



特集記事・5

鉄鋼業におけるAI・IoT技術の最前線

理論式の高精度化のためのパラメータ推定

Parameter Optimization for Accuracy Improvement of Theoretical Formula

名古屋大学
大学院工学研究科
教授

足立吉隆
Yoshitaka Adachi

名古屋大学
大学院工学研究科
研究員

Zhi-Lei Wang

1 はじめに

材料工学では各種理論式がこれまでに提唱されている。例えば、強度 (σ) については転位の摩擦抵抗の項 (σ_0) に加えて固溶強化 (1式, C:固溶量)、析出強化 (2式, G:剛性率, b:バーガースベクトル, L:粒子間隔, r:粒子サイズ, f:体積率)、転位強化 (3式, ρ :転位密度)、結晶粒微細化強化 (4式, d:結晶粒径)、複合組織強化にそれぞれ次の理論式が提唱されている。目的変数であるこれらの強化量を表現する理論式には、転位密度 (ρ) や結晶粒径 (d) などの説明変数に加えて、パラメータ (A, α , kで示される) が存在し、説明変数の測定精度とともにパラメータの最適化が目的変数を高精度で推定する際に重要となる。

$$\sigma = \sigma_0 + AC^{1/2} \dots\dots\dots (1)$$

$$\sigma = \sigma_0 + \alpha Gb / L, L = r(2/f)^{1/2} \dots\dots\dots (2)$$

$$\sigma = \sigma_0 + \alpha Gb \rho^{1/2} \dots\dots\dots (3)$$

$$\sigma = \sigma_0 + kd^{-1/2} \dots\dots\dots (4)$$

また、強化機構の加算側は一般的に (5) 式で表現される場合が多いようである。複数の強化機構が重畳して生じる強化量は各強化量の単純加算ではなく、各強化量に係数 (α , β , γ) とべき指数 (l, m, n) をつけて加算される¹⁾。実用材料では複数の強化機構が重畳して強化が図られており、その強化量を精度よく推定するためには、各強化量を精度良く求めることの重要性和同等に、加算則の係数とべき指数といったパラメータを精度良く求めることが大切である。

$$\sigma^l = \alpha \sigma_1^m + \beta \sigma_2^n \dots\dots\dots (5)$$

このような目的変数～説明変数の関係を推定する方法には

大きく分けて、1. 最小二乗法、2. 最尤法、3. 事後確率に基づく方法 (MCMC法)、4. ニューラルネットワーク法 (ANN) などの非線形機械学習法がある²⁾。最尤法はさらに分類され、(1) 誤差項が等分散正規 (ガウス) 分布の線形回帰 (一般線形モデル, linear model: lm)、(2) 誤差項が様々な分布 (ポアソン分布やガンマ分布) の線形回帰 (一般化線形モデル, Generalized linear model: glm)、がある。後述するように、最小二乗法と最尤法の一般線形モデルは等価である。1～3の推定方法は、理論式のパラメータを最適化することによりモデルの精度を高めようとする手法であり、物理モデルに基づいて目的変数～説明変数の関係を推定しようというものである。一方、4の推定方法は、柔軟な非線形関数を複数組み合わせ、その関係を推定しようとする順解析モデルであり、究極的な近似により推定精度を高めていることが特徴である。4の推定方法には、誤差項が様々な分布の非線形回帰 (一般化加法モデル, Generalized additive model: gam) やANNのような機械学習法がある。gamやANNは、まったく目的変数～説明変数の関係が不明な場合でも、柔軟な非線形 (伝達) 関数による究極的な近似で高精度順解析モデルを得ることができる特長を有する半面、材料工学の理論式に全く基づかないで関係を表現していることに留意する必要がある。他の表現をすると、1～3では因果関係に基づいた“予測”が行われるのに対して、4では相関関係を含めた全関係を使って“推定”が行われると説明できよう。

本解説では、目的変数～説明変数の関係を推定するこれらの各手法の特徴を、数学的な厳密性よりも、諸課題への適用、簡潔性を優先して説明することを試みる。

2 各種順解析モデル

本章では、目的変数 (z) と説明変数 (x, y) が (6) 式の関係式に従っているものと仮定して話を進める。a, b, cは係数

であり、 m, n はべき指数である。 ϵ は誤差項であり、後述するようにモデルによってその分散が異なるものと仮定する。 z の分散は ϵ で表現される。

$$z = ax^m + by^n + c + \epsilon \quad (6)$$

2.1 確率分布関数と確率密度関数

事前準備として、ここでは様々な確率関数について説明する²⁾。ある事情 x が起こる頻度の程度は、離散型データについては確率分布関数 $p(x)$ 、連続型データについては確率密度関数 $f(x)$ で表わされる。前者は、例えばさいころを振った時にある数 (x) がでる頻度 P の分布を表すのに対して、後者は確率分布関数の導関数であり、ある範囲 $a \leq x \leq b$ の確率 P は

$$P(a \leq x \leq b) = \int_a^b f(x) dx \quad (7)$$

で与えられる。

確率分布関数、確率密度関数 (煩雑になるのを防ぐため、以降とくにその違いを説明する時以外はまとめて確率関数 $p(x)$ と表現することにする) は複数あるが、本解説に関係するものを図1に示す。

