良設定問題:

観測点数がパラメータ数よりも少ない場合 ( $M < N$ ) は不良設定問題 (ill-posed problem) と呼ばれる。この場合最小二乗法だけでは、解が多数存在し、その中から 1 つの解を選ぶためには他にも条件が必要になる。SpM の 1 種である「L1 正則化」はその様な手法の 1 つである。

#### 3.3 L1 正則化

前述の様に不良設定問題を考える。

##### 3.3.1 0 次ノルム最小化

###### 0 次ノルム最小化

$\vec{\omega}$  がスパースであると仮定できるとき、単純に考えると  $\vec{y} = \vec{X} \vec{\omega}$

の条件を満たす複数の  $\vec{\omega}$  の解の中から、非ゼロ要素の「個数」(0 次ノルム:  $\|\vec{\omega}\|_0$ ) が最小になる  $\vec{\omega}$  を探すのが良さそうに思える。これは  $\vec{y}$  をより簡潔に説明することで、それを正則化という。

尚、0 次ノルム最小化でスパースな  $\vec{x}$  を完全に再構成できることは証明 [12, 13] されている。

しかし、0 次ノルム最小化を解くのに必要な計算量は、組み合わせ爆発が起こるため、 $\vec{\omega}$  の個数  $N$  に応じて指数的に増大し、高次元の問題に応用するのは現実的ではない。

##### 3.3.2 L1 正則化: LASSO 法

Tibshirani[14] は線形回帰の問題に 1 次ノルムの項を加える“LASSO 法” (Least Absolute Shrinkage and Selection Operator) という手法を提案した。

###### L1 正則化: LASSO 法

LASSO 法で取り扱う問題は (44) 式と書ける。

$$\hat{\vec{\omega}} = \arg \min_{\vec{\omega}} \left( \|\vec{y} - \vec{X} \vec{\omega}\|_2^2 + \alpha \|\vec{\omega}\|_1 \right) \quad (44)$$

- $\alpha = 0$ : 最小二乗法と同じ解を与えるが、それはノイズまで合わせに行く過適合 (over-fitting) な解である。
- $0 < \alpha < \infty$ : 適切な規範でスパース解を得る。
- $\alpha = \infty$ :  $\|\vec{\omega}\|_1 = 0$  が要請される。つまり非ゼロ要素はなく、データは全てノイズと解釈される。

LASSO法は、 $\vec{X}$ の列ベクトルの中で不要なものは係数 $\vec{w}$ をゼロにすることで過適合(over-fitting)を防ぎ、モデルの予測誤差を小さくすることを目的としている。また、 $\vec{w}$ の中でデータを説明するのに必要な変数だけを選択する手法(変数選択)としても使われる。

1次ノルム最小化の解=0次ノルム最小化の解

さらに、2000年代半ばには Candes & Tao[15] や Donoho & Tanner[16] が、 $M < N$ の問題(不良設定問題)で、ある条件下においては1次ノルム最小化によって得られる解が0次ノルム最小化で得られる解と同じになることを数学的に証明した。

1次ノルム最小化は従来からよく知られた最適化アルゴリズムである「線形計画法」で解くことができ、計算量の面では0次ノルム最小化より有利になる。

3.3.3 L1正則化で何故スパースな解が得られるか

直感的に理解するために、観測量 $\vec{y}$ が1つ(1次元:  $M = 1$ )でパラメータ $\vec{w}$ が2次元( $N = 2$ )の場合について考える。この場合(42)式の連立方程式は次式となる。

$$(y_1) = \begin{pmatrix} X_{11} & X_{12} \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} \rightarrow y_1 = X_{11}\omega_1 + X_{12}\omega_2,$$

もちろん方程式が1つで、未定係数が2つ( $\omega_1, \omega_2$ )なので、ユニークな解は得られない。

1次ノルム最小化の模式図

方程式は、

$$\omega_2 = \frac{X_{11}}{X_{12}}\omega_1 + \frac{y_1}{X_{12}}$$

の用に直線(図18の赤線)を表し、その直線上の任意の点で $\vec{y} = \vec{X}\vec{w}$ の関係を満たす。

一方図18の青線は内側から順に1次ノルムが大きい場合を示しているが、L1正則化は赤い直線の内、1次ノルムが最小のものを得る手法である。この例では一番内側のひし形がそれである。

