時系列データから、その支配方程式を導出するアルゴリズムの開発 *

稲　毛　真　一 **

Development of an algorithm to evaluate governing equation from time series data

Shin-ichi INAGE**

Abstract

A new algorithm to determine the master equation behind time-series data is proposed. For this algorithm, a library of potential terms for the master equation for a finite difference formula was given. Using time-series data, each time-series potential term was estimated. The master equation was written as a linear sum of each potential term. Additionally, each coefficient in front of the potential terms was determined to minimize Σ(linear sum)^2 during the given period using the genetic algorithm and Gauss– Seidel method. The genetic algorithm minimized both Σ(linear sum)^2 and number of potential terms, because the estimated master equation by the time-series data should be simple. This approach was applied to several functions and so-called Lorenz equations, which demonstrate typical chaotic motion. The time-series data was given through the numerical calculation of sample functions and Lorenz equations. The estimated master equation based on the proposed approach was compared with sample functions and the Lorenz equations. In each case, the coefficients in the equations were in good agreement. There was a small discrepancy between the coefficients of the original Lorenz equations and the estimated master equation. For the Lorenz equations, in the short-term, the attracters were in good agreement. By contrast, the discrepancies influenced the longterm attracters based on chaotic motion.

Key Words : BIG DATA, time-series data, physical equation

1. はじめに

本報告は、与えられた巨大な時系列データ、いわゆるBIG DATA を分析し、その背後に隠された物理現象の支配方程式を見出すアルゴリズムに関するものである。ここで述べる支配方程式はいわゆる近似式ではなく、微分方程式の形式を指している。

データから近似式を作成する手法は数多くの研究がなされており、(1)local approximation、(2)global approximation、(3)mid-range approximations に大別される [(1)-(3)]。特に、(2) のカテゴリーには、ニューラル・ネットワークによる機械学習などの応用も含まれ、昨今の一つの大きな流れとなっている。これらは近似関数やデジタルデータとしての出力を想定したものであり、データの背後に隠された物理法則までを評価するのは難しいと考える。他方、Schmidt 及び Lipson は、新たな試みとして人工知能、AI を活用して時系列データから物理法則を見出す研究をしており、興味深い [(4)]。本報告では、時系列データから直接、微分方程式の形で、理論形式の支配法則を見出す事をターゲットにしている。このような形式で支配方程式を見出すことができれば、その式を構成する各項を物理的に解釈するのは容易と考える。このような手法は、時系列データが複雑なほど有効と考える。本報告では、(1) 基本的なアルゴリズムの構築、(2) 基本的な関数、微分方程式により生成した時系列データによる検証、(3) カオス系の方程式、その一例としてのローレンツ方程式への適用と検証を目的とする。なお、開発者は本提案アルゴリズムを Data to Equation (D2E) アルゴリズムと呼んでいる。

* 令和元年 10 月 31 日受付

** 機械工学科

2. D2E の基本概念

本提案の基本概念を以下に示す。今、ここで、以下の時系列データを考える。

時間 : (t_0, t_1, t_2, t_3, ・・・, t_N)、データ : (f_0, f_1, f_2, f_3, ・・・, f_N).

ここで、時系列データの時間間隔を Δt とすると、各時刻のデータを用いて、データの一次～高次微分の時系列データを計算するのは、以下のように容易である。

一次微分 :

$$\left(\frac{df}{dt}\right)_i = \frac{f_i - f_{i-1}}{\Delta}, \tag{1}$$

二次微分 :

$$\left(\frac{d^2f}{dt^2}\right)_i = \frac{1}{\Delta}\left[\left(\frac{df}{dt}\right)_{i+1} - \left(\frac{df}{dt}\right)_i\right] = \frac{f_{i+1}+f_{i-1}-2f_i}{\Delta^2}, \tag{2}$$

