

Toward the Discovery of First Principle Based Scientific Law Equations

Takashi Washio and Hiroshi Motoda

Institute for Scientific and Industrial Research, Osaka University,
8-1, Mihogaoka, Ibarakishi, Osaka, 567-0047, Japan
{washio, motoda}@sanken.osaka-u.ac.jp

Abstract. Conventional work on scientific discovery such as BACON derives empirical law equations from experimental data. In recent years, SDS introducing mathematical admissibility constraints has been proposed to discover first principle based law equations, and it has been further extended to discover law equations from passively observed data. Furthermore, SSF has been proposed to discover the structure of a simultaneous equation model representing an objective process through experiments. In this report, the progress of these studies on the discovery of first principle based scientific law equations is summarized, and the future directions of this research are presented.

1 Introduction

Langley and others' BACON [1] is the most well known pioneering work to discover a complete equation representing scientific laws governing an objective process under experimental observations. FAHRENHEIT [2], ABACUS [3], etc. are the successors of BACON that use basically similar algorithms. However, a drawback of the BACON family, that is their low likelihood of the discovered equations being the first principle underlying the objective process, is reported. To alleviate the drawback, some systems, *e.g.*, ABACUS and COPER [4], utilize the information of the unit dimensions of quantities to prune the meaningless terms. However, many of these conventional scientific equation discovery systems have some critical limitations for the real world applications. First, the information of the unit dimension of each quantity in the data is needed to discover the first principle based equation. Second, the data must be acquired under “*active observations*” where the values of some quantities representing the objective process are observed for various process states by controlling the values of the other relevant quantities. Third, a complex equation model, especially a “*simultaneous equation model*”, to represent the process consisting of multiple mechanisms is hardly discovered due to the complexity of the search space.

To alleviate the first limitation, a law equation discovery system named SDS based on the mathematical constraints of “*scale-type*” and “*identity*” is proposed for the active observations [5]. Since the knowledge of scale-types of quantities is widely obtained in various domains, SDS is applicable to non-physics domains.

The equations discovered by SDS are highly likely to represent the first principle underlying the objective process. To address the second limitation, SDS has been further extended by introducing a novel principle named “*quasi-bi-variate fitting*” [6] to make it applicable to the “*passive observations*” where the quantities of the objective process can only be partially or even hardly controlled. Moreover, to overcome the third limitation, a simultaneous structure finding system named SSF has been proposed to discover a valid simultaneous equation structure under the active observations [7]. SSF identifies the number of equations needed to represent the objective process, and further identifies the sets of quantities to appear in each of the respective equations of the model while eliminating quantities irrelevant to the equations. The combination of SDS and SSF enables the discovery of the first principle based simultaneous equation model for the objective process under active observations. In this report, the principles and the performances of these recent scientific discovery systems are outlined, and the future directions of this research topic are discussed.

2 Required Conditions

Prior to the explanation of each scientific law equation discovery system, the conditions required for the application of the system are summarized. The approach of the SDS requires the following conditions.

- (1) All of the quantities except one dependent quantity can be controlled to their arbitrary values in the ranges of the quantities of our interest.
- (2) The objective system can be represented by a complete equation in the value range.
- (3) The scale-types of all quantities needed to represent the objective system are known.

The first condition is to ensure the application of the SDS to experimental environment in scientific laboratories. This condition is the requirement of the original BACON systems, and is also required by other BACON family. The second condition is also a common requirement in BACON family to search for a complete equation for every continuous region in the objective system. The third condition comes from the fact that the SDS uses the the information of the scale-types of the quantities to search mathematically admissible equation formulae to relate the quantities. As the scale-types of the measurement quantities are widely known based on the measurement theory [5], this condition does not restrict the applicability of the SDS.

The extended SDS discovers the law equations from the passively observed data, and thus the aforementioned condition (1) is not required. Instead, it needs the following condition in addition to the conditions (2) and (3).

- (4) The observed data are uniformly distributed over the possible states of the objective system.

Violation of the requirement of the uniform distribution over a certain value range of a quantity implies the low observability of the quantity [6],[8]. Any approaches such as the linear system identification and the neural network do not derive valid models under low observability. This limitation is generic, and further discussion on this issue is out of scope of this report.