図から直線と $\omega_1$ 軸と $\omega_2$ 軸と交わる点は2箇所あることが変わる。それらの交点ではともに $|\vec{w}|_0 = 1$ である。L1正則化は、その $|\vec{w}|_0 = 1$ のスパース解のうち、 $|\vec{w}|_1$ がより小さい $\omega_2$ 軸との交点が選ばれる。

この結果は、主要パラメータ抽出の観点から考えると、 $\omega_1$ よりも $\omega_2$ のほうが、 $\vec{y}$ を再現するには重要であることを意味している。

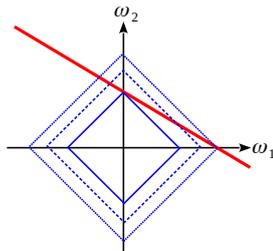


図18: 赤い直線が方程式を満たす線、青い線は1次ノルム一定の線である。

さらに(44)式について理解を深めておく。不良設定問題でなくなるが、

$$\vec{y} = \begin{pmatrix} -2 \\ 3 \end{pmatrix}, \quad \vec{X} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \quad \vec{w} = (\omega_1)$$

としよう。この場合、

$$\vec{y} - \vec{X}\vec{w} \rightarrow \begin{pmatrix} -2 - 2\omega_1 \\ 3 + \omega_1 \end{pmatrix},$$

$$\|\vec{y} - \vec{X}\vec{w}\|_2^2 = (-2 - 2\omega_1)^2 + (3 + \omega_1)^2 = 5\omega_1^2 + 14\omega_1 + 13$$

$$\|\vec{y} - \vec{X}\vec{w}\|_2^2 + \alpha\|\vec{w}\|_1 = \underbrace{5\omega_1^2 + 14\omega_1 + 13}_{\text{青}} + \underbrace{\alpha|\omega_1|}_{\text{緑}}$$

である。これらをグラフに示したものが図19である。ここでこれらのグラフ特徴をまとめておく。

- $\|\vec{y} - \vec{X}\vec{w}\|_2^2$ :  $\omega_1$ に関して上に開いた放物線関数であり、その極小値は必ず0以上である。

- $\alpha\|\vec{w}\|_1$ :

$\alpha \geq 0$ であるから、原点を極小値として上に開いたV字型関数である。この「原点を極小値」にとる特徴が $\alpha \rightarrow \infty$ で $\omega_1 \rightarrow 0$ を導き、パラメータ $\omega_1$ をゼロへ誘導しスパース化させる。

図19-(a)~(e)の赤破線で示した様に、

$$-1.4 = \arg \min_{\omega_1} \left( \|\vec{y} - \vec{X}\vec{w}\|_2^2 \right)$$

で極小値(図中の $\min_1$ )を取るのに対し、青破線で示した様に $\|\vec{y} - \vec{X}\vec{w}\|_2^2 + \alpha\|\vec{w}\|_1$ がとる極小値 $\hat{x}$ (図中の $\min_2$ )は $\alpha$ に依存する。これらの $\alpha$ 依存性を図19-(f)に示した。図19-(f)より、 $\alpha > 15$ では $\hat{\omega}_1 \approx 0$ となり、唯一非ゼロであった $\omega_1$ もゼロに漸近する。その結果、 $\vec{w} = 0$ となり、測定で得られた $\vec{y} = (-2, 3)$ が残差として残ることになり、主成分を発見できずにデータは全てノイズである解析となる。

3.4 時系列データの周期解析への応用

この節では図20の青線で示したような

$$y(t) = \sum_{j=1}^N (a_j \cos 2\pi f_j t + b_j \sin 2\pi f_j t) + n(t)$$

でモデル化出来るような時系列データを対象とする。

図20のデータは、 $f_1 = 0.8$  Hz,  $f_2 = 2.4$  Hz,  $b_{f_1} = 1.0$ ,  $a_{f_2} = 0.8$  で、 $\sigma_{\text{noise}} = 0.2$ の標準偏差を持つ正規分布ノイズ $N(0, \sigma_{\text{noise}})$ で、

$$y(t) = b_{f_1} \sin 2\pi f_1 t + a_{f_2} \cos 2\pi f_2 t + N(0, \sigma_{\text{noise}})$$

で合成したものである。

3.4.1 フーリエ変換基底

このような振動性の時系列データから振動周波数を推定する場合、フーリエ変換がよく用いられる。しかしフーリエ変換は完全性が担保される結果、データに必ず重畳するノイズまでも周波数分解してしまう。