・・・

更に、微分係数に時間 t の関数、データ f の関数を掛ける等の、新なる時系列データを作る事も容易である。本報告では、これらの時系列データを加工して作成する新たな時系列データをライブラリーと呼ぶことにする。与えられた時系列データが何らかの微分方程式系の支配方程式を有するのであれば、それは準備したライブラリー項を用いて、次のように書けるはずである。

$$C_1\left(\frac{df}{dt}\right)_i + C_2 f\left(\frac{df}{dt}\right)_i + C_3 f^2\left(\frac{df}{dt}\right)_i + \cdots\cdots + C_m\left(\frac{d^2f}{dt^2}\right)_i + \cdots\cdots = 0, \tag{3}$$

ここで、C_i は係数であり、現状では実数を想定している。すなわち、現状では複素数は取り扱わないものとする。上記の各項の内で、確実に支配方程式に含まれると思われる項を“参照項”とし、その係数で割れば、以下のように書ける（ここでは、例として一次微分項を参照項とした）。

$$\left(\frac{df}{dt}\right)_i + A_1 f\left(\frac{df}{dt}\right)_i + A_2 f^2\left(\frac{df}{dt}\right)_i + A_3 f^3\left(\frac{df}{dt}\right)_i + \cdots\cdots + A_m\left(\frac{d^2f}{dt^2}\right)_i + \cdots\cdots = 0, \tag{3'}$$

ここに、$A_i = C_i/C_1$ である。上記で想定したライブラリー項の中には、支配方程式として不要なものも含まれているために、実際に時系列データを (3') に当てはめても、左辺はゼロにはならない。すなわち、誤差を生じる。その誤差 e を各時刻に対して、以下のように書く。

$$e_i = \left(\frac{df}{dt}\right)_i + A_1 f\left(\frac{df}{dt}\right)_i + A_2 f^2\left(\frac{df}{dt}\right)_i + A_3 f^3\left(\frac{df}{dt}\right)_i + \cdots + A_m\left(\frac{d^2f}{dt^2}\right)_i + \cdots \tag{4}$$

時系列データを通じて、全体の時間数を合計したトータル誤差は以下で書ける。

$$E = \sum_0^N (e_i)^2 = \sum_0^N \left[\left(\frac{df}{dt}\right)_i + A_1 f\left(\frac{df}{dt}\right)_i + A_2 f^2\left(\frac{df}{dt}\right)_i + A_3 f^3\left(\frac{df}{dt}\right)_i + \cdots + A_m\left(\frac{d^2f}{dt^2}\right)_i + \cdots\right]^2. \tag{5}$$

この全体誤差を最小にする係数 A_1, A_2, ・・・の組合せを見出せれば、それは時系列データの支配方程式を見つけた事になると考える。すなわち、

$$\left(\frac{\partial E}{\partial A_1}\right) = \left(\frac{\partial E}{\partial A_2}\right) = \left(\frac{\partial E}{\partial A_3}\right) \cdots = 0. \tag{6}$$

例えば、A_1 では、

$$\sum_0^N \left(\frac{df}{dt}\right)_i \left[\left(\frac{df}{dt}\right)_i + A_1 f\left(\frac{df}{dt}\right)_i + A_2 f^2\left(\frac{df}{dt}\right)_i + A_3 f^3\left(\frac{df}{dt}\right)_i + \cdots + A_m\left(\frac{d^2f}{dt^2}\right)_i + \cdots\right] = 0 \;\therefore \left[\sum_0^N \left(\frac{df}{dt}\right)_i f\left(\frac{df}{dt}\right)_i\right] A_1 + \left[\sum_0^N \left(\frac{df}{dt}\right)_i f^2\left(\frac{df}{dt}\right)_i\right] A_2 + \cdots = -\sum_0^N \left(\frac{df}{dt}\right)_i \left(\frac{df}{dt}\right)_i. \tag{7}$$

同様に、A_i に関しても同様の計算ができる。式 (7) を始めとする A_i に関する式群は、A_1, A_2,…の一次線形方程式であり、ガウス・ザイデル法などにより、適切に解を求める事ができる。このアルゴリズムにより、時系列データからライブラリー項を生成し、その組合せの線形結合式の係数を見つける事ができる。これは、時系列データを支配する物理方程式に近い事が期待できる。

