Scientific Discovery of Dynamic Models Based on Scale–type Constraints

Fuminori Adachi*, Takashi Washio* and Hiroshi Motoda*

* *I.S.I.R., Osaka University*

This paper proposes a novel approach to discover dynamic models represented by simultaneous time differential law equations including hidden states from time series data measured in an objective process. This task has not been addressed in the past work though it is essentially important in scientific discovery since any behaviors of the objective processes emerge in time evolution. The promising performance of the proposed approach is demonstrated through the analysis of synthetic data.

1. Introduction

A set of well known pioneering approaches of scientific law equation discovery is called BACON family[1)∼4)]. They try to figure out a static equation on multiple quantities over a wide state range under a given laboratory experiment. More recent systems introduced unit dimension constraints and "*scale–type constraints*" to limit the search space to mathematically admissible equations reflecting the first principles[3),5)]. 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 an objective processes under "*passive observations*" where none of quantities are experimentally controllable[6)]. However, the discovery of "*hidden state variables*" in the objective processes has never been addressed in past work.

In this paper, we propose a novel approach named SCALETRACK (Scale–type and state TRACKing based discovery system) to discover a model of an objective process having the following features.

(1) The model is a simultaneous time differential equations representing the dynamic behavior of an objective process.
(2) The model is not an asymptotic approximated model but a model representing the first principles governing the objective process.
(3) The model can discover hidden state variables and their governing differential equations.
(4) The model is discovered without using background domain knowledge specific to the objective process.
(5) The model is discovered from passively observed data.

In the rest of this paper, the basic problem setting and the entire approach of SCALETRACK are outlined in Section 2, and the performance evaluations are shown in Section 3.

2. Outline

2.1 Basic Problem Setting

We adopt the following "*state space expression*" to model an objective processes and measurements without loss of generality.

$$\left\{\begin{array}{lcl}\dot{\boldsymbol{x}}(t) & = & \boldsymbol{f}(\boldsymbol{x}(t))+\boldsymbol{v}(t), \\ \boldsymbol{y}(t) & = & \boldsymbol{C}\boldsymbol{x}(t)+\boldsymbol{w}(t),\end{array}\right. \tag{1}$$
$$(\boldsymbol{v}(t)\sim N(0,\Sigma_v), \boldsymbol{w}(t)\sim N(0,\Sigma_w)),$$

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 "*state 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 nonlinear in general, and any state transition of $\boldsymbol{x}$ can be represented by this formulation. While $\boldsymbol{C}$ is a linear transformation representing measurement facilities to derive the measurement variables in $\boldsymbol{y}$ from the state variables in $\boldsymbol{x}$, the facilities are artificial and linear in most cases. Thus, this does not reduce the generality of this expression. If $\boldsymbol{C}$ is a unit matrix, all state variables are directly observable through the measurement. If $\boldsymbol{C}$ is column full rank, the value of all state variables with the measurement noise can be estimated by solving the measurement equation with $\boldsymbol{x}$. Otherwise, some state variables cannot be estimated by the measurement equation only. Such state variables are called "*hidden state variables.*"

In practical setting of discovery, $\boldsymbol{f}(\boldsymbol{x})$ and some elements of $\boldsymbol{x}$ are initially unknown. We can know only subvector $\boldsymbol{x}'(\subseteq \boldsymbol{x})$ measured by artificial measurement facilities. Thus only a submatrix $\boldsymbol{C}'(\subseteq \boldsymbol{C})$ representing a relation between $\boldsymbol{x}'$ and $\boldsymbol{y}$ is initially known. So our proposing method should identify the correct dimension of $\boldsymbol{x}$ including hidden state variables based on given measurement data at first. Subsequently, it searches plausible candidates of $\boldsymbol{f}(\boldsymbol{x})$ reflecting the first principles.