ここではフーリエ変換の基底関数を用いてSpMを試みる。時系列データが $\Delta t$ の等時間ステップで $M$ 個のデータ点として計測されたとすると、フーリエ変換の周波数分解能 $\Delta f$ は、 $\Delta f \equiv 1/(M\Delta t)$ である。ここで $f_n \equiv n\Delta f$ とすると、 $N = M/2$ としてフーリエ変換は

$$y(t) = \sum_{n=1}^N (a_n \cos 2\pi f_n t + b_n \sin 2\pi f_n t) \quad (45)$$

である。 $y(t)$ の時系列データを $\vec{y} \equiv \{y_1, y_2, \dots, y_M\}$ とし、 $\vec{w} \equiv \{a_1, a_2, \dots, a_N, b_1, b_2, \dots, b_N\}$ とすると、(45)式は、

$$\vec{X} \equiv \begin{pmatrix} \cos 2\pi f_1 t_1 & \dots & \cos 2\pi f_N t_1 & \sin 2\pi f_1 t_1 & \dots & \sin 2\pi f_N t_1 \\ \cos 2\pi f_1 t_2 & \dots & \cos 2\pi f_N t_2 & \sin 2\pi f_1 t_2 & \dots & \sin 2\pi f_N t_2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cos 2\pi f_1 t_M & \dots & \cos 2\pi f_N t_M & \sin 2\pi f_1 t_M & \dots & \sin 2\pi f_N t_M \end{pmatrix}$$

を用いて、 $\vec{y} = \vec{X}\vec{w}$ と書ける。ここで $\vec{X}$ を図21に示した。この縦一列が各周波数の基底関数である。

3.4.2  $\alpha$ に依存した非ゼロ要素の変化

図22に $\vec{w}$ 係数(=  $\{a_1, a_2, \dots, a_N, b_1, b_2, \dots, b_N\}$ )の非ゼロ要素が $\alpha$ の減少とともに分岐する様子を示した。

$\alpha$ の大きい右側では適切に2つの振動モードが抽出されているのに対し、 $\alpha$ の小さい左側では、小さな非ゼロ要素が多数分岐し始めることが分かる。その領域は、データに重畳したノイズ成分をフーリエ基底で再現しようとする過適合領域である。

図23は、 $\alpha$ を横軸(対数)にして、○は $\vec{w}$ の非ゼロ要素数 $\|\vec{w}\|_0$ の変化を、●は得られたスパース解 $\hat{\vec{w}}$ による再現データ $\vec{X}\hat{\vec{w}}$ と解析対象データ $\vec{y}$ との残差二乗平均平方根(RMSD: Root Mean-Square of Deviations)の変化を示した。水平破線はデータに重畳させた正規ノイズの標準偏差である。