ポアソン分布は離散型データが対象であり、パラメータ λ のみを使って確率変数 x が生じる確率が与えられる。ポアソン分布の平均、分散ともに λ に等しいという特徴を有している。ランダムな事情が単位時間に起こる回数を表わす。

正規分布は、ガウス分布とも呼ばれ、連続型データを対象とし、平均 μ と分散 σ^2 の二つのパラメータを使って表わされる。 $\mu = 0, \sigma^2 = 1$ の時、特に標準正規分布と呼ばれる。平均 μ からのずれが $-1\sigma \sim +1\sigma$ の範囲に x が含まれる確率は 68.27% となる ($\pm 2\sigma$ の場合は 95.45%, $\pm 3\sigma$ の場合は 99.73%)。自然現象で正規分布に従うものは多いとされる。確率変数 x そのままでは正規分布に従わない場合でも、その

対数が正規分布に従う場合も多いとされている。このような確率変数の対数が正規分布に従う場合を対数正規分布という。確率変数の範囲は $-\infty \sim +\infty$ と想定する。ランダムなノイズの分布を表す。

ガンマ分布は、連続型データを対象とし、形状母数 k と尺度母数 θ の二つのパラメータで表わされる。確率変数の範囲が $x > 0$ を対象とする。ガンマ分布の平均、分散はそれぞれ $k\theta, k\theta^2$ となる。期間 θ ごとに 1 回ぐらい起こる事象が k 回起こるまでの時間の分布を表す。

2.2 尤度

パラメータ $A = (\mu, \sigma^2)$ を変化させたときにデータ $X = (x_1, x_2, x_3, \dots, x_n)$ が生じる確率 (密度) の分布は、(8) 式で示すように各データが生じる確率 $p(x_i)$ の同時確率で与えられる。いま、データ X が正規分布 (平均 μ 、分散 σ^2) に従うとすると、

$$p(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right) \quad (8)$$

であるので、

$$P(x_1, x_2, \dots, x_n) = p(x_1)p(x_2), \dots, p(x_n) \\ = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right) \quad (9)$$

となる。

ここまでの説明ではパラメータを μ, σ^2 と固定した場合のデータ X が生じる同時確率を考えたが、考え方を考えて、データ X を固定としパラメータを変更させることを考えると、あるパラメータの時に尤もらしいデータの分布が得られる (式 (10))。式の右辺は (9) 式と同じである。これを尤度 $L(\mu, \sigma)$ という。例えば、データ X の分布が図2 (a) のようになっている時に、パラメータ $\mu = 0$ の時の確率密度関数は実線で表される。この時尤度は小さい。 $\mu = 3$ になる (図2 (b)) と尤度が最も高くなる。この尤度が最も高くなる時のパラメータが最適値であると考える手法を最尤法と呼ぶ。

$$L(\mu, \sigma) = \prod_{i=1}^n p(x_i | \mu, \sigma^2) \\ = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right) \quad (10)$$

この尤度の対数をとると右辺の計算が積から和になり計算しやすくなる ((11) 式)。このような尤度を対数尤度と呼ぶ。

$$\log\left(\prod_{i=1}^n p(x_i | \mu, \sigma^2)\right) \\ = \log\left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right)\right) \\ = \sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right)\right) \quad (11) \\ = -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log(\sigma^2) - \frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2}$$

確率(密度)関数

- ポアソン分布 (パラメータ: $\lambda = \mu$ (平均) = σ (分散), 離散データ, ランダムな事象が単位時間に起きる回数)

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

- 正規分布 (パラメータ: μ, σ^2 , 連続データ, ランダムノイズ)

$$p(x_i) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}$$

- ガンマ分布 (パラメータ: t, n , 期間 θ ごとに 1 回ぐらい起こるランダムな事情が k 回起こるまでの時間の分布)

$$p(x) = x^{k-1} \frac{e^{-x/\theta}}{\Gamma(k)\theta^k} \quad (x \geq 0)$$

図1 代表的な確率 (密度) 関数

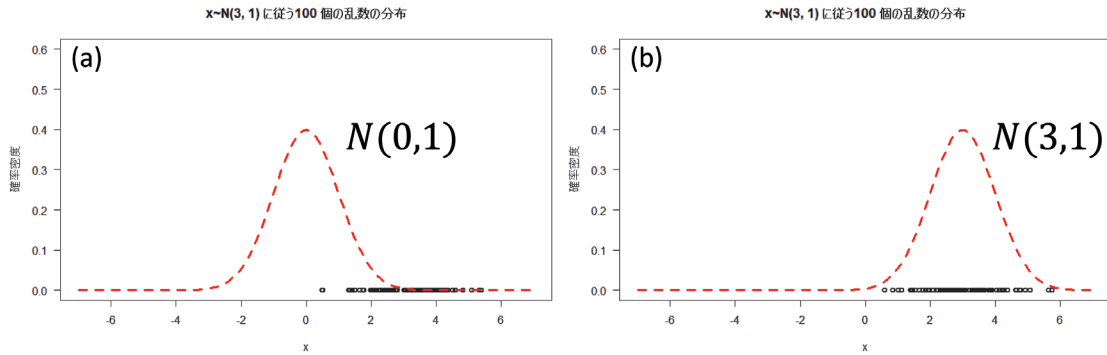


図2 最尤法による正規分布のパラメータの平均値の決定

上の最終式の第一項はパラメータ μ と σ^2 に関係しないので、最大化のときは無視できる。すなわち、

$$-\frac{N}{2} \log(\sigma^2) - \frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2} \dots\dots\dots (12)$$

