Discovering Time Differential Law Equations Containing Hidden State Variables and Chaotic Dynamics[*]

Takashi Washio, Fuminori Adachi and **Hiroshi Motoda**
Institute of Scientific and Industrial Research, Osaka University,
8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan.
washio@ar.sankan.osaka-u.ac.jp

Abstract

This paper proposes a novel approach to discover simultaneous time differential law equations having high plausibility to represent first principles underlying objective processes. The approach has the power to identify law equations containing hidden state variables and/or representing chaotic dynamics without using any detailed domain knowledge.

1 Introduction

A set of well known pioneering approaches of scientific law equation discovery is called BACON family [Langley *et al.*, 1987]. They try to figure out a static equation on multiple quantities over a wide state range under a given laboratory experiment. Some approaches introduced unit dimension constraints and "*scale-type constraints*" to limit the search space to mathematically admissible equations reflecting the first principles [Falkenhainer and Michalski, 1986],[Washio and Motoda, 1997]. Especially, the scale-type constraints have wider applicability since it does not require any unit information of quantities. Subsequently, LAGRANGE addressed the discovery of "*simultaneous time differential law equations*" reflecting the dynamics of objective processes under "*passive observations*" where none of quantities are experimentally controllable [Dzeroski and Todorovski, 1995]. Its extended version called LAGRAMGE introduced domain knowledge of the objective process to limit the search space within plausible law equations [Todorovski and Dzeroski, 1997]. Extended IPM having similar functions with LAGRAMGE further identified plausible law equations containing "*hidden state variables*" when the variables are known in the domain knowledge [Langley *et al.*, 2003]. PRET identified "*chaotic dynamics*" under similar conditions with these approaches where rich domain knowledge is available [Bradley *et al.*, 1998]. However, scientists and engineers can develop good models of the objective dynamics without using the discovery approaches in many practical cases when the detailed domain knowledge is available. Accordingly, the main applications of the discovery approaches are to identify simultaneous time differential equations reflecting the first principles under passive observation and "*little domain knowledge*."

In this paper, we propose a novel approach called SCALETRACK (SCALE-types and state TRACKing based discovery system) to discover a model of an objective process under the following requirements.

(1) The model is simultaneous time differential equations representing the dynamics of an objective process.

(2) The model is not an approximation but a plausible candidate to represent the underlying first principles.

(3) The model is discovered from passively observed data without using domain knowledge specific to the objective process.

(4) The model can include hidden state variables.

(5) The model can represent chaotic dynamics.

2 Outline

2.1 Basic Problem Setting

We adopt the following "*state space model*" of objective dynamics and measurements without loss of generality.

$$\begin{aligned}\dot{\boldsymbol{x}}(t) &= \boldsymbol{f}(\boldsymbol{x}(t)) + \boldsymbol{v}(t) \quad (\boldsymbol{v}(t) \sim N(0, \boldsymbol{\Sigma}_v)), \text{ and} \\ \boldsymbol{y}(t) &= \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{w}(t) \quad (\boldsymbol{w}(t) \sim N(0, \boldsymbol{\Sigma}_w)),\end{aligned}$$

where the first equation is called a "*state equation*" and the second a "*measurement equation*." $\boldsymbol{x}$ is called a state vector, $\boldsymbol{f}(\boldsymbol{x})$ a system function, $\boldsymbol{v}$ a process noise vector, $\boldsymbol{y}$ a measurement vector, $\boldsymbol{C}$ a measurement matrix, $\boldsymbol{w}$ a measurement noise and t a time index. $\boldsymbol{f}(\boldsymbol{x})$ is not limited to linear formulae in general. $\boldsymbol{C}$ is represented by a linear transformation matrix, since the measurement facilities are artificial and linear in most cases. If $\boldsymbol{C}$ is column full rank, the values of all state variables are estimated by solving the measurement equation with $\boldsymbol{x}$. Otherwise, some state variables are not estimated within the measurement equation, and these variables are called "*hidden state variables*." In the scientific law equation discovery, $\boldsymbol{f}(\boldsymbol{x})$ is initially unknown, and even $\boldsymbol{x}$ is not known correctly. Only a state subvector $\boldsymbol{x}'(\subseteq \boldsymbol{x})$ and a submatrix $\boldsymbol{C}'(\subseteq \boldsymbol{C})$ are initially known. To derive $\boldsymbol{C}$ from $\boldsymbol{C}'$, SCALETRACK must identify the dimension of $\boldsymbol{x}$ at first. Then, it searches plausible candidates of $\boldsymbol{f}(\boldsymbol{x})$ from the measurement time series data.

