Discovery of Differential Equations from Numerical Data

Niijima, Koichi
Department of Informatics Kyushu University

Uchida, Hidemi
Department of Informatics Kyushu University

Hirowatari, Eiju
Department of Informatics Kyushu University

Arikawa, Setsuo
Department of Informatics Kyushu University

https://hdl.handle.net/2324/3007

出版情報：DOI Technical Report. 138, 1997-09-24. Department of Informatics, Kyushu University
バージョン：
権利関係：

DOI-TR-138

DOI Technical Report

Discovery of Differential Equations from Numerical Data

by

K. Niijima, H. Uchida, E. Hirowatari and
S. Arikawa

September 24, 1997

Department of Informatics
Kyushu University
Fukuoka 812-81, Japan

Email: niijima@i.kyushu-u.ac.jp Phone: +81-92-583-7631

Discovery of Differential Equations from Numerical Data

Koichi Niijima
Department of Informatics
Kyushu University
Fukuoka 812-81
Japan
niijima@i.kyushu-u.ac.jp

Hidemi Uchida
Department of Informatics
Kyushu University
Fukuoka 812-81
Japan
hidemi@i.kyushu-u.ac.jp
(At present, NTT Corporation)

Eiju Hirowatari
Department of Informatics
Kyushu University
Fukuoka 812-81
Japan
eiju@i.kyushu-u.ac.jp

Setsuo Arikawa
Department of Informatics
Kyushu University
Fukuoka 812-81
Japan
arikawa@i.kyushu-u.ac.jp

Abstract

This paper proposes a method of discovering some kinds of differential equations with interval coefficients, which characterize or explain numerical data obtained by scientific observations and experiments. Such numerical data inevitably involve some ranges of errors, and hence they are represented by closed intervals in this paper. Based on these intervals, we design some interval inclusions which approximate integral equations equivalent to the differential equations. Interval coefficients of the differential equations are determined by solving the interval inclusions. Many combinations of interval coefficients can be obtained from numerical data. We find out a differential equation whose coefficients consist of intersections of the computed interval coefficients. The refutability of the differential equation is also discussed. Our discovering method is verified by some simulations.

1 Introduction

By rapid progress of observation and experiment systems, it has become possible to get a large number of numerical data. It is very important to discover some rules from these numerical data. Up to now, many scientists have found such rules in the form of mathematical expressions such as differential and integral equations based on their past experience. However, a large number of numerical data make it difficult. Rules contained in such data need to be discovered automatically with the aid of computer. Langley et al.(1981,1983a,1983b) developed the BACON system for discovering conservation laws in scientific fields. In the paper (Falkenhainer & Michalski, 1986), the ABACUS system was invented to realize quantitative and qualitative discovery. Zembowicz & Żytkow (1992) constructed a system for finding equations.

The present paper proposes a method of discovering some kinds of differential equations with interval coefficients, which characterize or explain numerical data obtained by scientific observations and experiments. Such numerical data inevitably involve some ranges of errors, and hence we represent the numerical data by closed intervals. Using these intervals, we design some interval inclusions which approximate integral equations equivalent to the differential equations. Interval coefficients in the differential equations are determined by solving the interval inclusions using Hansen's method. We can obtain many combinations of interval coefficients from numerical data. It is desirable to discover a differential equation satisfying all the numerical data. Thus we construct a differential equation whose coefficients consist of intersections of the computed interval coefficients provided that they are not empty. This is just a differential equation identified from all the numerical data. If at least one of the intersection coefficients is empty, then we show that the differential equations, each of which has a combination of interval coefficients, do not possess common solutions. By virtue of this fact, we can refute all the differential equations in the present searching class and proceed to search a desirable equation in a larger class of differential equations.

Until now, identification of systems by differential equations has been done using numerical data themselves. Based on these numerical data, parameters occurring in the differential equations are determined by applying the least square methods. Such approaches, however, do not yield the concept of refutability of differential equations. Our identification method can refute the differential equations by checking the emptiness of intersections of computed interval coefficients. This fact enables us to search a larger class of differential equations for a target equation.