2.2 Outline of Approach

The outline of our proposing method is shown in **Fig. 1**. Given a set of measurement data and knowledge on scale–types of measurement variables, the dimension of $\boldsymbol{x}$ is identified through a statistical analysis called "*correlation dimension analysis.*"[7] For each element of $\boldsymbol{y}$, the locus of its temporal change is mapped to a phase space constructed by time-delayed values of the element, and the degree of freedom which is dimension of $\boldsymbol{x}$ is estimated by calculating the sparseness of the locus in the phase space. Once the dimension of $\boldsymbol{x}$ is known, all possible combinations of scale–types of the elements in $\boldsymbol{x}$ are enumerated based on scale–type constraints from the known measurement submatrix $\boldsymbol{C}'$ and the scale–types of the elements in $\boldsymbol{y}$. Then for all combination, the admissible candidate equations of $\boldsymbol{f}(\boldsymbol{x})$ are generated. Subsequently, the validity of the candidate is tested through a simulation based tracking method called "*Sequential Importance Sampling/Resampling Monte Carlo filter*(SIS/RMC filter)"[8] on the given measurement data. Simulation based tracking is repeated for each candidate to optimize the coefficients in the candidate equations. Then, the combination of candidate $\boldsymbol{f}$ and its coefficients providing highly accurate tracking, in terms of "*mean square error* (MSE)", is selected as the discovered dynamic model of the objective process. Through these steps, SCALETRACK discovers the first principle based state space model of the objective process from passively observed data without detailed domain knowledge except for scale–types and measurement facilities.

3. Performance Evaluation

3.1 Basic Performance

The evaluation is made in terms of scale–types of state variables, hidden state variables and measurement noise levels by using two dimensional artificial formulae named RR and RI. Their equations are the followings.

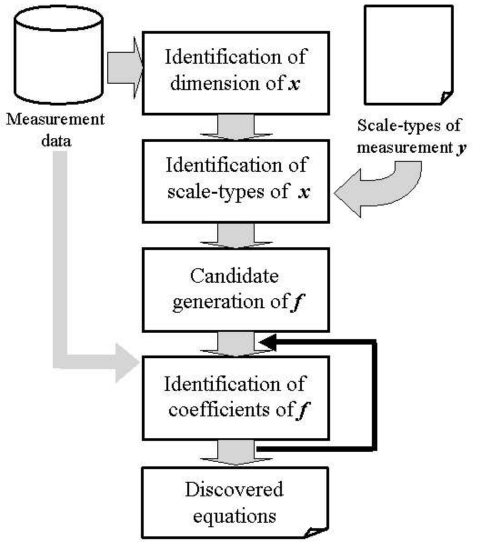
Measurement data
Identification of dimension of x
Scale-types of measurement y
Identification of scale-types of x
Candidate generation of f
Identification of coefficients of f
Discovered equations

Fig. 1 Block Diagram of Approach.

1. Model RR:
$$\begin{aligned}\dot{x}_1(t) &= x_1(t)x_2(t),\\ \dot{x}_2(t) &= -0.5x_2(t),\end{aligned}$$
$$\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} + \boldsymbol{w}_t,$$
where $y_1(t) = x_1(t)$ and $y_2(t) = x_2(t)$ are Ratio scale. The measurement data were generated by the simulations under one time step $\Delta t = 0.005$ and total steps $n = 600$.
2. Model RI:
$$\begin{aligned}\dot{x}_1(t) &= 0.4x_1(t)(x_2(t)+0.2),\\ \dot{x}_2(t) &= -0.1(x_2(t)+0.6),\end{aligned}$$
$$\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} + \boldsymbol{w}_t,$$
where $y_1(t) = x_1(t)$ is Ratio scale and $y_2(t) = x_2(t)$ Interval scale. The measurement data were generated by the simulations under one time step $\Delta t = 0.05$ and total steps $n = 600$.