[*] The support of JSPS Grant-in-aid for Scientifc Research, No.16300045 is acknowledged.

2.2 Entire Approach

The entire approach of SCALETRACK is outlined in Figure 1. Given a set of measurement time series data, the dimension of $\boldsymbol{x}$ is identified through a statistical analysis called "*correlation dimension analysis*" [Berge *et al.*, 1984]. For each element of $\boldsymbol{y}$, its time trajectory is mapped to a phase space constructed by time lagged values of the element, and the degree of freedom, *i.e.*, the dimension of $\boldsymbol{x}$, embedded in the time trajectory is estimated by computing the sparseness of the trajectory in the space.

Once the dimension is known, all possible combinations of scale-types of the elements in $\boldsymbol{x}$ are enumerated based on scale-type constraints, the known measurement submatrix $\boldsymbol{C'}$ and the scale-types of the elements in $\boldsymbol{y}$. The representative scale-types of quantities are ratio scale and interval scale. The examples of the ratio scale quantities are physical mass and absolute temperature where each has an absolute origin, while the examples of the interval scale quantities are temperature in Celsius and sound pitch where their origins are not absolute and arbitrary changed by human's definitions. Due to these natures, the quantitative relations among the quantities are strongly constrained [Luce, 1959], and these constraints are used to determine the scale types of the elements in $\boldsymbol{x}$ from $\boldsymbol{y}$. After every combination of the scale types in $\boldsymbol{x}$ is derived, the candidate formulae of a state equation are generated for each combination based on "Extended Product Theorem" [Washio and Motoda, 1997] limiting the admissible formulae of the equation based on the scale-type constraints.

Subsequently, through a set of state tracking simulations called "SIS/RMC filter" on the given measurement time series data, the parameter values and the states in every candidate state equation are estimated [Doucet *et al.*, 2000]. This state tracking has many advantages comparing with the other nonlinear state tracking approaches such as the conventional Extended Kalman Filter [Haykin, 2001] and the qualitative reasoning based PRET [Bradley *et al.*, 1998]. The former using the linearization of the state equations does not work well when the equations include some singular points and/or some state regions having strong sensitivity to the tracking error. The latter faces a combinatorial explosion of qualitative states when the dimension and/or the complexity of the state space structure are high. In contrast, SIS/RMC filter does not require any approximation to be spoiled by the singularity and the strong nonlinearity, and does not face the combinatorial explosion of the states to be considered, because it tracks the state probability distributions by using its direct and sequential Monte Carlo integration within Bayesian framework. In our approach, the estimated parameter values are rounded off to integers when the values are close enough to the integers within the expected estimation errors, since the parameters tend to be integers in many physical processes. Finally, some state equations providing highly accurate tracking in terms of "*Mean Square Error* (MSE)" are selected as the plausible candidates of first principle based and dynamic state space models of the objective process.

2.3 Implementation

The evaluation of candidate state equations by the SIS/RMC filter is the most time consuming step. Any search can not be skipped, since the search space is nonmonotonic. We experienced that one run of stand alone SCALETRACK to discover a simple state equation took more than a month even if we used an efficient algorithm. Accordingly, the current SCALETRACK introduced a simple grid computing framework using a PC cluster consisting of a control server and 10 personal computers where each has an Athlon XP 1900+ (1.6 GHz) CPU and 2GB RAM. The server computes the first three steps and then allocates the task to evaluate 10% of candidate state equations to each computer. Because this task is mutually independent, and occupies the most of computation in SCALETRACK, this implementation accelerates the run speed almost 10 times.