The condition required by the SSF for the derivation of the simultaneous law equation models is as follows.

- (5) The simultaneous equation model under consideration is not over-constrained where the number of the equations is no more than the number of quantities in the model.

This condition always holds for the models in scientific and engineering domains, since the over-constrained state does not exist in any real world process.

3 Smart Discovery System (SDS)

The information required from the user besides the actual measurements in SDS is a list of the quantities and their scale-types. The rigorous definition of scale-type was given by Stevens[9]. The quantitative scale-types are interval scale, ratio scale and absolute scale, and these are the majorities of the quantities. Examples of the interval scale quantities are temperature in Celsius and sound tone where the origins of their scales are not absolute, and are changeable by human's definitions. Its admissible unit conversion follows "*Generic linear group*: $x' = kx + c$ ". Examples of the ratio scale quantities are physical mass and absolute temperature where each has an absolute zero point. Its admissible unit conversion follows "*Similarity group*: $x' = kx$ ". Examples of the absolute scale quantities are dimensionless quantities. It follows "*Identity group*: $x' = x$ ".

The first step of the algorithm of SDS searches bi-variate relations among the quantities by using the "*scale-type constraint*." Two well-known theorems which are deduced from the group structures of the scale-types provides the basis of this step [5].

Theorem 1 (Extended Buckingham Π -theorem) *If $\phi(x_1, x_2, x_3, \dots) = 0$ is a complete equation where each $x_i \in Q$, the set of quantities given in the experiment, and if each argument is one of interval, ratio and absolute scale-types, then the solution can be written in the form*

$$F(\Pi_1, \Pi_2, \dots, \Pi_{n-w}) = 0,$$

where n is the number of arguments of ϕ , w is the basic number of bases in x_1, x_2, x_3, \dots , respectively. For all i , Π_i is an absolute scale-type quantity.

A base is such a basic scaling factor which has a degree of freedom independent of the other bases in the given ϕ . For instance, length [L], mass [M] and time [T] of physical dimension and the origin of the temperature in Celsius are the examples of the bases. The relation of each Π_i to the arguments of ϕ is given by the following theorem [5].

Theorem 2 (Extended Product Theorem) *Let primary quantities in a set RQ are ratio scale-type and those in another set IQ interval scale-type, the function ρ relating a secondary quantity Π to $x_i \in RQ \cup IQ$ has the forms:*

$$\Pi = \left(\prod_{x_i \in R} |x_i|^{a_i} \right) \left(\prod_{I_k \subseteq I} \left(\sum_{x_j \in I_k} b_{kj} |x_j| + c_k \right)^{a_k} \right)$$

$$\Pi = \sum_{x_i \in R} a_i \log |x_i| + \sum_{I_k \subseteq I} a_k \log \left(\sum_{x_j \in I_k} b_{kj} |x_j| + c_k \right) + \sum_{x_\ell \in I_g \subseteq I} b_{g\ell} |x_\ell| + c_g$$

where all coefficients except Π are constants and $I_k \cap I_g = \phi$.

These theorems state that any meaningful complete equation consisting of the arguments of interval, ratio and absolute scale-types can be decomposed into an equation of absolute scale-type quantities having an arbitrary form and some equations of interval and ratio scale-type quantities in products and logarithmic form. The former $F(\Pi_1, \Pi_2, \dots, \Pi_{n-w}) = 0$ is called an “ensemble” and the latter $\Pi = \rho(x_1, x_2, x_3, \dots)$ s “regime”s.

If any pair of interval and/or ratio scale quantities $\{x, y\}$ in a given complete equation is to belong to an identical regime, they has to have a relation that follows the Theorem 2. Conversely, SDS searches bi-variate relations in the set of quantities Q where the relations have the following product, linear or logarithmic forms which are deduced from the Theorem 2.

$$x^a y = b, \quad \text{where } x, y \text{ are ratio scale,} \quad (1)$$

$$ax + y = b, \quad \text{where } x, y \text{ are interval scale, and} \quad (2)$$

$$a \log x + y = b \quad \text{or} \quad cx^a + y = b, \quad (3)$$

where x is ratio scale, and y interval scale.