In Section 2, some notations to be used in this paper are given. Section 3 is devoted to derive interval inclusions from differential equations and to describe Hansen's method. We discuss in Section 4 the refutability of differential equations. In Section 5, we extend the results obtained in Sections 3 and 4 to a system of differential equations. Simulation results are given in Section 6.

2 Preliminary

Numerical data obtained by observation and experiment systems usually contain various kinds of noises such as ones by the systems themselves, noises coming from outside in measurement, and measurement errors. We are often obliged to discover reasonable differential equations from such noisy numerical data. So far, the discovery of differential equations has been carried out using noisy numerical data themselves. In this paper, we represent the noisy numerical data by closed intervals which may contain true values. Let x be a numerical datum, and $[x_l, x_r]$ a closed interval including x. We denote $[x_l, x_r]$ by $[x]$ which is called an interval number. For a vector $v = (v_1, v_2, ..., v_m)$, we define an interval vector $[v]$ by

$$[v] = ([v_1], [v_2], ..., [v_m]),$$

where $[v_i]$ indicates an interval number. Similarly, we can define an interval matrix $[M]$ by

$$[M] = ([M_{ij}]),$$

where the (i, j)-component $[M_{ij}]$ indicates an interval number.

3 Derivation of interval inclusions from differential equations

We consider a single differential equation of the form

$$\frac{dx}{dt} = \sum_{i=0}^{n} a_i\ x^i, \tag{1}$$

where $x = x(t)$. The right hand side is restricted to a polynomial of x and takes a linear form with respect to the parameters a_j. Although modeling by single differential equations is not so useful, a system of such differential equations has been used for modeling in many application fields. Such systems will be treated in Section 5.

The present problem is to identify the parameters $a_0, a_1, ..., a_n$ in the interval form from the interval numbers $[x(t)]$. We rewrite (1) as

$$\sum_{i=0}^{n} x^i\ a_i = \frac{dx}{dt}. \tag{2}$$

We hit on an idea that approximates (2) by a difference inclusion

$$\sum_{i=0}^{n} [x(t_k)]^i\ [a_i] - \frac{[x(t_{k+1})] - [x(t_k)]}{t_{k+1} - t_k} \supseteq 0, \tag{3}$$

where t_k denote observation points. However, the difference quotient of interval numbers
$([x(t_{k+1}] - [x(t_k)])/(t_{k+1} - t_k)$ causes large numerical errors and the errors propagate in the latter interval arithmetic.

To cope with this problem, we derive an integral equation by integrating both side of (2) from t_0 to t_N:

$$\sum_{i=0}^{n} \int_{t_0}^{t_N} x^i(t)dt\ a_i = x(t_N) - x(t_0). \tag{4}$$

Approximating the integral terms using the trapezoidal rule and replacing both side by interval numbers yield an interval inclusion

$$\sum_{i=0}^{n} [X_i(t_0, t_N)][a_i] - ([x(t_N)] - [x(t_0)]) \supseteq 0, \tag{5}$$

where $[X_i(t_0, t_N)]$ denotes

$$[X_i(t_0, t_N)] = \frac{h}{2}([x(t_0)]^i + [x(t_N)]^i) + h \sum_{k=1}^{N-1} [x(t_k)]^i \tag{6}$$

and the observation points t_k are assumed to be equidistant with mesh size h.