具体的に、f=sin (2t/π) で検討する。この三角関数で、t=0–2 の間で時系列データを与える。ライブラリーとして、一次微分、二次微分及び f の二次関数を考える。その場合、(4) で定義した誤差は

$$e_i = \left(\frac{d^2f}{dt^2}\right)_i + A_1\left(\frac{df}{dt}\right)_i + A_2 f + A_3 f^2. \tag{8}$$

ここでは、二次関数を参照項としている。式 (5) のトータル誤差を最小にする条件で、係数 A_1–A_3 は以下の連立式で求まる。

$$\begin{bmatrix} \sum\left(\frac{df}{dt}\right)_i^2 & \sum f_i\left(\frac{df}{dt}\right)_i & \sum f_i^2\left(\frac{df}{dt}\right)_i \\ \sum f_i\left(\frac{df}{dt}\right)_i & \sum f_i^2 & \sum f_i^3 \\ \sum f_i^2\left(\frac{df}{dt}\right)_i & \sum f_i^3 & \sum f_i^4 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} = \begin{bmatrix} -\sum\left(\frac{df}{dt}\right)_i\left(\frac{d^2f}{dt^2}\right)_i \\ -\sum f_i\left(\frac{d^2f}{dt^2}\right)_i \\ -\sum f_i^2\left(\frac{d^2f}{dt^2}\right)_i \end{bmatrix}. \tag{9}$$

式 (9) をガウス・ザイデル法で解く。時系列データの間隔 Δt は、Δt=0.01 である。すると、求められた係数

は、表 1 の通りであり、A_1 及び A_3 は≒ 0 であり、A_2 ≒ 0.4052496 を得る。すなわち、満足すべき支配方程式は、以下と考える。

$$\left(\frac{d^2f}{dt^2}\right) = -0.4052496f. \quad (10)$$

この理論的な係数は、$(2/\pi)_2$=0.40528 であり、十分な精度で係数を再現している。トータル誤差そのものも計算でき、5.730635e-17 であり、ほぼゼロとなる事を確認できた。なお、問題によっては、トータル誤差がゼロでなく、ある値に収束することはあり得る。これは、支配方程式に定数項が含まれていると解釈が可能である。

表 1 ：係数の評価結果

A_1	4.091419e-11
A_2	0.405284
A_3	-6.843723e-11

3. 遺伝的アルゴリズムとの組合せ

上述の通り、基本的概念は時系列データからライブラリー項の時系列データを生成し、その線形結合の係数を最小化する組合せを見つけるものである。他方、この手法のみでは､支配方程式として冗長になる可能性がある。すなわち、例えば、微分方程式 (10) は f =sin (2t/π) の支配方程式であるが、式 (10) を更に微分した、

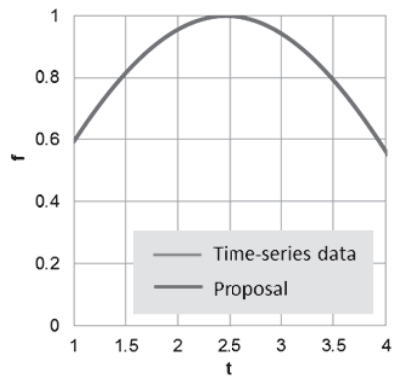
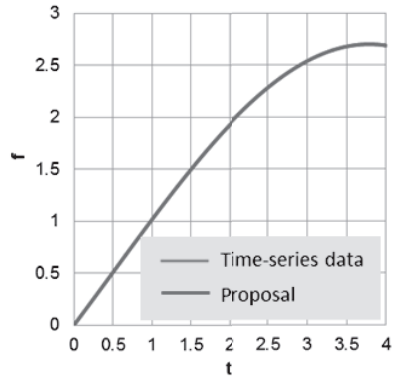
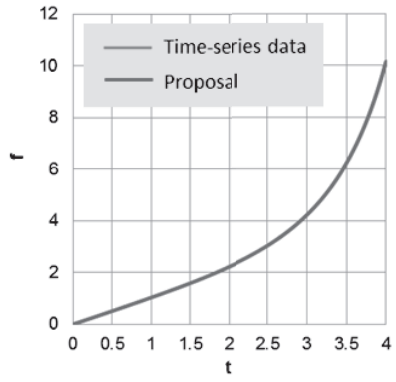
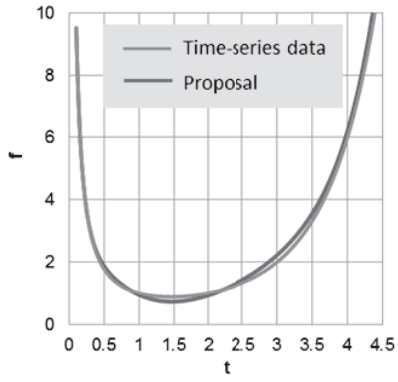
$$\left(\frac{d^3f}{dt^3}\right) = -0.4052496\left(\frac{df}{dt}\right). \quad (11)$$