The value of the constant a in each formula must be independent of any other quantities according to Extended Product Theorem, while the constants b and c are dependent on the other quantities in the regime. SDS applies the least square fitting of these relations to the bi-variate experimental data of x and y that are measured while holding the other quantities constant in Q , and determines the values of coefficients in the bi-variate relations. Subsequently, the judgment is made whether this equation fits the data well enough by some statistical tests.

This procedure is now demonstrated by an example of a complex system depicted in Figure 1. This is a circuit of photo-meter to measure the rate of increase of photo intensity within a certain time period. The resistance and switch parallel to the capacitor and the current meter are to reset the operation of this circuit. The actual model of this system is represented by the following complex equation involving 18 quantities.

$$\left(\frac{R_3 h_{fe_2}}{R_3 h_{fe_2} + h_{ie_2}} \frac{R_2 h_{fe_1}}{R_2 h_{fe_1} + h_{ie_1}} \frac{r L^2}{r L^2 + R_1} \right) (V_1 - V_2) - \frac{Q}{C} - \frac{K h_{ie_3} X}{B h_{fe_3}} = 0. \quad (4)$$

Here, L and r are photo intensity and sensitivity of the Csd device. X , K and B are position of indicator, spring constant and intensity of magnetic field of the

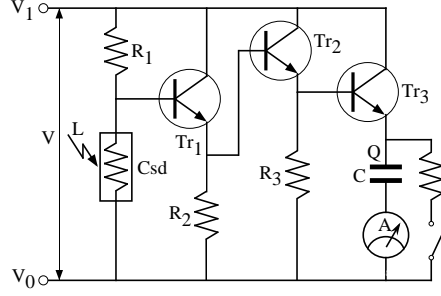


Fig. 1. A circuit of photo-meter

current meter respectively. h_{ie_i} is input impedance of the base of the i -th transistor. h_{fe_i} is gain ratio of the currents at the base and the collector of the i -th transistor. The definitions of the other quantities follow the standard symbolic representations of electric circuit. h_{fe_i} are absolute scale, V_1 and V_2 interval scale, and the rest ratio scale. X is the dependent quantity in this circuit, and the others are independently controllable by the change of boundary conditions and the replacement of devices. SDS requests the bi-variate change of quantities to the experimental environment. When a quantity is dependent (not controllable) during the search process, SDS modifies its request to control the other independent quantity. A simulation based experimental environment was designed and build for the circuit system. $\pm 4\%$ (std.) of relative Gaussian noise was added to both of the control quantity (input) and the measured quantity (output) in every bi-variate test. First, SDS set the set of interval scale quantities IQ as $\{V_1, V_2\}$, ratio scale quantities RQ as $\{L, r, R_1, R_2, R_3, h_{ie_1}, h_{ie_2}, h_{ie_3}, Q, C, X, K, B\}$ and absolute scale quantities AQ as $\{h_{fe_1}, h_{fe_2}, h_{fe_3}\}$ based on the input information on scale-types. Next, it performed the bi-variate fitting of a linear form Eq. (2) to the experimental data among the quantities in IQ , and applied the statistical tests. Then, SDS figured out a set consisting of a bi-variate equation $IE = \{\Theta_0 = 1.000V_1 - 1.000V_0\}$ quickly. Subsequently, a product form Eq. (1) among the quantities in $RQ \cup \{\Theta_0\}$ is searched, and applied the statistical tests.

The resultant set of the bi-variate equations RE that passed the tests was as follows.

$$RE = \{L^{1.999}r = b_1, L^{-1.999}R_1 = b_2, r^{-1.000}R_1 = b_3, R_2^{-1.000}h_{ie_1} = b_4, R_3^{-1.000}h_{ie_2} = b_5, Q^{-1.000}C = b_6, h_{ie_3}^{1.000}X = b_7, h_{ie_3}^{1.000}K = b_8, h_{ie_3}^{-1.000}B = b_9, X^{1.000}K = b_{10}, X^{-0.999}B = b_{11}, K^{-1.000}B = b_{12}\}$$

The bi-variate fitting of Eq. (3) for the other pairs across IQ and RQ have also been conducted. But no equations have passed the statistical tests.