From (5) for $N = k, k+1, ..., k+n$, where $k \geq 1$, we obtain a simultaneous interval inclusion

$$[X^k][a^k] - [r^k] \supseteq 0, \tag{7}$$

where $[X^k]$ is an $(n+1) \times (n+1)$ matrix whose components consist of $[X_i(t_0, t_N]$ for $0 \leq i \leq n$ and $k \leq N \leq k+n$, and $[a^k]$ and $[r^k]$ denote $([a_0^k], [a_1^k], ..., [a_n^k])$ and $([x(t_k)] - [x(t_0)], [x(t_{k+1})] - [x(t_0)], ..., [x(t_{k+n})] - [x(t_0)])$, respectively. There are many solutions $[a^k]$ of (7) so that the following inclusion is fulfilled:

$$\{\ a^k = (a_0^k, a_1^k, ..., a_n^k) \mid X^k a^k = r^k, \quad X^k \in [X^k], \ r^k \in [r^k]\ \} \subseteq [a^k]. \tag{8}$$

As a method for obtaining a solution $[a^k]$ of (7) satisfying (8), we know the Gaussian elimination process (Alefeld & Herzberger 1983). This process, however, contains division in the interval arithmetic, which causes numerical error propagations. We notice here that the left hand side of (8) may be rewritten as

$$\{\ a^k = (a_0^k, a_1^k, ..., a_n^k) \mid a^k = (X^k)^{-1} r^k, \quad X^k \in [X^k], \ r^k \in [r^k]\ \}.$$

We shall remember Hansen's method (Hansen, 1965) which is a technique for computing an interval matrix U such that $\{\ (X^k)^{-1} \mid X^k \in [X^k]\ \} \subseteq U$. The merit of this method is to be able to estimate the bound of an interval matrix H defined by

$$U = \{\ (X^k)^{-1} \mid X^k \in [X^k]\ \} + H.$$

Hansen's method enables us to compute $U[r^k]$ satisfying $[a^k] \subseteq U[r^k]$, and to make the width of the interval vector $[X^k][a^k] - [r^k]$ in (7) as small as possible. In simulations to be presented in Section 6, we use this method to compute $U[r^k]$.

4 Refutation of differential equations with interval coefficients

Solving (7) by Hansen's method and denoting the solution by $[a^k] = ([a_0^k], [a_1^k], ..., [a_n^k])$, we get a differential equation

$$\frac{dx}{dt} = \sum_{i=0}^{n} [a_i^k]\ x^i. \tag{9}$$

Let us define the set S_k:

$$S_k = \{\ x = x(t) \mid \frac{dx}{dt} = \sum_{i=0}^{n} a_i^k x^i, \quad t_0 \leq t \leq T, \quad a_i^k \in [a_i^k], \ x(t_0) \in [x(t_0)]\ \}. \tag{10}$$

It can be shown by Peano's theorem (Zeidler, 1993) that the initial value problem appeared in the set S_k has solutions, that is, the set S_k is not empty.

Usually, the number of numerical data are much more than that of parameters contained in the differential equation. Differential equations should be identified from all the numerical data.

We assume that $M+n+1$ numerical data have been obtained. This means that we can consider M combinations of data of the form $(x(t_0), x(t_k), ..., x(t_{k+n}))$. Put

$$[b_i] = \cap_{k=1}^{M} [a_i^k]. \tag{11}$$

If all $[b_i]$, $i = 0, 1, ..., n$, are not empty, we can say that

$$\frac{dx}{dt} = \sum_{i=0}^{n} [b_i] x^i \tag{12}$$

was identified from all the numerical data.

The following theorem holds.

Theorem 1. *Let S_k be defined by (10) and assume that the set $\cap_{k=1}^{M} S_k$ is not empty and does not contain constant solutions. We suppose that all the interval numbers $[b_i]$, $i = 0, 1, ..., n$ defined by (11) are not empty and put*

$$I = \{ \ x = x(t) \mid \ \frac{dx}{dt} = \sum_{i=0}^{n} b_i x^i, \ \ t_0 \le t \le T, \quad b_i \in [b_i], \ x(t_0) \in [x(t_0)] \ \}. \tag{13}$$

Then we have

$$I = \cap_{k=1}^{M} S_k.$$

Proof: We first prove $I \subseteq \cap_{k=1}^{M} S_k$. Choose any $x \in I$. The function $x = x(t)$ satisfies the differential equation

$$\frac{dx}{dt} = \sum_{i=0}^{n} b_i x^i.$$

Since $b_i \in [b_i] = \cap_{k=1}^{M} [a_i^k]$, it follows that $b_i \in [a_i^k]$ for all k. This implies that x belongs to S_k for all k, that is, $x \in \cap_{k=1}^{M} S_k$ which proves $I \subseteq \cap_{k=1}^{M} S_k$.