も支配方程式であり、(10)、(11) の線形和も時系列データを満足する支配方程式である｡そのため､ライブラリー項に、一番低次の支配方程式に含まれる項を更に微分して得られる項が含まれると、D2E では、その線形和を支配方程式と求めてしまい、式をいたずらに冗長にしてしまう。これを避けるために、支配方定期の項数を最小にしながら、誤差を最小化するアルゴリズムとした。その項数最小化を遺伝的アルゴリズムに依った。遺伝的アルゴリズムは多くの提案があるが、本報告では、渡邊らによる Neighborhood Cultivation Genetic Algorithm (NCGA) を適用している [5]。

4. 検証事例：

基本的な関数の事例を対象に、D2E を検証した結果を表 2 に示す。最上段は、上述の f =sin (2t/π) の事例である。二番目、三番目は、二次の線形、非線形微分方程式を数値的に解き、生成した時系列データを用いて、D2E アルゴリズムを適用したものである。計算時間の範囲、タイム･ステップ及び生成した時系列データ数は、左の欄に記載している。いずれも、元の微分方程式と同じ項のみが残り、係数の一致度も良い。図中、青線はオリジナルの時系列データ、赤線は D2E より求められた支配方程式を数値的に解いたものである。一致度は極めて良い。最下段は、時系列データとして、ガンマ関数を用いたものである。ガンマ関数は支配方程式（微分方程式）を持たない事が、既に 19 世紀に数学的に証明されている [6]。それに対して、敢えて D2E を適用した場合である。近似的ながら、表中に示す微分方程式を得ることができる。それを数値的に解くと、グラフの赤線のようになり、青線のオリジナルのガンマ関数をよく近似できていることが分かる。このように、D2E は近似的な関数を生成する上でも有効と考える。ただし、この近似はあくまで、時系列データの時間間隔の中で成立するものであり、それを超えての適用は限界があると考える。

表２：D2Eの検証事例

Function or Differential equation	Theoretical master equation	Estimated master equation	Validation
$f = \sin(\frac{2t}{\pi})$ Period: t=1-4 Δt=1e-3 Data number:3000	$\frac{d^2f}{dt^2} = -\left(\frac{2}{\pi}\right)^2 f$	$\frac{d^2f}{dt^2} = -0.40528f$ +4.7685E-15	
$\frac{d^2f}{dt^2} = 0.1\frac{df}{dt} - 0.2f$ Period: t=0-4 Δt=1e-3 Data number:4000	Same as on the left	$\frac{d^2f}{dt^2} = 0.1\frac{df}{dt} - 0.2f$ -5.59101E-10	
$\frac{d^2f}{dt^2} = 0.1\frac{df}{dt} - 0.2f$ $+0.2f^2$ Period: t=0-4 Δt=1e-3 Data number:4000	Same as on the left	$\frac{d^2f}{dt^2} = 0.1\frac{df}{dt} - 0.2f$ $+0.2f^2$ -5.72618E-09	
f= Γ(t) $= \int_0^\infty x^{t-1}e^{-x}dx$ Period: 0.1-4 Δt=1e-3 Data number:3900	None	$\frac{d^2f}{dt^2}$ $+ (7.61115500799875$ $- 0.11864f)\frac{df}{dt}$ $- 3.876703t$ $- 2.6139007t^2$ $+ 16.12076f$ $- 3.885793f^2$ $= 0.1112640$	