In the next step, triplet consistency tests are applied to every triplet of equations in $IE \cup RE$. In case of a triplet of the power form equations, $x^{a_{xy}}y = b_{xy}$, $y^{a_{yz}}z = b_{yz}$, $x^{a_{xz}}z = b_{xz}$, by substituting y in the first to y in the second, we obtain $x^{-a_{yz}a_{xy}}z = b_{xy}^{-a_{yz}}b_{yz}$. Thus, the following condition must be met.

$$a_{xz} = -a_{yz}a_{xy}.$$

However, if any of the three equations are not correct due to the noise and error of data fitting, this relation may not hold. Thus, a statistical test judges if the three of the equations are mutually consistent in terms of as . For the triplet of the linear form equations, a similar test is applied based on the identical principle. SDS applies this test to every triplet of equations in $IE \cup RE$, and search every maximal convex set MCS where each triplet of equations among the quantities in this set are mutually consistent. The value of each a is evaluated by its average \bar{a} over the equations in the MCS . When the value of \bar{a} is close enough to its nearest integer within its statistical error bound, it is set to the integer value. This operation is based on the observation that the majority of the first principle based equations have integer power coefficients. Finally, the merged quantities are replaced by the term of each equation of the derived regime in $IQ \cup RQ$. In the example in Figure 1, the final forms of regimes were represented by the merged terms in $IQ \cup RQ$ as follows.

$$IQ \cup RQ = \{\Pi_1 = R_1 r^{-1} L^{-2}, \Pi_2 = h_{ie_1} R_2^{-1}, \Pi_3 = h_{ie_2} R_3^{-1}, \Pi_4 = h_{ie_3} XKB^{-1}, \Pi_5 = QC^{-1}, \Theta_0 = V_1 - V_0\}$$

Once all regimes are identified, new terms are further generated by merging these regimes in $IQ \cup RQ \cup AQ$. SDS searches bi-variate relations having one of the formulae $x^a y = b$ (product form) and $ax + y = b$ (linear form) by adopting the least square fitting of these formulae. Then, the statistical tests of the goodness of the fitting are applied. In the example of the circuit, the product form was applied first. They were merged to the following new terms.

$$\begin{aligned} \Theta_1 &= \Pi_1 h_{fe_1} = R_1 r^{-1.0} L^{-2.0} h_{fe_1}, \\ \Theta_2 &= \Pi_2 h_{fe_2} = h_{ie_1} R_2^{-1.0} h_{fe_2}, \\ \Theta_3 &= \Pi_3 h_{fe_3} = h_{ie_2} R_3^{-1.0} h_{fe_3}. \end{aligned}$$

Next, the linear form was tested, then a form was found.

$$\Theta_4 = \Pi_4 + \Pi_5 = h_{ie_3} XKB^{-1.0} + QC^{-1.0}$$

Thus, $IQ \cup RQ \cup AQ$ became as $\{\Theta_0, \Theta_1, \Theta_2, \Theta_3, \Theta_4\}$. Again, by applying the linear form, another was newly generated.

$$\Theta_5 = \Theta_0 \Theta_4^{-1.0} = (V_1 - V_0)(h_{ie_3} XKB^{-1.0} + QC^{-1.0})^{-1.0}$$

Thus, $IQ \cup RQ \cup AQ = \{\Theta_1, \Theta_2, \Theta_3, \Theta_5\}$. As no new terms became available, this step was finished.

In the final step, the “*identity constraint*” are applied to further merge terms. The basic principle of the identity constraints comes by answering the question that “*what is the relation among Θ_h , Θ_i and Θ_j , if $\Theta_i = f_{\Theta_j}(\Theta_h)$ and $\Theta_j = f_{\Theta_i}(\Theta_h)$ are known?*” For example, if $a(\Theta_j)\Theta_h + \Theta_i = b(\Theta_j)$ and $a(\Theta_i)\Theta_h + \Theta_j = b(\Theta_i)$ are given, the following relation is deduced.

$$\Theta_h + \alpha_1 \Theta_i \Theta_j + \beta_1 \Theta_i + \alpha_2 \Theta_j + \beta_2 = 0$$

This principle is generalized to various relations among multiple terms. In the example of the circuit, SDS found a set of the bi-variate linear relations on the combinations of $\{\Theta_1, \Theta_5\}$, $\{\Theta_2, \Theta_5\}$ and $\{\Theta_3, \Theta_5\}$. By applying the identity constraint, the following multi-linear formula has been obtained.

$$\Theta_1\Theta_2\Theta_3 + \Theta_1\Theta_2 + \Theta_2\Theta_3 + \Theta_1\Theta_3 + \Theta_1 + \Theta_2 + \Theta_3 + \Theta_5 + 1 = 0$$

Because every coefficient is independent of any terms, this is considered to be the ensemble equation. The equivalence of this result to Eq. (4) is easily checked by substituting the intermediate terms to this ensemble equation.