We next show that $\cap_{k=1}^{M} S_k \subseteq I$. Choose any $x \in \cap_{k=1}^{M} S_k$. Then x belongs to S_k for all k. Therefore, this x satisfies

$$\frac{dx}{dt} = \sum_{i=0}^{n} a_i^k x^i, \qquad a_i^k \in [a_i^k] \tag{14}$$

for all k. It is easily shown that $a_i^k = a_i^\ell$ holds for any $k \neq \ell$ and for all i. Indeed, if $a_i^k \neq a_i^\ell$ for some pair $k \neq \ell$ and for some i, then we have from (14),

$$\sum_{i=0}^{n} (a_i^k - a_i^\ell) x^i = 0.$$

This implies that $x = x(t)$ is a constant solution, which contradicts the assumption.

We put $a_i = a_i^k$. Then the differential equation satisfied by x can be written as

$$\frac{dx}{dt} = \sum_{i=0}^{n} a_i x^i, \qquad a_i \in [a_i^k]$$

for all k. Therefore, we have $a_i \in [b_i] = \cap_{k=1}^{M} [a_i^k]$. This means $x \in I$, which finishes the proof.

This theorem justifies the construction of differential equations by taking intersections of interval coefficients. It is valuable to note that these intersection interval numbers decrease monotonically.

In the next, we consider the case that some of $[b_i]$ defined by (11) are empty. Then we can prove the following theorem.

Theorem 2. *Suppose that some of $[b_i]$ defined by (11) are empty. We assume that $\cap_{k=1}^{M} S_k$ does not contain constant solutions. Then $\cap_{k=1}^{M} S_k$ is empty.*

Proof: Assume that $\cap_{k=1}^{M} S_k$ is not empty. Then there exists $x \in \cap_{k=1}^{M} S_k$. Therefore, this $x = x(t)$ satisfies the differential equations

$$\frac{dx}{dt} = \sum_{i=0}^{n} a_i^k x^i, \quad a_i^k \in [a_i^k], \qquad k = 1, 2, ..., M$$

from which we get

$$\sum_{i=0}^{n} (a_i^k - a_i^\ell) x^i = 0. \tag{15}$$

By the first assumption of Theorem 2, we have $[b_{i_0}] = \phi$ for some i_0, where the symbol ϕ denotes the empty set, and hence there exists some pair (k, ℓ) such that $a_{i_0}^k \neq a_{i_0}^\ell$. This fact shows that (15) has only constant solutions, which contradicts the second assumption of Theorem 2.

Theorem 2 shows that when some $[b_i]$ defined by (11) are empty, there are no differential equations identifiable from all the numerical data. This suggests that when at least one of $[b_i]$ defined by (11) is empty, we can refute the set of solutions $\cap_{k=1}^{M} S_k$ and proceed to search differential equations in a new set $\cap_{k=1}^{M} S_k$ obtained by replacing n by $n+1$ in the definition of S_k.

The discovery algorithm for differential equations is as follows:

(i) Put $n = 0$ and $m = 2$.

(ii) Check whether the intervals $\cap_{k=1}^{m} [a_i^k]$ for $0 \leq i \leq n$ are empty or not.

(iii) If all the intervals are not empty and $m < M$, then go to 2 after replacing m by $m+1$. If $m = M$, then stop.

(iv) If the interval $\cap_{k=1}^{m} [a_i^k]$ for some i is empty, then refute all the differential equations appearing in the set $\cup_{k=1}^{m} S_k$. Next, replace n by $n+1$, reset $m = 2$ and go to 2.