という関数を最大化するパラメータ μ と σ^2 を見つけるということになる。

以下では確率関数と尤度の基本概念を使って、回帰式のパラメータ最適化について説明を進める。後半になるほど一般化された回帰式の導出法となる。

2.3 最小二乗法によるパラメータ推定

パラメータ推定法で最もよく用いられているのが最小二乗法である。いま入力変数 x と出力変数 y の関係 (図3 (a)) が (13) 式で与えられるとした時に、誤差関数 E (14) 式をもっとも最小にするパラメータ a 、 b を求める。

$$y = f(x) = ax + b \dots\dots\dots (13)$$

$$E = \sum_i (y_i - f(x_i))^2 \dots\dots\dots (14)$$

2.4 誤差項が等分散正規 (ガウス) 分布の線形回帰 (一般線形モデル、linear model : lm)

最小二乗法では $y_i = f(x_i)$ は一定値であると考えた。ここでは $y_i = f(x_i)$ はある分布を持っていると考える (図3 (b))。いま x_i を固定した時に、 y_i 値は平均 μ 、分散 σ^2 の正規分布を持っていると仮定する (このことを $N(\mu, \sigma^2)$ と書く)。また分散はどのデータに対しても一定値 (等分散) であるとする。目的変数の平均 μ と入力変数間の関係は (15) 式で示される線形であるとする。

$$y_i = f(x_i) = \mu = ax + b \dots\dots\dots (15)$$

この場合対数尤度は図1中の該当式 (ただし、 $x_i \rightarrow y_i$) で与え

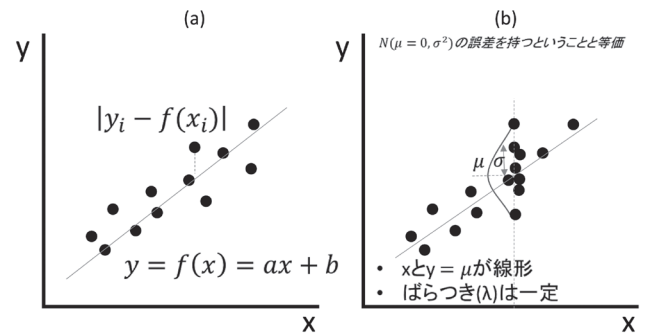


図3 最小二乗法 (a) と誤差項が等分散正規分布の一般線形モデル (b) によるパラメータ推定の比較

られる。よって尤度を最大にする平均 $f(x_i) = \mu$ とは $\sum_i (y_i - \mu)^2$ を最小にするものということになり、最小二乗法と等価であるといえる。

ここでは単純な線形回帰を考えているので、 $g(f(x_i))$ で与えられるリンク関数 g は定数1 (恒等リンク関数という) である。ただし、リンク関数がいつも固定というわけではない。

2.5 誤差項が様々な分布 (ポアソン分布やガンマ分布) の線形回帰 (一般化線形モデル、Generalized linear model : glm)

データ分布が正規分布と仮定した場合の線形回帰が一般線形回帰であるが、データ分布は様々な分布に従う可能性があり、線形回帰の方法を拡張する必要がある。ここではデータがポアソン分布に従う場合を考える。

ポアソン分布の確率密度関数は図1に掲げているように、

$$p(y_i) = \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!} \dots\dots\dots (16)$$

であるので、対数尤度は次式で与えられる。

$$\log \prod_i p(y_i) = \log \prod_i \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!} = \sum_i (y_i \log \lambda - \lambda - \sum_k y_k \log k) \dots\dots\dots (17)$$

前述したようにポアソン分布の平均 μ = 標準偏差 $\sigma = \lambda$ である。

目的変数の平均 $\mu = \lambda$ と入力変数間の関係は (18) 式で示される線形であるとする。

$$y_i = f(x_i) = \mu = \lambda_i = \exp(a + bx_i) \quad (18)$$

この場合 $f(x_i)$ と $\exp(a + bx_i)$ が比例関係であり、 $\exp$ のように比例関係になるように変換する関数を基底関数という。両辺の対数をとった場合は、 $g(f(x_i))$ で表されるリンク関数 g は対数 $\log$ である。ただし、リンク関数がいつも固定というわけではない。

2.6 誤差項が様々な分布の非線形回帰 (一般化加法モデル、Generalized additive model : gam)