4 Extended SDS

As noted in the previous section, the bi-variate fitting requires experimental control of some quantities, and is not applicable to the passive observation environments. To overcome this difficulty, the “*quasi-bi-variate fitting*” procedure depicted in Figure 4 is used to extract a bi-variate relation between two quantities under the approximated constant values of the other quantities. Let $OBS = \{X_1, X_2, \dots, X_n\}$ be a set of observations where each X is a m -dimensional vector of m quantities. The fitting of a candidate bi-variate formula for a pair of two quantities $P_{ij} = \{x_i, x_j\} (\subseteq X)$ is applied to a subset of OBS . This subset OBS_{ijg} is chosen in such a way that every quantity $x_k \in (X - P_{ij})$ takes a value in the vicinity of the value of x_{kg} , where $X_g = \{x_{1g}, x_{2g}, \dots, x_{mg}\} \in OBS$ is an arbitrary chosen observation vector. The vicinity of x_{kg} is defined as

$$\Delta x_k = |x_k - x_{kg}| < \epsilon_k.$$

ϵ_k determines the size of the vicinity. This vicinity is indicated by a rectangular cube in the left figure of Fig. 4. Every admissible bi-variate formula is generally represented in the form

$$F_{ij}(P_{ij}, a_{ij}, G_{ij}(X - P_{ij}), H_{ij}(X - P_{ij})) = 0. \quad (5)$$

Here, G_{ij} and H_{ij} are dependent on the quantities in $X - P_{ij}$, while a_{ij} remains constant. Given an OBS_{ijg} , if each ϵ_k is moderately small, the values of G_{ij} and H_{ij} become almost constant. The least square fitting of Eq.(5) approximately provides the functional relation within P_{ij} and the coefficient a_{ij} as depicted in the bottom figure of Fig. 4 while almost excluding the influence of the other dimensions $X - P_{ij}$. The goodness of the fitting is judged by some statistical tests. For the bi-variate relations of the identity constraints, the similar scheme of the quasi-bi-variate fitting is applied.

The proposed method has been applied to a real world problem. The objective of the application is to discover a generic law formula governing the mental preference of people on their houses subject to the cost to buy the house and the social risk at the place of the house. We designed a questionnaire sheet to ask the preference of the house in the trade off between the frequency of huge

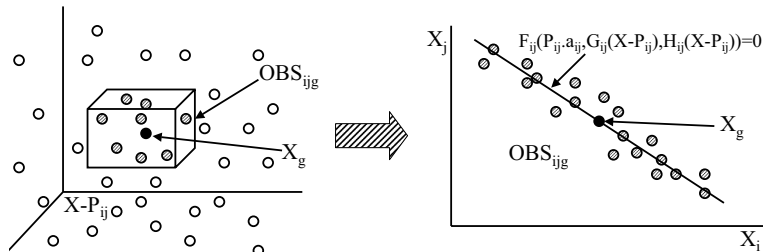


Fig. 2. Outline of quasi-bi-variate fitting

earthquakes, x_1 (earthquake/year), and the cost to buy, x_2 (\$). These are ratio scale quantities. In the questionnaire, 9 combinations of the cost and the earthquake frequency are presented, and each person chooses its preference from the 7 levels for each combination. We distributed this questionnaire sheet to the people owning their houses in the suburb area of Tokyo, and totally 400 answer sheets are collected back. The answer data has been processed by following the method of successive categories which is widely used in the experimental psychology to compose an interval scale preference index y [10], and $OBS = \{X_1, X_2, \dots, X_{400}\}$ where $X_i = [x_{1i}, x_{2i}, y_i]$ is obtained. The expected basic structure of the law equation governing the data is $y = f(x_1, x_2)$. The quasi-bi-variate fitting between x_1 and x_2 was applied, and $x_1 = a(y)x_2^{-0.25}$ have been obtained. Next, the formulae between x_1 and y have been identified as either one of $y = a(x_2)x_1^{-0.23} + b(x_2)$ and $y = 0.62 \log x_1 + b(x_2)$. Similar search has been made for x_2 and y , and, likewise, two candidate equations $y = a(x_1)x_2^{0.026} + b(x_1)$ and $y = 0.34 \log x_2 + b(x_1)$ have been derived. Subsequently, the triplet-test among $\{x_1, x_2, y\}$ is conducted, and only the following two candidates have passed the test.