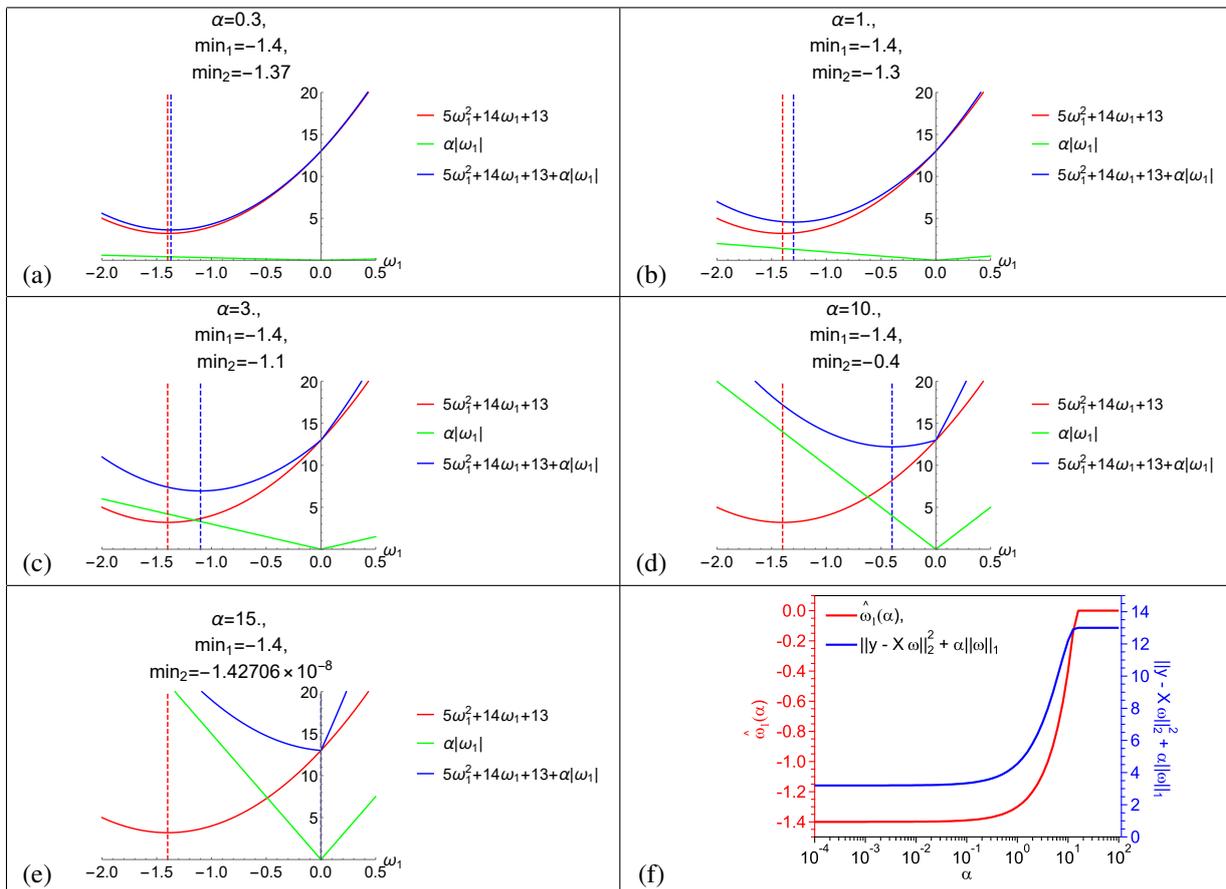


図19: (a)~(e)  $\alpha = 0.3 \sim 15$ . (f)  $\alpha$  依存性

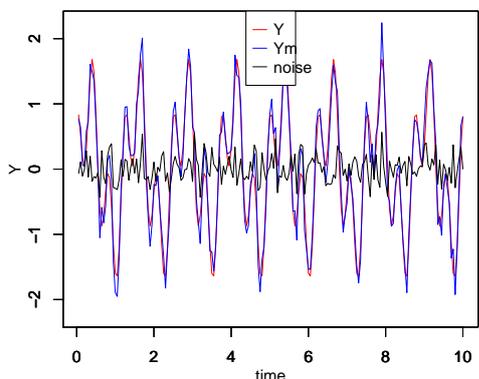


図20: 時系列データの周期解析対象のデータ

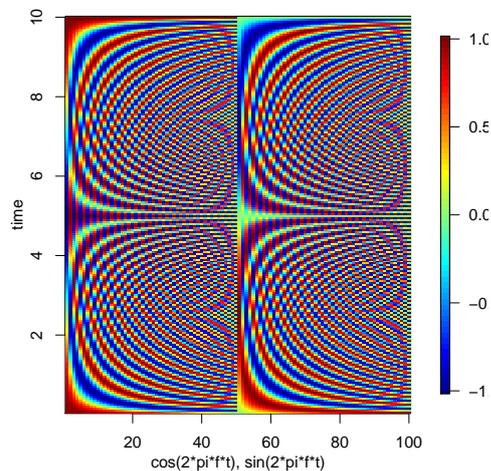


図21: 線形写像行列のイメージプロット

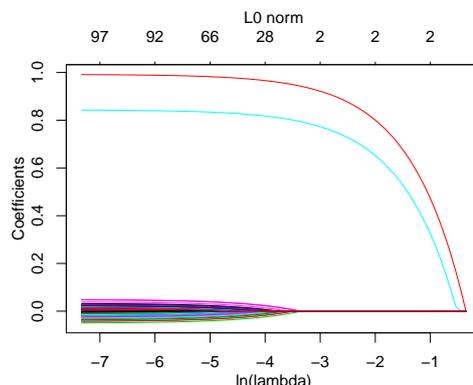


図22:  $\bar{\omega}$  の  $\alpha$  に依存した分岐の変化。横軸は  $\alpha$  を対数目盛りで示してある。

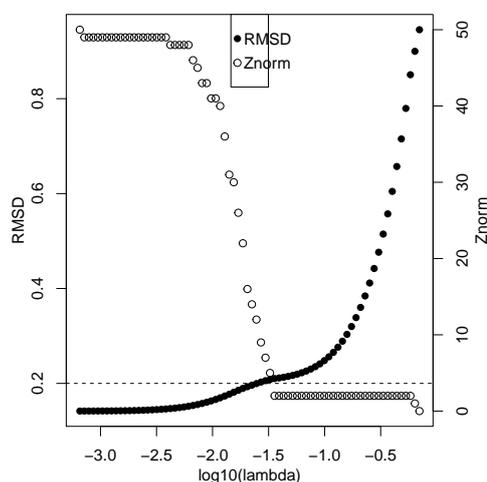


図23: 0次ノルムと RMSD の  $\alpha$  による変化のグラフ

$\log_{10}(\alpha) > -1.5$  では2つの振動モードが抽出され、 $\alpha$  の減少とともに RMSD が減少して再現性が向上していることがわかる。一方、 $\log_{10}(\alpha) < -1.5$  ではノイズを再現するため0次ノルムが急激に増加し、RMSD が重畳させた正規ノイズの標準偏差を下回る過適合になっていることが分かる。

### 3.4.3 交差検定

SpM では、データを説明する成分の冗長性とデータの再現性両方を考慮した適切な情報量規準にしたがって  $\alpha$  を決定する。 $\alpha$  を決める方法として、交差検定 [17] などともよく用いられる。