一般化線形回帰モデルでは目的変数と入力変数の関係を線形とした。しかし、より複雑な関係である場合もあり、それを取り扱うためには関係式 $f(x_i)$ を非線形とする拡張が必要となる。非線形を表現する関数は複数あり^{*1}、例えば、高次まで項を並べた多項式回帰 (例: $y = a_1 + a_2x + a_3x^2 + a_4x^3 + a_5x^4 \dots$) による手法があるが、全データ範囲に対して一つの回帰式を当てはめようとするところに無理がある。ここでは gam で広く使われている非線形回帰をする平滑化スプライン関数を用いる。平滑化スプライン関数は、データにできるだけフィットさせながらも、過度なフィッティングは抑制し平滑化とバランスさせながら曲線を見つける手法で、後述するニューラルネットワークモデルの最適化にも使われるペナルティ付き損失関数 (具体的には一般化交差検証値: GCV) を最小化するように曲線を求める手法である。詳細は専門書を参照願いたい。平滑スプライン法では線形回帰などのように平均や分散といったパラメータはなく、これをノンパラメトリック法という (線形回帰の場合はパラメトリック法)。リンク関数は恒等リンク関数や対数など複数のものが使われる。

2.7 事後分布を使ったパラメータ最適化 (MCMC 法)

上述した回帰手法はいずれも最尤法に基づいて最適なパラメータが求められている。最尤法では、より大きなデータを対象にした真の確率密度関数が存在することは一切考慮しないで計算されるのに対して、ベイズ定理 (あるいはベイズ更新) を使う MCMC 法では限られたデータに対するパラメータは一義的には決まっておらず更新されるほど真の確率密度関数に近づきより精度の高いパラメータ値が得られるという考え方に基づいている。ベイズ定理によるデータ y が与えら

れた時のパラメータ a の事後分布 $p(a|y)$ は (19) 式で与えられる。

$$p(a|y) = \frac{p(y|a)p(a)}{p(y)} \quad (19)$$

ここで $p(y|a)$ は尤度、 $p(a)$ は事前分布である。 $p(y)$ は y が生じる確率密度であるが、用いるデータの中では常に一定であるので、パラメータの変化に伴う事後分布の変化を議論する際には重要ではない。 $p(a)$ は当初不明であるので、例えば一様分布とする。この場合尤度と事後分布は同じになる。しかし、一度この計算を行うと、求められた $p(a|y)$ を $p(a)$ に代入して、新たなパラメータ a' に対して再度ベイズの定理を計算する (ベイズ更新という) ことにより、より信頼度の高いパラメータの確率密度関数が得られる。このベイズ更新を繰り返すことにより、より精度が高まった確率密度関数より次に述べる事後期待値 (EAP) よりパラメータが求まる。

事後分布が高くなる方向にパラメータを変更しながら最適値を求める手法がマルコフ連鎖モンテカルロ法 (MCMC) である。MCMC 法は大別して、パラメータをランダムに変更しながら事後分布が低くなる時はある確率で棄却するメトロポリス法や、 x , y 方向にのみ条件付確率に従っていくことを許可するギブスサンプラー法がある。いずれも、ひとつ前の状況と比較して事後分布が上がるか、下がるかを検証している。

一般化線形回帰モデルのところで設定したものと同一環境を考える。すなわち、目的変数と入力変数は線形の線形の関係 (係数 a 、切片 b) にあり、目的変数 (y) は正規分布に従うものと仮定する。

したがって、

$$p(a|y) = \frac{p(y|a)p(a)}{p(y)} = \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(y-\mu)^2}{\sigma^2}\right) p(a)}{p(y)} \quad (20)$$

が事後分布である。

a の事後期待値 (Expected a Posterior : EAP) は、 $f(a|y)$ $\times a$ を全ての a に対して積算した値となり、次式で与えられる。

$$\begin{aligned} \text{EAP 推定量} &= \int (p(a|y) \times a) da \\ &= \int \left(\frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(y-\mu)^2}{\sigma^2}\right) p(a)}{p(y)} \times a \right) da \end{aligned} \quad (21)$$

この積分計算を直接計算するのは困難な場合が多いので直接

*1: 局所回帰関数、平滑化スプライン、B スプライン、自然スプラインなどがある。

計算するのではなく、 a をMCMCで変更しながらEAP推定値を求める。

MCMC法によるシステムパラメータの実際の更新手順を以下にまとめる。ここではメトロポリス法を対象に説明する。

- 更新する前 ($t=t$) のシステムパラメータ a_t を変えてみて尤度 $p(D|a_t)$ をしらみつぶしに計算する。ベイズ定理 $p(a_t|D) \propto p(D|a_t) \cdot p(a_t)$ に基づいて事後分布 $p(a_t|D)$ を計算する。 a_t となる分布 (事前分布) $p(a_t)$ は当初不明であるので、例えば一様分布とする。一様分布は、データの範囲が $a < D < \beta$ の場合、離散型データ、連続型データともに確率分布、確率密度関数は $\frac{1}{\beta-a}$ で与えられる。
- 変更したシステムパラメータ a' に対して尤度 $p(D|a')$ を計算する。ベイズ定理 (ベイズ更新) に基づいて事後分布 $p(a'|D)$ を計算する。
- 更新前後の事後分布の比 $a=p(a'|D)/p(a_t|D)$ を求める。区間 $Un(0,1)$ での一様分布 u と比較し、 $a > u$ であれば $a_{t+1}=a'$ として採用。 $a < u$ であれば $a_{t+1}=a$ として更新しない。これを繰り返す。