5. カオス系への適用—ローレンツ方程式への応用

カオス的なふるまいをする事例として、以下のローレンツ方程式を考える。これは、関数 X、Y、Z が非線形な形で互いに影響を与え合うものである [7]。

$$\frac{dX}{dt} = -pX + pY \tag{12}$$

$$\frac{dY}{dt} = -XZ + rX - Y \tag{13}$$

$$\frac{dZ}{dt} = XY - bZ, \tag{14}$$

ここで、p, r 及び b は係数である。この式を対象にすると、D2E の形式は、各 X、Y、Z に対して定義され、以下となる。

$$e_{1(i)} = \frac{dX}{dt} + A_1X + A_2Y + \cdots \tag{15}$$

$$e_{2(i)} = \frac{dY}{dt} + B_1XZ + B_2X + B_3Y + \cdots \tag{16}$$

$$e_{3(i)} = \frac{dZ}{dt} + C_1XY + C_2Z + \cdots \tag{17}$$

X、Y、Z それぞれに対するトータル誤差も、同様に、以下で定義できる。

$$\mathrm{E}_1 = \sum_0^N \left(e_{1(i)}\right)^2 \tag{18}$$

$$\mathrm{E}_2 = \sum_0^N \left(e_{2(i)}\right)^2 \tag{19}$$

$$\mathrm{E}_3 = \sum_0^N \left(e_{3(i)}\right)^2. \tag{20}$$

式 (12) － (14) に適切な初期値 X_0=0.1、Y_0=0.15 及び Z_0=0.3) を与えて作成した、X、Y、Z の３つ時系列データを、式 (15) － (17) に対して適用したライブラリー項を表３にまとめる。ライブラリー項として、オリジナルのローレンツ方程式に含まれない複数の項を追加している。それらの項を考慮した上で、D2E を適用した場合の各ライブラリー項の係数も併せて、表３にまとめている。これから、D2E は正しくローレンツ方程式の各係数をトレースできている事が分かる。また、オリジナルに含まれない項の係数はゼロに収束している。

オリジナルローレンツ方程式と D2E で求めた支配方程式を同じ初期値 (X_0=0.1, Y_0=0.15 及び Z_0=0.3) で数値的に解いた時系列のグラフ及びアトラクターをそれぞれ、図１、２で比較する。図 1 で、横軸は時間、縦軸は X、Y、Z の値を示す。点はオリジナルのローレンツ式の結果、線は D2E で求めた支配方程式による解である。図のように X、Y、Z はいずれもカオス的な振舞いをしており、この時系列データのみが与えられた場合、人の手で支配方程式を求めるのは不可能に思われる。D2E で求めた方程式の解はオリジナルの値を良く再現している。他方、ローレンツ方程式は典型的なカオスの事例であり、各項係数や初期条件に敏感であることが知られている。表３の通り、係数にも非常小さいながら差異があるので、このまま計算を継続すれば互いの解はずれてくると思われる。しかし、短時間間隔に限れば D2E カオス的な振舞いに対しても十分適用できる事を確認できたと考える。図２のアトラクターでも、点はオリジナルのローレンツ式、線は D2E の支配方程式による解である。こ時間範囲内ではアトラクターも十分再現できている。

以上から、提案した D2E アルゴリズムは、適切な時間解像度における時系列データがあれば、その背後に隠された支配方程式を求める事が可能である事を検証できたと考える。