5 Extension to a system of differential equations

The results obtained in Section 4 are easily extended to a system of differential equations. We consider a system of differential equations

$$\frac{dx_\ell}{dt} = \sum_{0\le|i|\le n_\ell} a_{i,\ell}\, x^i, \qquad \ell = 1, 2, ..., m, \tag{16}$$

where x^i indicates $x_1^{i_1} x_2^{i_2} \dots x_m^{i_m}$ with $i = (i_1, i_2, ..., i_m)$ and $|i| = \sum_{j=1}^m i_j$. In a similar way as in Section 3, we get an interval inclusion:

$$\sum_{0\le|i|\le n_\ell} [X_i(t_0, t_N)][a_{i,\ell}] - ([x_\ell(t_N)] - [x_\ell(t_0)]) \supseteq 0, \tag{17}$$

where $[X_i(t_0, t_N)]$ has the same form as (6), but now $[x(t_k)]^i = [x_1(t_k)]^{i_1}[x_2(t_k)]^{i_2} \dots [x_m(t_k)]^{i_m}$.

Let d_ℓ be the number of i satisfying $0 \le |i| \le n_\ell$, and let $d = \sum_{\ell=1}^m d_\ell$. For each ℓ, we consider the inclusion (17) for $k \le N \le k + d_\ell - 1$. Moreover, we denote a $d \times d$ matrix $([X_i(t_0, t_N)])_{\,0\le|i|\le n_\ell,\; k\le N\le k+d_\ell-1;\; 1\le\ell\le m}$ again by $[X^k]$, and a vector $([a_{i,\ell}])_{\,0\le|i|\le n_\ell;\; 1\le\ell\le m}$ again by $[a^k]$. By $[r^k]$ we also denote a vector $([x_\ell(t_N)] - [x_\ell(t_0)])_{\,k\le N\le k+d_\ell-1;\; 1\le\ell\le m}$. Then we get a simultaneous interval inclusion

$$[x^k][a^k] - [r^k] \supseteq 0$$

which has the same form as (7). Therefore, we can proceed the latter discussion in the same way as in Sections 3 and 4.

6 Simulations

We carry out some simulations using the discovering system developed previously. Our system works following the discovering algorithm given in Section 4. As the degree of polynomials occurring in the differential equations becomes high, the size of simultaneous inclusions for determining interval coefficients becomes large. This causes the extent of intervals to be found in the interval arithmetic. So we do not calculate all the interval coefficients simultaneously, but solve subsystems of inclusions with changing the combination of the interval terms. In our simulations, numerical data are made artificially from the solutions of differential equations to be identified.

Example 1 (single differential equation).

Let us consider $x(t) = e^t$ in the interval $0 \le t \le 10$. We divide this interval into 100 equidistant subintervals and denote the mesh points by $t_k = 0.1k$. Numerical data are given as $x(t_k) = e^{t_k}$ and interval numbers are constructed by adding 1% error to each of the numerical data. Based on these interval numbers, our system first tries to identify the differential equation

$$\frac{dx}{dt} = [a_0].$$

However, this equation is refuted by Theorem 2 because of $[a_0^1] \cap [a_0^2] = \phi$. Next, our system is going to find out the differential equation

$$\frac{dx}{dt} = [a_1]x.$$

In this case, we can obtain the intersection

$$\cap_{k=1}^{99}[a_0^k] = [0.991371, 1.00265].$$

Therefore, the differential equation

$$\frac{dx}{dt} = [0.991371, 1.00265]x$$

was discovered from the given numerical data. From this equation, we can guess the differential equation $dx/dt = x$ which is satisfied by the given function $x(t) = e^t$.

Example 2 (single differential equation).