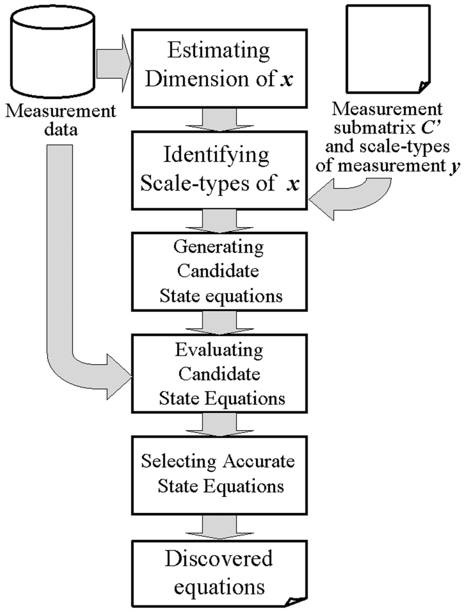
Measurement data
Estimating Dimension of $\boldsymbol{x}$
Measurement submatrix C' and scale-types of measurement $\boldsymbol{y}$
Identifying Scale-types of $\boldsymbol{x}$
Generating Candidate State equations
Evaluating Candidate State Equations
Selecting Accurate State Equations
Discovered equations

Figure 1: Outline of Entire Approach.

Table 1: Basic Performance.

case	ν	ct (h)	$\sigma_w(\%)$				
	$(\sigma_w = 1.0\%)$		0.1	0.5	1.0	2.0	5.0
RR	2.21	1.5	+	±	±	±	−
RRH	2.21	5.5	±	±	−	−	−
RI	2.19	4.0	+	±	±	±	−
RIH	2.19	5.5	+	±	−	−	−

3 Result

3.1 Basic Performance Evaluation

The evaluation is made in terms of scale-types of state variables, hidden state variables and measurement noise levels by using the following two dimensional artificial formulae.

$$\left.\begin{array}{rcl} \dot{x}_1(t) & = & x_1(t)x_2(t) \\ \dot{x}_2(t) & = & -0.5x_2(t) \end{array}\right\} RR,$$

where $y_1 = x_1$ and $y_2 = x_2$ are ratio scale.

$$\left.\begin{array}{rcl} \dot{x}_1(t) & = & 0.4x_1(t)(x_2(t) + 0.2) \\ \dot{x}_2(t) & = & -0.1(x_2(t) + 0.6) \end{array}\right\} RI,$$

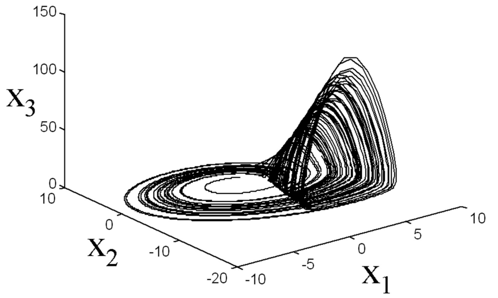

Figure 2: An Attractor of Altered Rossler Chaos.

where $y_1 = x_1$ is ratio scale and $y_2 = x_2$ interval scale. Table 1 shows the evaluation result. All state variables are observed in RR and RI. Whereas, the measurement variable y_2 is not available in RRH and RIH respectively, and hence a hidden state variable exists in these cases. The correlation dimension analysis properly estimated the dimension ν of state vectors as nearly 2 in each case. The computation times ct required for RRH, RI and RIH were longer than that of RR, because the variety of admissible formulae containing interval scale variables is larger than that of ratio scale variables. The result in that the formula having the correct shape is top ranked by the accuracy is marked by $+$. If the correct formula is derived within the top five solutions, it is marked by $\pm$, otherwise it is marked by $-$. The table shows that almost $\sigma_w = 2.0\%$ relative noise is acceptable for no hidden state cases, while noise less than 1.0% is required for hidden state cases. Since $0.5-2.0\%$ noise is the most widely seen in many applications, the performance of SCALETRACK is practical for no hidden state cases and some hidden state cases.