表３．ローレンツ方程式と D2E による評価の比較

Time-series data	Considered library terms	Coefficients	
		X_0=0.1, Y_0=0.15, Z_0=0.3	X_0=0.2, Y_0=0.3, Z_0=0.1
X	$\frac{dX}{dt}$	1	1
	X	9.991003	9.991551
	Y	-9.999383	-9.999348
	X^2	0	0
	$X\frac{dX}{dt}$	0	0
	$YZ\frac{dX}{dt}$	1.7142e-06	1.610772e-06
Y	$\frac{dY}{dt}$	1	1
	X	-27.896400	-27.89692593
	Y	0.970553	0.970211
	Y^2	-0.000629888	-0.000583009
	XZ	0.997455	0.997408
	$Y\frac{dY}{dt}$	0	0
	$XZ\frac{dY}{dt}$	-2.3374E-06	-2.49382E-06
Z	$\frac{dZ}{dt}$	1	1
	Z	2.684706	2.684748
	Z^2	-0.000727547	-0.000728772
	XY	-0.999108	-0.999111
	$Z\frac{dZ}{dt}$	-7.58534e-05	-7.59098e-05
	$XY\frac{dZ}{dt}$	3.48495e-06	3.485e-06

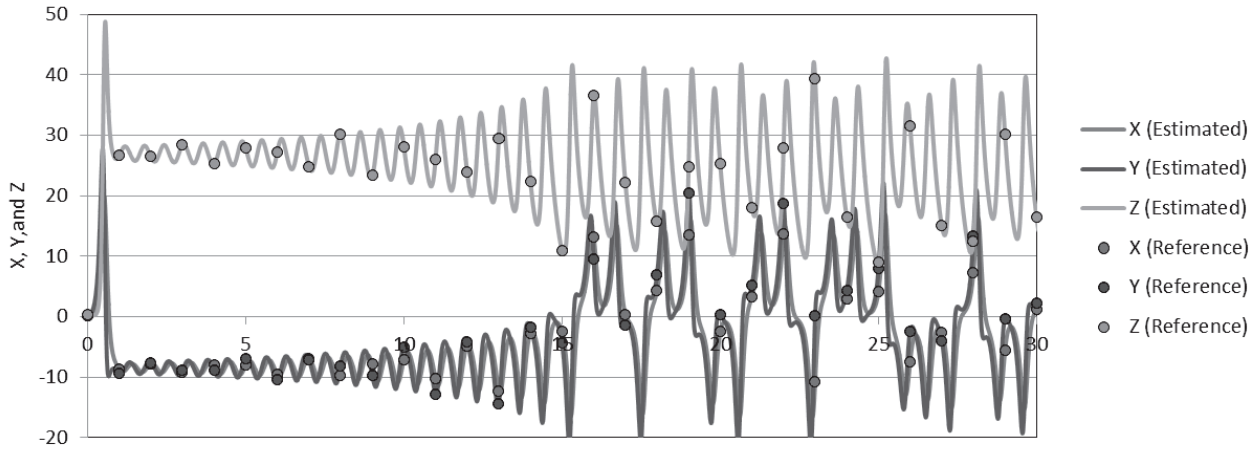
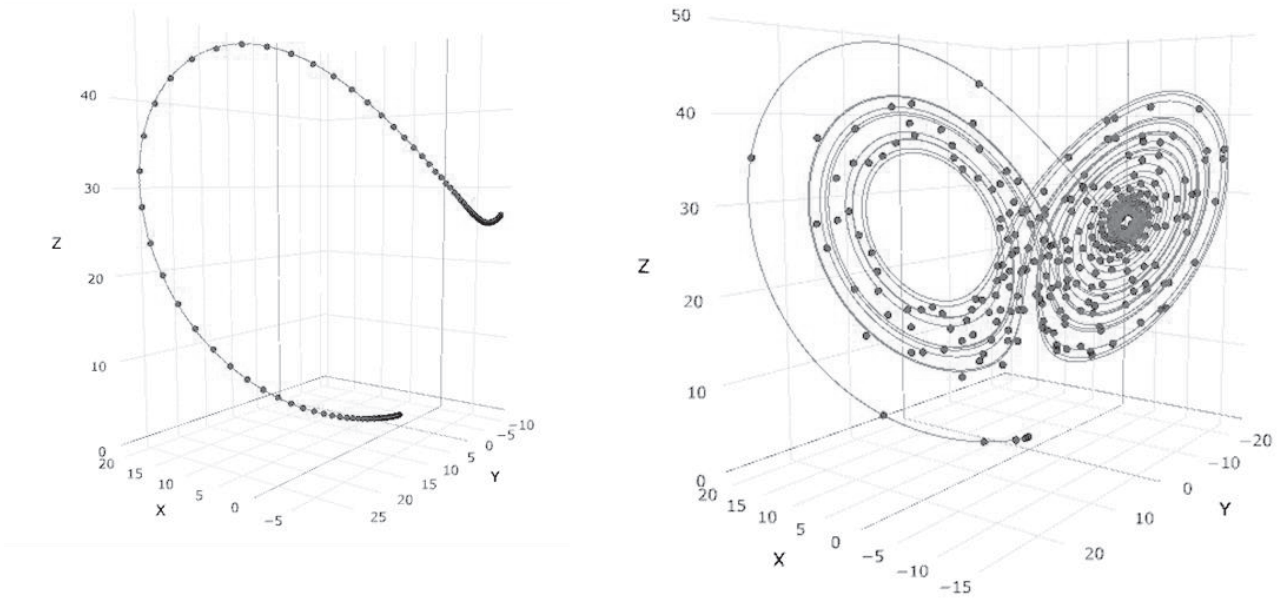