パラメータを更新する最初の頃は事前分布 $p(a_t)$ を適当な確率関数を与えたので信用できない事後確率が回答されるが、更新を繰り返すうちにある事後確率に収れんする。事後分布が収れんした以降の範囲 (その臨界回数をburn-inという) のデータのみを使って、事後分布の頻度グラフを作成し、最も頻度が高くなる事後分布を与えるパラメータ a_t を最適値とする。

2.7 ハイパーパラメータを最適化したニューラルネットワーク回帰 (ANN)

最尤法、MCMC法では、理論式のパラメータを確率密度関数を使って評価し最適値を見つける手法を通して、目的変数～入力変数の関係を精度良く表現しようとする手法であった。そこには理論式に基づく物理が背景に存在した。しかし、複雑系では、その関係が複雑であり、理論式では対応しきれない場合がある。そのような場合に、因果関係だけではなく相関関係もすべて使って究極的な近似式を作り上げようとするアプローチとして様々な識別器を使った機械学習法がある。材料学的な理論を背景に持たないまま、非線形の関数を組み合わせて、目的変数～入力変数の関係を究極的な近似式で表現する手法といえる。したがって、現象の理解には向かない手法であるが、現象の推定には絶大な威力を発揮する。

ここでは代表的な機械学手法の例としてニューラルネットワーク法を取り上げる。ニューラルネットワーク法の詳細は専門書に譲るとして、要点は、入力変数と出力変数を直接重み係数などで結びつけるのではなく、中間層を設けてその中のノードと入力変数間の関係式をつくり、またノードと出力

変数間の関係式を作り、できるだけ入力変数と出力変数の相関を高めるように、中間層、ノードの数、重み係数などのハイパーパラメータの最適化を行い精度の高い順解析モデルを作ろうとするということである。入力変数と中間層のノード間の関係は線形ではなく、シグモイド関数やtanhなどの非線形伝達関数が用いられる。データへのフィッティング精度が高い反面、全体的な傾向を見落とす過学習が生じることが最大の課題であり、それを抑制するために、ペナルティ付き損失関数を最小化することによりフィッティングとスムージングのバランスをとっている。そのバランスは重み減衰率係数で与えられ、それもハイパーパラメータの一つである。

柔軟にデータフィッティングをするニューラルネットワークモデルとは言え、ハイパーパラメータの最適化がそのモデルの精度を決める。ハイパーパラメータの最適化には、ベイズ的最適化といった効率的探索法や、グリッドサーチ法などの全探索法があるが、詳細は別報告を参照願いたい。

3 疑似データを使った各回帰手法の比較検証

$x \sim N(0,1)$, $y \sim N(0,1)$ ($N(0,1)$: 標準正規分布) のデータが (22) 式の関係式に従うことが分かっているという前提で、一般化線形モデル、一般化加法モデル、ベイズ更新法 (MCMC)、を使った手法でパラメータ a , b , c , m , n を求め、精度を比較する。

$$z = ax^m + by^n + c + k * rnorm(x, 0.001) \dots\dots\dots (22)$$

$$k = 1, 10, 100, 1000$$

以下の計算では係数、べき指数は未知であるが、疑似データを作成するにあたっては a , b , c , m , n 値はそれぞれ 5, 10, 10, 3, 2 とした。すなわちこれらの数値を以下の手法で推定することを試みた。あわせて、 x , y 間の関係が全く不明な場合に、その相関関係を推定する順解析モデルをニューラルネットワーク法で推定した。計算はプログラミング言語 R にて行った。一般化線形モデル、一般化加法モデル、ベイズ更新法の計算では、データは正規分布 (rnorm) のノイズを仮定した。ノイズの大きさを変えるために、係数 k を 1~1000 の範囲で変更したデータを作成した。リンク関数は恒等リンク関数を用いた。べき指数 m , n の最適値を求めるにあたっては、for 文で m , n 値を 1 のステップで変化させ、その与えた m , n 値の条件で赤池情報量規準 (AIC)³⁾ に基づいて最適な a , b , c 値を求め、それを各 m , n 値の条件で繰り返し計算し、最も AIC が小さくなる時の m , n 値と a , b , c 値を最適値とした。

一般化線形モデル一般化加法モデルには mgcv パッケージ^{4,7)}、ベイズ更新法には MCMCpack パッケージ⁸⁾、ニュー

ラルネットワーク法にはnnetパッケージ⁹⁾を利用した。ニューラルネットワークのハイパーパラメータ最適化にはrBayesianOptimizationパッケージ¹⁰⁻¹³⁾を利用した。

データノイズが今回最も小さい $k=1$ の場合(図4) の実際の z 値と各モデルで推定した z 値 (glm-estimatedなど) の関係を、推定アルゴリズムごとに図5 (a) ~ (d) に示す。(a) glmモデル、(b) gamモデル、(c) MCMCモデルともに大差