## 3.5 EXAFS 解析への応用

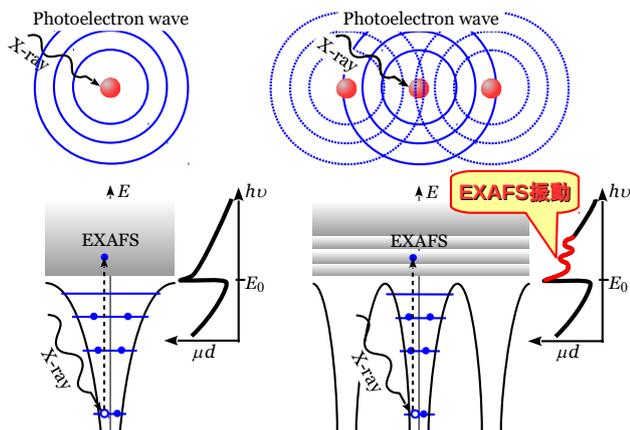


図24: 広域 X 線吸収微細構造 (EXAFS)

LASSO 法などの L1 正則化による SpM は、抽出したい特徴量が少数 (スパース) である場合、従来解析法を凌駕する強力な解析法である。

赤井等はその優位性に注目し、放射光を用いて計測される広域 X 線吸収微細構造 (EXAFS: Extended X-ray Absorption Fine Structure) の解析に L1 正則化を適用 [18] した。

図24に示した様に、EXAFS は、X 線吸収端以上への電子励起による光電子波の近接原子による散乱と干渉によって、吸収スペクトルに現れる振動構造である。EXAFS では、X 線吸収端エネルギーで選択された注目原子近傍の原子による散乱が重要であるが、本来近接原子の配位は、物質の化学結合や構造を反映して、距離に対してスパースである。赤井等はそのスパース性に注目した。

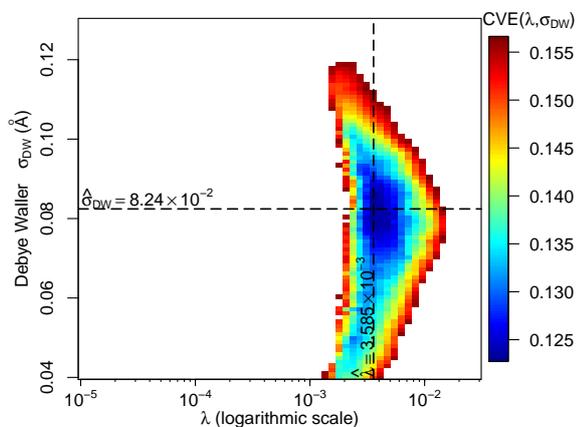


図25: 交差検定誤差のヒートマップ

EXAFS では、注目原子近傍の近距離構造の動径分布関数 (RDF: Radial Distribution Function) を解明する目的で解析されるが、近接配位原子の構造ゆらぎや可動性を表すデバイ・ワラー因子を推定するには、本来未解明の近距離構造を仮定する必要があった。それに対し、L1 正則化を用い