The elements of measurement noise $\boldsymbol{w}_t$ are determined as follows.

$$w_t^h \sim N\left(0, \sigma_w x^h(t)\right),$$

where w_t^h is the h–th element of $\boldsymbol{w}_t$, $x^h(t)$ the h–th element of $\boldsymbol{x}(t)$, and σ_w a relative ampli-

Table 1 Basic Performance

case	*ct*	$\sigma_w(\%)$				
	(hrs.)	0.1	0.5	1.0	2.0	5.0∼
RR	1.5	++	+	+	+	-
RRH	5.5	+	+	-	-	-
RI	4.0	++	+	+	+	-
RIH	5.5	++	+	-	-	-

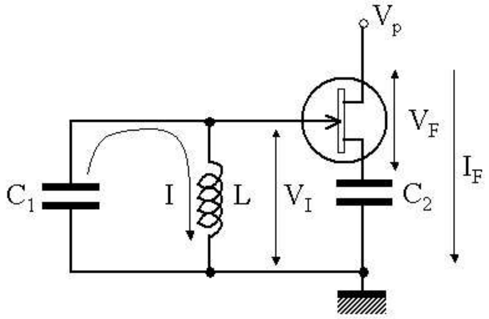

Fig. 2 An LC and FET Circuit

tude of measurement noise. The second measurement variable, y_2, is not available in RRH and RIH, and hence a hidden state variable exists. On the other hand, all state variables are observed in RR and RI. The correlation dimension analysis properly estimated the dimension of state vectors as nearly 2 in each case. The computation times *ct* required for RRH, RI and RIH were far longer than that of RR, because the variety of admissible formulae containing interval scale variables is far 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 +, 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 $0.1 - 0.5\%$ is required for hidden state cases.

3.2 Discovery of Circuit Dynamics

SCALETRACK has been applied to synthetic data of an electric circuit consisting of LC and Field Effect Transistor (FET) as shown in **Fig. 2**. Its state equation is represented as follows.

$$\begin{aligned}\dot{V}_I(t) &= -\frac{I(t)}{C_1} &&= -100I(t),\\ \dot{I}(t) &= \frac{V_I(t)}{L} &&= 50V_I(t),\\ \dot{V}_F(t) &= \frac{V_I(t)V_F(t)}{rC_2} &&= 250.0V_I(t)V_F(t),\end{aligned}$$

where the definitions of V_I, I, V_F, $L = 20$mH, $C_1 = 10$mF and $C_2 = 1$mF are clear in the figure and $r = 4.0\Omega$V a voltage–resistance coefficient of FET. All state variables are Ratio scale, and can be measured via corresponding Ratio scale measurement variables respectively. The measurement data were sampled under one time step $\Delta t = 0.001$, total time steps $n = 800$ and the relative measurement noise $\sigma_w = 0.1\%$. Because the dimension of the state vector, 2.94, was obtained in correlation dimension analysis, the state equation consisting of three state variables was searched.

In case that every state variables are directly measured, the following state equation having the best accuracy was derived.

$$\begin{aligned}\dot{V}_I(t) &= -133.3I(t),\\ \dot{I}(t) &= 6.94V_I(t)V_F(t),\\ \dot{V}_F(t) &= 249.0V_I(t)V_F(t).\end{aligned}$$

The shapes of the first and the third expressions of the equation are identical with those in the original equation though the values of coefficients are slightly different from the original.

Subsequently, the measurement of I was omitted to make I a hidden state variable. The following correct formula except for the discrepancy of coefficient values showed up within the solutions having top five accuracies.

$$\begin{aligned}\dot{V}_I(t) &= -26.9I(t),\\ \dot{I}(t) &= 298.0V_I(t),\\ \dot{V}_F(t) &= 250.0V_I(t)V_F(t).\end{aligned}$$

These results indicate that SCALETRACK has ability to discover state equations of engineering objects having three-dimensional dynamics at least.