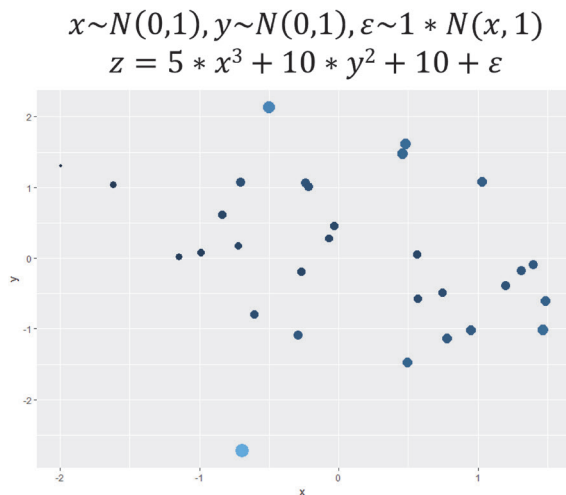


図4 ノイズを含む入力データ (x,y) と出力データ (z) の関係の一例

なく相関係数 (CC) 0.9990の高精度で z 値の推定ができている。ANNモデル (d) の場合は、やや精度が落ちる (CC = 0.9834) がモデル式を全く与えていないにも拘らずこの精度が出ていることは注目に値する。

データノイズが増えて $k=10$ とした場合の結果を図6 (a) ~ (d) に示す。いずれのモデルも推定精度が落ちている。glm, gam, MCMCモデルに比べてANNモデルの精度が落ちる傾向はノイズが小さい場合と変わらない。

$k=1000$ とノイズが大きくなると、もはやglm, gam, MCMCモデルの推定精度 (CC) は0.5程度まで急落する(図7 (a) ~ (c))。一方、ANNモデル(図7 (d)) の精度は依然CC = 0.8226を保っていることは意外である。

推定精度 (CC) を各モデルで比較した結果を図8に示す。ノイズが小さい時は、モデル式の係数推定法であるglm, MCMCモデル、モデル式を想定しないgam, ANNモデルともにほぼ同じ推定精度であるが、ノイズが大きくなるとgam, ANNの方が推定精度がよい。入力、出力データ間の関係をモデル式に基づいてその係数を最適化するという物理モデルに基づいた手法はデータノイズが小さい範囲では有益であるが、データノイズが大きい場合には究極的な近似式を与えるANNモデルの方がむしろ精度が高くなっている。ANN解析に要する時間は上に述べた各種係数推定法と大差なく、

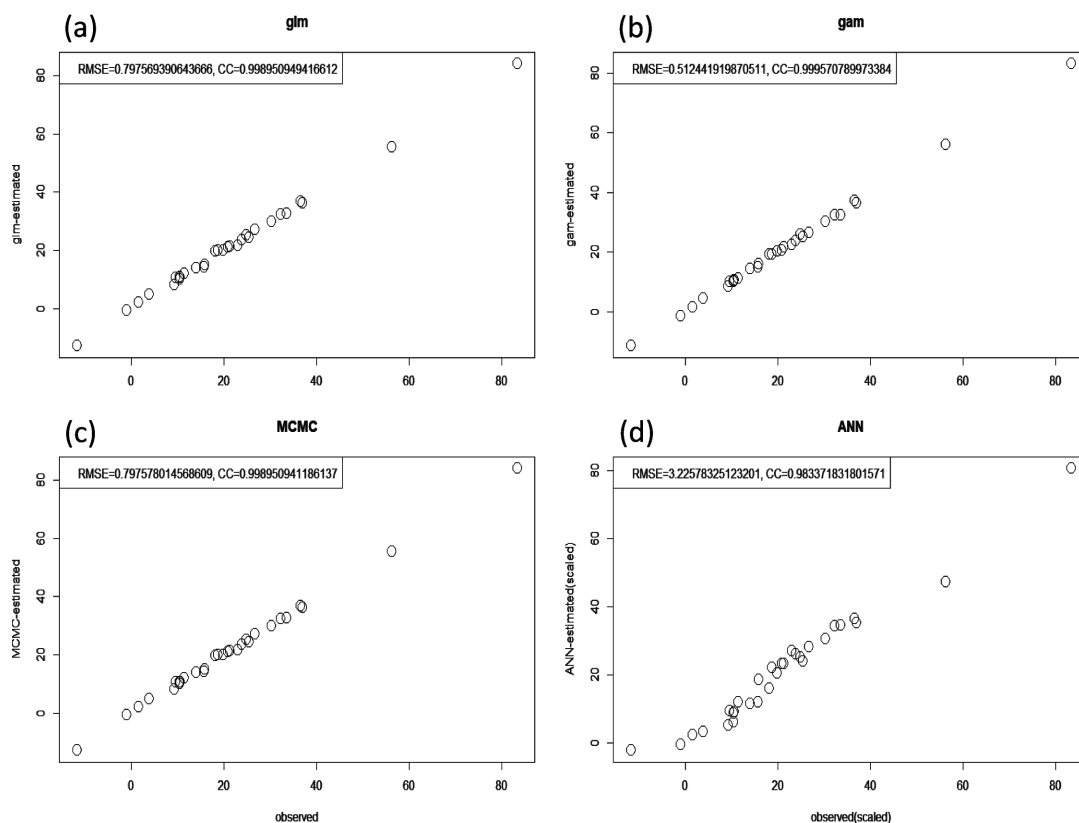


図5 glm,gam,MCMC,ANN で得た順解析モデルの精度比較 (ノイズの程度: 小の場合)