図１：ローレンツ方程式と D2E による支配方程式の解時系列の比較.



1）During t = 0－1　　2）During t = 0－30

図２：ローレンツ方程式と D2E による支配方程式の解のアトラクターの比較

6. おわりに

時系列データから、その背後に隠された支配方程式を理論形式で求めることを目的に、D2E アルゴリズムを開発し、検証した。決定論的な支配方程式による時系列データのみならず、微分方程式形式で記述ができないガンマ関数や、カオス的な振舞いをするローレンツ方程式にも適用可能な事を確認した。今後の応用として、流体複雑現象の解明等に適用したいと考えている。特に乱流では、計算機環境の進化により、ナビエ・ストークス方程式を大規模に直接解く事ができるようになっており、実験結果も含めて膨大なデータも公開もされている。そのような BIG DATA に対して、今回の D2E を適用する事で乱流の素過程を記述できるシンプルな支配方程式を導出できないかを検討している。また、他の応用として、本手法を逆問題の手法と考えれば、電力系統の代表ノードでの電力情報(位相､振幅など)の計測データがあれば、系統のインピーダンスを逆算できる可能性もある｡今後、種々の事例に適用していく予定である。

参考文献

(1) Kurt Hornik , Maxwell Stinchcombe , Halbert White , Multilayer feedforward networks are universal approximators, Neural Networks, Volume 2, Issue 5 1989, pp. 359-366
(2) Martin Langkvist, Lars Karlsson, Amy Loutfi, A review of unsupervised feature learning and deep learning for time-series modeling., Pattern Recognition Letters 42, 2014, pp. 11-24.
(3) Abdulrahman Baqais, Generic Algorithm for function approximation: An experimental investigation, International Journal of Artificial Intelligence and Applications (IJAIA), Vol. 7, No. 3, May 2016.
(4) Michael Schmidt and Hod Lipson, Distilling Free-Form Natural Laws from Experimental Data.” ,Science, Vol. 324, April 3, 2009.
(5) Otto Ludwig Hölder, "Über die Eigenschaft der Gammafunction keiner algebraischen Differentialgleichung zu genügen," Math. Ann., 28, (1887) pp. 1-13. doi:10.1007/BF02430507.
(6) Shinya Watanabe, Tomoyuki Hiroyasu, Mitsunori Miki, NCGA : Neighborhood Cultivation Genetic Algorithm for Multi-Objective Optimization Problems, Late Breaking papers at the Genetic and Evolutionary Computation Conference (GECC-2002), New York, USA, 9-13.
(7) Lorenz, E. N. N.., Deterministic Nonperiodic Flow, Journal of Atmospheric Sciences, Vol.20, pp.130 -141,1963.