$$y = 0.63 \log x_1 + 0.34 \log x_2 - 2.9 \quad (6)$$

$$y = -0.61 x_1^{-0.23} x_2^{0.026} + 3.2 \quad (8)$$

Though both equations are admissible as law equations based on the mathematical constraints and the given data, Eq. (6) is preferred as a law equation in terms of the principle of parsimony, because it gives less error for the questionnaire data, and has less number of parameters.

5 Simultaneous Structure Finder (SSF)

The principle to discover the simultaneous equation structure from experimental data is based on some fundamental and generic characteristics of simultaneous equation models presented in the past work [7]. The principle is briefly explained though an example electric circuit depicted in Fig. 3. This can be represented

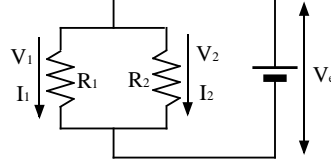


Fig. 3. An circuit of parallel resistances.

by the following simultaneous equation model.

$$V_1 = I_1 R_1 \text{ \#1}, V_2 = I_2 R_2 \text{ \#2}, V_e = V_1 \text{ \#3 and } V_e = V_2 \text{ \#4}, \quad (9)$$

where R_1, R_2 : two resistances, V_1, V_2 : voltage differences across resistances, I_1, I_2 : electric current going through resistances and V_e : voltage of a battery. We consider an experiment to externally control some values of the quantities in this model. For example, the quantities R_1 and V_e can be externally controlled by the specification of the resistance and the battery. If we specify these values in Eq.(9), the values of the other quantities, V_1, V_2 and I_1 , that are involved in the first, the third and the fourth equations, #1, #3 and #4, are determined since the number of the quantities which are not externally specified is equal to the number of the equations. But, this external control does not determine the values of R_2 and I_2 through the equation #2. Thus, the equation set {#1, #3, #4} is considered to represent a mechanism which determines the state of a part of the objective process. We introduce the following definition to characterize this mechanism in the simultaneous equation model.

Definition 1 (complete subset) *Given a set of equations, E , let the set of all quantities be Q appearing in the equations in E . Given a quantity set $SQ(\subset Q)$ for external specification, when the values of all quantities in $NQ = CQ - SQ$ are determined where CQ ($SQ \subset CQ \subset Q$) is a set of all quantities appearing in a set of equations $CE(\subseteq E)$, CE is called a “complete subset”. The cardinality $|CE| = |NQ|$ is called the “order” of the complete subset.*

The equation set {#1, #3, #4} is a complete subset of the order 3. Under any external control of two quantities among R_1, V_e, V_1, V_2 and I_1 , {#1, #3, #4} always determines the values of the remained three quantities. Thus, the complete subset is “invariant” for the selection of the externally controlled quantities.

The complete subset gives an important foundation to discover the structure of the simultaneous equation model, which appropriately reflects the dependency, embedded in the observation of quantities. For example, the circuit in Fig. 3 can be represented by the following different simultaneous equation formulae

$$I_1 R_1 = I_2 R_2 \text{ \#1'}, V_2 = I_2 R_2 \text{ \#2}, V_e = V_1 \text{ \#3 and } V_e = V_2 \text{ \#4}. \quad (10)$$

If the same specification on V_e and R_1 is made in Eq.(10), a different complete subset {#3, #4} is obtained, and any complete subset to determine the value of I_1 does not exist since the equation #1' cannot determine the value of I_1 without the constraint of #2. #1' and #2 that include the undetermined quantities I_2

and R_2 do not satisfy Definition 1. In the real experiment on the electric circuit, the value of I_1 is physically determined, and this fact contradicts the consequence derived by the analysis on Eq.(10). In contrast, the model of Eq.(9) always gives correct answers on the determination of quantities for any external specifications of quantities. The model having the complete subsets, which are isomorphic with the actual dependency among quantities, is named a “*structural form*”.