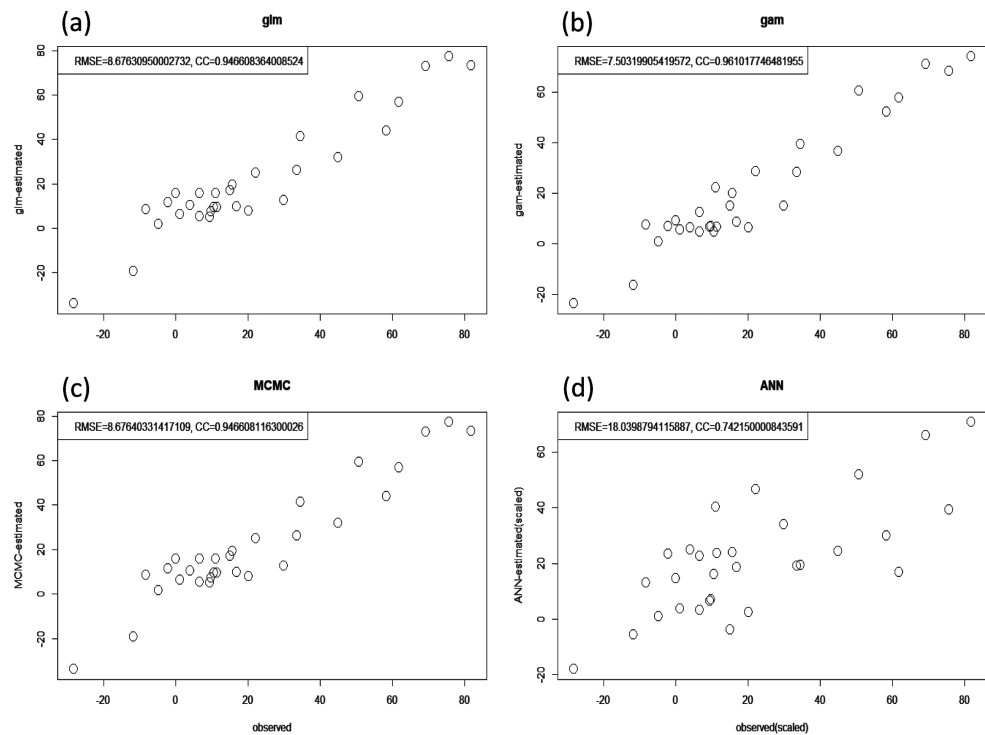


図6 glm,gam,MCMC,ANN で得た順解析モデルの精度比較（ノイズの程度：中の場合）

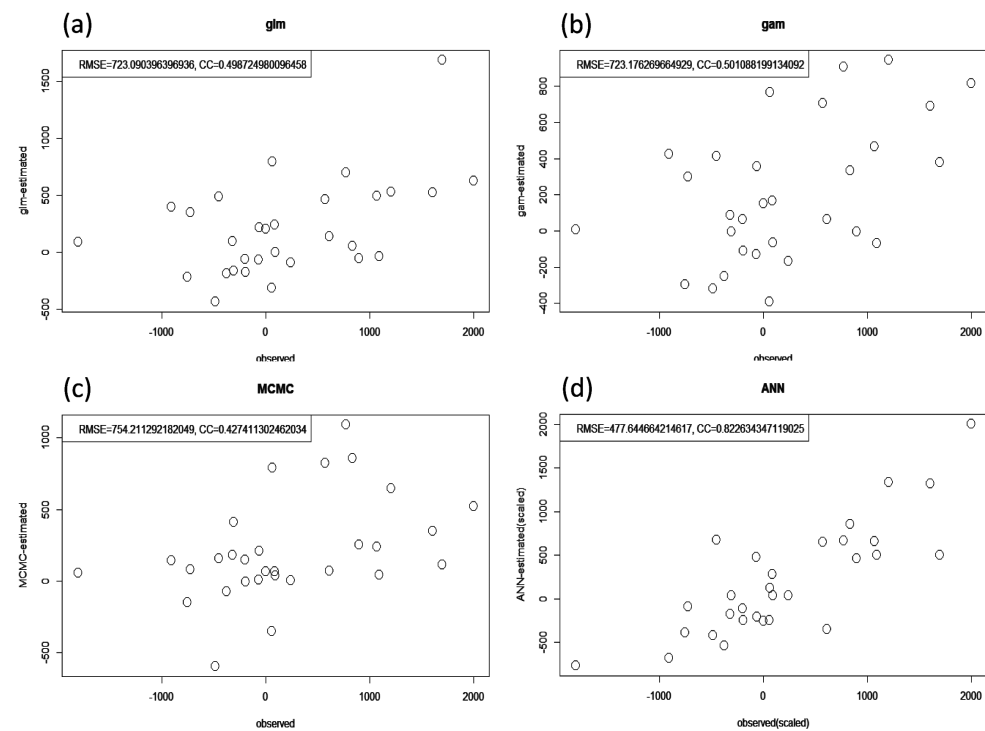


図7 glm, gam, MCMC, ANN で得た順解析モデルの精度比較（ノイズの程度：大の場合）

データがノイズを含む場合の順解析モデルの一つの選択肢として有益と思われる。
本稿で行った各種最尤法、MCMC法、ニューラルネット

ワーク回帰法は、Theory_designerモジュールとしてAI搭載材料情報統合システムの一つであるrMIPHA_studio^{*2}（プログラミング言語：R）に実装済みである。

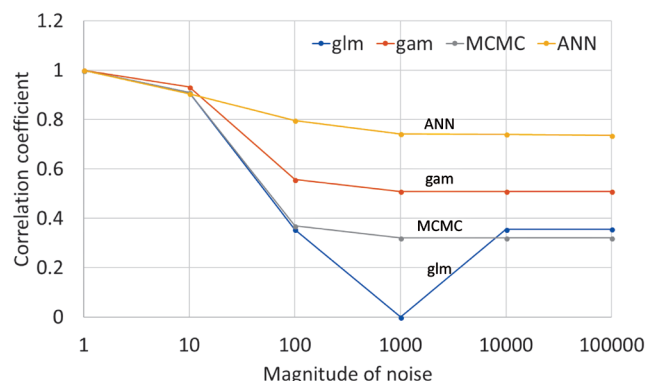


図8 各解析モデルの精度に及ぼすノイズの大きさの影響

4 おわりに

物理モデルにできるだけ従って現象を予測するというのが理想的である。単純系では、最尤法やMCMC事後分布法が物理モデルを表現したモデル式の係数、べき指数の高精度推定に有益であることを説明した。ただし、データが大きな誤差を含むほどその推定精度は顕著に低下する。この場合、物理モデルは全く含まない究極的な近似法であるニューラルネットワーク法が係数推定法よりも推定精度が高くなる場合がある事を示した。

因果関係を重視した物理モデルに従う予測法、相関関係を重視した究極的な近似による推定法ともに重要である。両者は補完的であり、互いの精度の検証のために併用することが解析の信用を増すことにつながるものと思われる。「機械学習はブラックボックスである。」と否定的にとらえる考え方もあるかもしれないが、最後に残された順解析モデル推定法としてその特徴を理解したうえで選択肢として残しておくことが肝要であろう。人工知能が新たなレベルに上がった今日、材料工学も人工知能を取り入れる絶好の機会であり、特に若手研究者、技術者にとって新分野を切り開くことができる魅力ある領域と著者は考えている。人工知能を取り入れることが一層魅力ある材料工学の新たな切り口になることを大いに期待したい。

参考文献

- 1) 加藤雅治：入門転位論，裳華房，(1999)，150.
- 2) C.M.Bishop 著，元田浩，栗田多喜男，樋口知之，松本祐治，村田昇 監訳：パターン認識と機械学習 上，丸善出版，(2012)
- 3) H.Akaike：Proceedings of the 2nd International Symposium on Information Theory, ed. by B.N.Petrov and F.Caski, Akademiai Kiado, Budapest, (1973), 267.