4. Discussion

In this paper, we proposed a method named SCALETRACK which discovers the first principle based dynamic models of an objective process represented by simultaneous time differential equations. According to the experiments, SCALETRACK has an ability to discover state equations even if hidden states exist. SCALETRACK accepts at least 2.0% measurement noise in relative amplitude when hidden states do not exist. This is comparable with the noise level in practical cases where 1.0–2.0% measurement noise is the most widely seen in scientific and engineering applications. Even when a hidden state exists, 0.5% measurement noise in relative amplitude can be accepted by SCALETRACK. This noise level can also be achieved by using proper measurement facilities in many applications. The performance of SCALETRACK shows robustness against measurement noise to some extent.

Computational complexity of SCALETRACK is NP–hard in terms of the number of state variables, because the number of possible combinations of scale–types, the number of candidate state equations and the number of the possible values of the coefficients to be searched show combinational explosions when the number of state variables increase. This fact is reflected to the computation time required by SCALETRACK, where it took over 3 days to complete the search of the solutions having three state variables of Ratio scales. Although the computational time can be reduced by more limiting the search of the coefficients, the correctness of the solutions is also reduced. More efficient search algorithm should be studied in future work.

Another issue remained in this work is the noise robustness. This problem is also very important to establish wilder applicability of SCALETRACK to noisy situations. Some approaches to reduce the noise effect should be introduced to the estimation of the probability distribution of the states in the SIS/RMC filter in future study.

The advantage of SCALETRACK is that the equations discovered by the SCALETRACK are guaranteed that they are the first principle based equations because the candidates generated in SCALETRACK are constrained by the scale–types of variables and Extended Product Theorem. Scientists can easily avoid the solutions not reflecting the underlying first principles by using this method.

5. Conclusion

We showed a novel method to discover a simultaneous time differential equations representing the first principles governing dynamic behavior of an objective process from passively observed data. The significant features of this approach are the discovery without strong bias of the domain knowledge due to no use of knowledge specific to the objective process and the wide applicability to the cases including hidden state variables in the objective process. The remained major issue is to overcome the computational time complexity of the search. Under the current environment, the derivation of models consisting of more than a few differential equations are not very practical. The study to significantly increase the search speed is currently underway.

References

1) Langley, et al.: *Scientific discovery; Computational explorations of the creative process*. Cambridge, MA: MIT Press (1987).

2) Koehn, B. and Zytkow, J. M.: Experimenting and theorizing in theory formation. *Proceedings of the International Symposium on Methodologies for Intelligent Systems*, pp. 296–307, ACM SIGART Press (1986).

3) Falkenhainer B. C. and Michalski R. S.: Integrating Quantitative and Qualitative Discovery: The ABACUS System, *Machine Learning*, pp. 367–401, Boston, Kluwer Academic Publishers (1986).

4) Nordhausen, B. and Langlay, P.W.: An Integrated Approach to Empirical Discovery, *Computational Models of Scientific Discovery and Theory Formation*, Morgan Kaufman Publishers, Inc, San Mateo, California (1990).

5) Washio, T. and Motoda, H.: Discovering Admissible Models of Complex Systems Based on Scale–types and Identity Constraints, *Proceedings of IJCAI'97*, Vol.2, Nagoya, pp. 810–817 (1997).

6) Dzeroski, S. and Todorovski, L.: Discovering Dynamics: From Inductive Logic Programming to Machine Discovery, *Journal of Intelligent Information Systems*, Kluwer Academic Publishers, pp. 1–20 (1994).

7) Berge, P., Vidal, C. and Pomeau, Y.: translated by Aisawa, A.: Order within Chaos - Towards a Deterministic Approach to Turbulence, Sangyo-Tosho, ISBN: 4782800681, pp.133–147 (1992) (in Japanese)

8) Doucet, A., Godsill, S. and Andriew, C. : On Sequential Monte Carlo Sampling Methods for Bayesian Filtering, *Statistics and Computing*, Vol.10, Issue 3, pp.197–208 (2000).