た解析では、一切の事前仮定を必要とせず、データだけから RDF とデバイ・ワラー因子の推定に成功した。その例を図25に示した。この横軸はスパース性を制御する  $\alpha$  を対数で示し、縦軸はデバイ・ワラー因子で、交差検定誤差をヒートマップで示したものである。交差検定誤差の最小化によってデバイ・ワラー因子が推定可能である。

## 謝辞

本文の一部は、JST, CREST, JPMJCR1861 の支援を受けたものである。

## 参考文献

- [1] T. Bayes and R. Price. *An Essay towards Solving a Problem in the Doctrine of Chances*. *Phil. Trans. Roy. Soc.*, **53**, pp. 370–418, (1763).
- [2] K. Nagata, S. Sugita, and M. Okada. *Bayesian spectral deconvolution with the exchange Monte Carlo method*. *Neural Networks*, **28**, p. 82, (2012).
- [3] P. J. Green. *Reversible jump Markov chain monte carlo computation and Bayesian model determination*. *Biometrika*, **82**, p. 711, (1995).
- [4] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. *Equation of state calculations by fast computing machines*. *J. Chem. Phys.*, **21**, pp. 1087–1092, (1953).
- [5] C. Andrieu, N. D. Freitas, A. Doucet, and M. I. Jordan. *An Introduction to MCMC for Machine Learning*. *Machine Learning*, **50**, pp. 5–43, (2003).
- [6] S. Aihara, M. Hamamoto, K. Iwamitsu, M. Okada, and I. Akai. *High precision modeling of a damped oscillation in coherent phonon signals by Bayesian inference*. *AIP Adv.*, **7**, p. 045107, (2017).
- [7] R. H. Swendsen and J-S. Wang. *Replica Monte Carlo Simulation of Spin-Glasses*. *Phys. Rev. Lett.*, **57**, pp. 2607–2609, (1986).
- [8] K. Hukushima and K. Nemoto. *Exchange Monte Carlo Method and Application to Spin Glass Simulations*. *J. Phys. Soc. Jpn.*, **65**, pp. 1604–1608, (1996).
- [9] Y. Mototake, M. Mizumaki, I. Akai, and M. Okada. *Bayesian Hamiltonian Selection in X-ray Photoelectron Spectroscopy*. *J. Phys. Soc. Jpn.*, **88**, p. 034004, (2019).
- [10] 岡田真人. 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」, 2014. <http://sparse-modeling.jp/>.
- [11] 植村誠. 天文屋のための HOW TO スパースモデリング, 連立 1 次方程式を解く, 2016. [http://home.hiroshima-u.ac.jp/uemuram/How2SM/01\\_demo.pdf](http://home.hiroshima-u.ac.jp/uemuram/How2SM/01_demo.pdf).
- [12] 和田山正. 圧縮センシングにおける完全再現十分条件について. *日本神経回路学会誌*, **17**, pp. 63–69, (2010).
- [13] 田中利幸. 圧縮センシングの数理. *IEICE Fundamentals Review*, **4**, pp. 39–47, (2010).
- [14] R. Tibshirani. *Regression shrinkage and selection via the lasso*. *J. Roy. Stat. Soc. B*, **58**, pp. 267–288, (1996).
- [15] E. J. Candes and T. Tao. *Near-optimal signal recovery from random projections: Universal encoding strategies?* *IEEE Transactions, Information Theory*, **52**, pp. 5406–5425, (2006).
- [16] D. L. Donoho. *Compressed sensing*. *IEEE Transactions, Information Theory*, **52**, pp. 1289–1306, (2006).
- [17] K. P. Murphy. *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series)*. The MIT press, 2012. ISBN 0262018020.
- [18] I. Akai, K. Iwamitsu, Y. Igarashi, M. Okada, H. Setoyama, T. Okajima, and Y. Hirai. *Sparse Modeling of an Extended X-Ray Absorption Fine-Structure Spectrum Based on a Single-Scattering Formalism*. *J. Phys. Soc. Jpn.*, **87**, p. 074003, (2018).