Next, we consider the logistic curve $x(t) = (1 + 10e^t)^{-1}$ in the interval $0 \leq t \leq 10$. The division of this interval and the mesh points are the same as in Example 1. Numerical data are given as $x(t_k) = (1+10e^{t_k})^{-1}$ and interval numbers are constructed by adding 1% error to each of the numerical data. The searching of a target equation was tried by changing the combination of the terms in the differential equations. As a result, we succeeded to find out the differential equation:

$$\frac{dx}{dt} = [-1.5006, -0.962858]x + [0.930201, 1.06570]x^2.$$

This equation contains the differential equation

$$\frac{dx}{dt} = -x + x^2$$

which is satisfied by the given logistic curve.

Example 3 (system of differential equations).

Finally, we consider two functions $x(t) = -2e^{-t}$ and $y(t) = e^{-t}(\sin t + \cos t)$ in the interval $0 \leq t \leq 10$. In the same way in Example 1, we divide this interval and denote the mesh points by t_k. Numerical data are given as $x(t_k) = e^{t_k}$ and $y(t_k) = e^{-t_k}(\sin t_k + \cos t_k)$. We construct interval numbers by adding 1% error to each of the numerical data as in Example 1. By changing the combination of the terms in the differential equations, our system succeeded to find out finally the system of differential equations

$$\begin{cases} \dfrac{dx}{dt} = [-2.03367, -1.96269]x + [-2.00822, -1.96263]y, \\[2ex] \dfrac{dy}{dt} = [0.988213, 1.00629]x. \end{cases}$$

This system contains a target system

$$\begin{cases} \dfrac{dx}{dt} = -2x - 2y, \\ \\ \dfrac{dy}{dt} = x. \end{cases}$$

7 Conclusion

In this paper, we proposed a method of discovering some differential equations from given numerical data. The feature of our approach lies in constructing interval inclusions from an integral equation equivalent to a differential equation. By solving these interval inclusions, we can obtain various combinations of coefficients in the differential equation in an interval form. If the intersections of these interval coefficients are not empty, we can identify a differential equation having these intersection coefficients. The justification of this was given in Theorem 1. If not the case, we refute all the differential equations in the present searching class and try to search a target equation in a larger class of differential equations. This refutation was justified by Theorem 2. Since our method contains interval arithmetic, the accuracy of interval coefficients goes down in proportion to the size of interval inclusions. In the simulation, therefore, the size of interval inclusions was reduced by changing the combination of terms in the differential equation.

There are some problems to be solved in the future. This paper restricts a searching class to a class of differential equations whose right hand side has a polynomial form. It is a future work to extend the polynomial form to a more general mathematical expression. One more problem is to clarify a relation between the solutions of a discovered differential equation and base interval numbers constructed from the given numerical data. This is also a future work.

References

Alefeld, G. & Herzberger, J. 1983. *Introduction to Interval Computations.* Academic Press, London.

Falkenhainer, B. C. & Michalski, R. S. 1986. Integrating quantitative and qualitative discovery: the ABACUS system. *Machine Learning*, 1:367–401.

Hansen, E. 1965. Interval arithmetic in matrix computations. part I. *SIAM J. Numerical Analysis.*, 2:308–320.

Langley, P. G., Bradshow, L. & Simon, H. A. 1981. BACON:5 The Discovery of Conservation Laws. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pages 121–126.

Langley, P. G., Bradshow, L. & Simon, H. A. 1983. Rediscovering Chemistry with the Bacon System. In name, editor,*Machine Learning: An Artificial Intelligence Approach*, publisher, place.

Langley, P., Zytkow, J., Bradshaw, G.L. & Simon, H. A. 1983. Mechanisms for Qualitative and Quantitative Discovery. In *Proceedings of the International Machine Learning Workshop*, pages 12–32.

Zeidler, E. 1993. *Nonlinear Functional Analysis and its Applications I*, Springer-Verlag, New York.

Zembowicz, R. & Żytkow, J. M. 1992. Discovery of equations: experimental evaluation of convergence. In *Proceedings Tenth National Conference on Artificial Intelligence*, pages 70–75.