3.2 Discovery of Chaos

The state equation to be discovered is the following Altered Rossler Chaos equation.

$$\dot{x}_1 = -x_2 - x_3, \quad \dot{x}_2 = x_1 + 0.36x_2, \text{ and}$$
$$\dot{x}_3 = 0.01(x_1 - 4.5)(x_1 + 1000x_3 - 4.5).$$

This has an attractor in a (x_1, x_2, x_3)-phase space as depicted in Figure 2. All state variables are interval scale, and can be measured through the corresponding interval scale measurement variables respectively. $\nu = 3.33$ was obtained in the correlation dimension analysis, and hence the state equations consisting of three state variables were searched. The required computation time was 15.0 hours, and the following most accurate state equation was resulted. This formula has an identical shape with the original except some discrepancies of coefficients. This indicates the high ability of SCALETRACK to discover the Chaotic dynamics reflecting the underlying first principles.

$$\dot{x}_1 = -x_2 - x_3, \quad \dot{x}_2 = x_1 + 0.33x_2, \text{ and}$$
$$\dot{x}_3 = 0.064(x_1 - 6.34)(x_1 + 1002x_3 - 4.75).$$

4 Conclusion

SCALETRACK achieved three advantages which have not been addressed in any past work of mathematics, physics and engineering not limited to scientific discovery. The first is the discovery of simultaneous time differential equations having plausibility to represent first principles. The second is the discovery of hidden state variables. The third is the discovery of chaotic dynamics. These discoveries are done without using detailed domain knowledge. These advantages are essentially important in many scientific and engineering fields due to the wide existence of such dynamics in nature.

References

[Berge *et al.*, 1984] Pierre Berge, Yves Pomeau, and Christian Vidal. *Order in Chaos - For understanding turbulent flow*. Hermann, Paris, France, 1984.

[Bradley *et al.*, 1998] Elizabeth A. Bradley, Agnes A. O'Gallagher, and Janet E. Rogers. Global solutions for nonlinear systems using qualitative reasoning. *Annals of Mathematics and Artificial Intelligence*, 23:211–228, 1998.

[Doucet *et al.*, 2000] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10:197–208, 2000.

[Dzeroski and Todorovski, 1995] Saso Dzeroski and Ljupco Todorovski. Discovering dynamics: from inductive logic programing to machine discovery. *Journal of Intelligent Information Systems*, 4:89–108, 1995.

[Falkenhainer and Michalski, 1986] Brian C. Falkenhainer and Ryszard S. Michalski. Integrating quantitative and qualitative discovery: The abacus system. *Machine Learning*, 1:367–401, 1986.

[Haykin, 2001] Simon S. Haykin. *Kalman Filtering and Neural Networks*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2001.

[Langley *et al.*, 1987] Pat W. Langley, Herbert A. Simon, Gary L. Bradshaw, and Jan M. Zytkow. *Scientific Discovery; Computational Explorations of the Creative Process*. MIT Press, Cambridge, Massachusetts, 1987.

[Langley *et al.*, 2003] Pat Langley, Dileep George, Stephen Bay, and Kazumi Saito. Robust induction of process models from time-series data. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 432–439, Menlo Park, California, August 2003. The AAAI Press.

[Luce, 1959] Duncan R. Luce. On the possible psychological laws. *Psychological Review*, 66(2):81–95, 1959.

[Todorovski and Dzeroski, 1997] Ljupco Todorovski and Saso Dzeroski. Declarative bias in equation discovery. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 376–384, San Mateo, California, July 1997. Morgan Kaufmann.

[Washio and Motoda, 1997] Takashi Washio and Hiroshi Motoda. Discovering admissible models of complex systems based on scale-types and identity constraints. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 810–817, Nagoya, Japan, August 1997.