Conversely, if we identify all complete subsets from the experimental observation of quantities in the objective process, and compose a simultaneous equation model consisting of these complete subsets, the model is ensured to be the structural form. The following theorem provides a basis for the composition [7].

Theorem 3 (modular lattice theorem) *Given a model of an objective process consisting of equations E , the set of all complete subsets of the model, i.e., $L = \{\forall CE_i \subseteq E\}$, forms a modular lattice of the sets for the order of the complete subsets, i.e., $\forall CE_i, CE_j \in L, CE_i \cup CE_j \in L, CE_i \cap CE_j \in L$ and $n(CE_i \cup CE_j) = n(CE_i) + n(CE_j) - n(CE_i \cap CE_j)$ where n is the order of a given complete subset.*

For instance, the following four complete subsets having the modular lattice structure can be found in the example of Eq.(9).

$$\begin{aligned} &\{\#3, \#4\}(n = 2), \{\#1, \#3, \#4\}(n = 3), \\ &\{\#2, \#3, \#4\}(n = 3), \{\#1, \#2, \#3, \#4\}(n = 4). \end{aligned}$$

Because the complete subsets of an objective process mutually overlap in the modular lattice, the redundant overlaps must be removed in the model composition by introducing the following definition of independent component.

Definition 2 (independent component of a complete subset) *The independent component DE_i of the complete subset CE_i is defined as*

$$DE_i = CE_i - \bigcup_{\forall CE_j \subset CE_i \text{ and } CE_j \in L} CE_j,$$

where L is the set of all complete subsets of the model. The order δn_i of DE_i is defined as $\delta n_i = |DE_i|$.

For instance, the following independent components can be found for Eq.(9).

$$\begin{aligned} DE_1 &= \{\#3, \#4\} - \phi = \{\#3, \#4\}, \delta n_1 = 2 - 0 = 2, \\ DE_2 &= \{\#1, \#3, \#4\} - \{\#3, \#4\} = \{\#1\}, \delta n_2 = 3 - 2 = 1, \\ DE_3 &= \{\#2, \#3, \#4\} - \{\#3, \#4\} = \{\#2\}, \delta n_3 = 3 - 2 = 1. \end{aligned}$$

Because the independent components do not overlap, their collection represents the structure of the simultaneous equation model.

However, the issue on the ambiguity of the representation of the structural form still remains. For example, the set of equations $\{V_1 = I_1 R_1 \#1, V_e = V_1 \#3, V_e = V_2 \#4\}$ in Eq.(9) which is a complete subset of order 3 can be transformed by the linear transformation as follows.

$$2V_e + V_1 + V_2 = 4I_1R_1 \#1, \quad 2V_e = 2V_1 - V_2 + I_1R_1 \#3,$$

$$\text{and } 3V_e = -V_1 + 2V_2 + 2I_1R_1 \#4.$$

This transformation preserves the complete subset, and the model remains as a structural form. This ambiguity of the equation representation in a complete subset can cause combinatorial explosion in the enumeration of the structural forms. As indicated in the above example, if the set of all quantities, CQ , appearing in a complete subset CE is preserved through some transformation maintaining quantitative equivalence, the complete subset is also preserved [7]. Accordingly, only the following formula of a complete subset is focused in the search.

Definition 3 (canonical form of a complete subset) *Given a complete subset CE , the “canonical form” of CE is the form where all quantities in CQ appears in each equation in CE .*

An example of the canonical form is Eq.(11). Based on this definition, the structural canonical form of a simultaneous equation model is further defined.