- 4) S.N.Wood：J. Royal Statistical Society, 73B (2011) 1, 3.
- 5) S.N.Wood, N. Pya and B. Saefken：J. American Statistical Association, 111 (2016), 1548.
- 6) S.N.Wood：J. American Statistical Association, 99 (2004), 673.
- 7) S.N.Wood：An Introduction with R (2nd edition) . Chapman and Hall/CRC, (2017)
- 8) A.D.Martin, K.M.Quinn and J.H.Park：J. Statistical Software, 42 (2011) 9, 1.
- 9) W.N.Venables and B.D.Ripley：Modern Applied Statistics with S. Fourth Edition. Springer, New York, (2002)
- 10) J.Mockus：Theory and applications., Kluwer Academic, (2013)
- 11) J.Mockus：Optimization Techniques, (1974), 400.
- 12) J.Mockus：Proc. IFIP Congress, (1977), 195.
- 13) J.Mockus：Bayesian Approach to Global Optimization. Kluwer Academic Publishers, Dordrecht, (1989)
- 14) 足立吉隆：ふえらむ, 23 (2018) 6, 271.

(2018年7月6日受付)

*2: rMIPHA_studioは、rMIPHA¹⁴⁾、Theory_designer、Microstructure Image Editor で構成されスタンドアロン型のAI搭載材料情報統合システム (Material Genome Integration System for Phase and Property Analysis: MIPHA, プログラミング言語: R) である。それぞれ、豊富な機械学習を搭載した順・逆解析、理論式の係数推定、画像の類似性評価を行うことを得意とするモジュールである。この他に、2D3D 画像の特徴量抽出、順解析、逆解析を一貫して行えるスタンドアロン型のMIPHA (プログラミング言語: Visual Basic) がある。