Definition 4 (structural canonical form of simultaneous equations) *The “canonical form” of simultaneous equations consists of the equations in $\cup_{i=1}^b DE_i$ where each equation in DE_i is represented by the canonical form in the complete subset CE_i , and b is the total number of DE_i . If the canonical form of simultaneous equations is derived to be a “structural form”, then the form is named “structural canonical form”.*

The structural form of Eq.(9) is shown as follows.

$$DE_1 = \{f_{11}(V_e, V_1, V_2) = 0 \#3, \quad f_{12}(V_e, V_1, V_2) = 0 \#4\},$$

$$DE_2 = \{f_2(V_e, V_1, V_2, I_1, R_1) = 0 \#1\}, \quad DE_3 = \{f_3(V_e, V_1, V_2, I_2, R_2) = 0 \#2\} \quad (11)$$

where $f(\bullet) = 0$ is an arbitrary formula to represent a quantitative relation. Because Eq.(9) is a structural form, Eq.(11) is the structural canonical form. The concrete shape of each formula can be derived by the aforementioned SDS, once the structure of simultaneous equations is given.

6 Discussion and Conclusion

Dzeroski and Todorovski developed LAGRANGE [11] and LAGRAMGE [12] which discover simultaneous equation models from observed data. However, the mathematical admissibility is not introduced sufficiently in the discovery process, and many redundant representations of simultaneous equations can be derived at an expense of high computational complexity. COPER, which also discovers simultaneous equations, uses very strong mathematical constraints based on the unit dimensions to prune the meaningless terms [4]. However, it inevitably requires the unit information which is not frequently obtained in non-physical domains. The major advantages of our proposing methods in comparison with the past approaches are the efficiency of the equation search, the soundness of

the discovery in terms of the first principle and the wide applicability not limited to the physical domain. These are achieved by introducing the criteria of generic mathematical admissibility.

The future directions of our work are

- (1) Discovery of simultaneous equations from passively observed data and
- (2) Discovery of dynamic law equations from passively observed time series data.

For the former purpose, the principle of SSF should be extended to discover the structure of simultaneous equations from passively observed data. This may become possible by developing a new data sampling technique similar to the quasi-bi-variate fitting of the extended SDS. The latter approach is expected to discover the first principle based differential equations which appear in various scientific and engineering domains. The major issue to overcome will be the statistical information processing on the noise contained in the time series data which are widely seen in many problems. These scientific law equation discovery methods will provide a new measure of system modeling based on data in wide scientific and engineering domains.

References

- [1] Langley, P. W., Simon, H. A., Bradshaw, G. L. and Zytkow, J. M.: *Scientific Discovery; Computational Explorations of the Creative Process*, MIT Press, Cambridge, Massachusetts (1987)
- [2] Koehn, B. and Zytkow, J. M.: Experimenting and theorizing in theory formation. In *Proceedings of the International Symposium on Methodologies for Intelligent Systems*, ACM SIGART Press (1986) 296–307
- [3] Falkenhainer, Br. C. and Michalski, R. S.: Integrating Quantitative and Qualitative Discovery: The ABACUS System. In *Machine Learning*, Boston, Kluwer Academic Publishers (1986) 367–401
- [4] Kokar, M. M.: Determining Arguments of Invariant Functional Descriptions. In *Machine Learning*, Boston, Kluwer Academic Publishers (1986) 403–422
- [5] Washio, T. and Motoda, H.: Discovering Admissible Models of Complex Systems Based on Scale-Types and Identity Constraints, In *Proceedings of IJCAI'97*, Vol.2, Nagoya (1997) 810–817
- [6] Washio, T., Motoda, H. and Niwa, Y.: Discovering admissible model equations from observed data based on scale-types and identity constraints. In *Proceedings of IJCAI'99*, Vol.2 (1999) 772–779
- [7] Washio T. and Motoda, H.: Discovering Admissible Simultaneous Equations of Large Scale Systems, In *Proceedings of AAAI'98*, Madison (1998) 189–196
- [8] Ljung, L.: *System Identification*, P T R Prentice-Hall (1987)
- [9] Stevens S.S.: On the Theory of Scales of Measurement, In *Science* (1946) 677–680
- [10] Torgerson, W. S.: In *Theory and Methods of Scaling*, N.Y.: J. Wiley (1958)
- [11] Dzeroski, S. and Todorovski, L.: Discovering Dynamics: From Inductive Logic Programming to Machine Discovery. In *Journal of Intelligent Information Systems*, Boston, Kluwer Academic Publishers (1994) 1–20
- [12] Todorovski, L. and Dzeroski, S.: Declarative Bias in Equation Discovery, In *Proceeding of the fourteenth International Conference on Machine Learning*, San Mateo, CA, Morgan Kaufmann (1997) 376–384