Discovering Admissible Models of Complex Systems Based on Scale-Types and Identity Constraints

Takashi Washio and Hiroshi Motoda
Institute for the Scientific and Industrial Research, Osaka University
8-1, Mihogaoka, Ibarakishi, Osaka 567, JAPAN
e-mail: {washio, motoda}@sanken.osaka-u.ac.jp

Abstract

SDS is a discovery system from numeric measurement data. It outperforms the existing systems in every aspect of search efficiency, noise tolerancy, credibility of the resulting equations and complexity of the target system that it can handle. The power of SDS comes from the use of the scale-types of the measurement data and mathematical property of identity by which to constrain the admissible solutions. Its algorithm is described with a complex working example and the performance comparison with other systems are discussed.

1 Introduction

Finding regularities in the data is a basis of knowledge acquisition by induction. One such typical and challenging task is inducing quantitative formulae of scientific laws from measurement data. Langley and others' BACON systems [Langley *et al.*, 1985] are most well known as the pioneering work. They founded the succeeding BACON family. FAHRENHEIT [Koehn and Zytkow, 1986], ABACUS [Falkenhainer and Michalski, 1985], IDS [Nordhausen and Langlay, 1990] and KEPLER [Wu and Wang, 1989] are such successors that basically use similar algorithms to BACON in search for a complete equation governing the data measured in a continuous process. However, recent work reports that there is considerable ambiguity in their results under noisy data even for the relations among small number of quantities [Schaffer, 1990; Huang and Zytkow, 1996]. Another drawback of the BACON family is the complexity of hypothesis generation. This also limits their applicability to find a complex relation that holds among many quantities.

To alleviate these drawbacks, some members of the BACON family, e.g. ABACUS, utilizes the information of the quantity dimension to prune the meaningless terms based on the principle of dimensional homogeneity. However, this heuristic still leaves many types of equations in candidates. COPER [Kokar, 1985], another type of equation finding systems based on a principle of dimensional analysis called "*Buckingham's* Π*-theorem*" [Buckingham, 1914], can significantly reduce the candidate generation by explicit use of the information about the quantity dimension. Its another significant advantage is higher credibility of the solution that it is not merely an experimental equation but is indeed a first principle equation. However, these approaches are not applicable when the information of the quantity dimension is not available. This fact strongly limits their applicability to non-physics domains.

The primary objective of this study is to establish a method to discover an admissible complete equation governing a complex system where its domain is not limited to physics ensuring as much as possible its property being the first principles. Any other technical areas, including system identification theory [Ljung, 1987], have not addressed to automatically derive first principle based models of complex systems from measurement data. Our goal if attained will provide an advantageous means not only for the field of scientific discovery but also for the analysis of complex systems in engineering. As a step towards this goal, we developed a quantitative model discovery system "*Smart Discovery System (SDS)*" implementing our new approach. SDS utilizes newly introduced constraints of *scale-type* and *identity* both of which highly constrain the generation of candidate terms. Because these are not heuristics but mathematical constraints, the generated candidates are highly credible. SDS also adopts bi-variate equation generation based on data fitting. But what makes SDS different from BACON family is that it employs *triplet checking* of the validity of those bi-variate equations, a quite strong mathematical constraint. It should be emphasized that SDS does not require the information about quantity dimension. The information required besides the measurements is the knowledge of scale-type of each quantity. This feature expands the scope of its applicability since the knowledge of scale-types is widely obtained in various domains including psychophysics, sociology and etc.

2 Outline of Method

SDS requires two assumptions on the feature of the objective system to be analyzed. One is that the objective system can be represented by a single quantitative, con-

tinuous and complete equation for the quantity ranges of our interest. Another is that all of the quantities in the equation can be measured, and all of the quantities except one dependent quantity can be controlled to their arbitrary values in the range. The latter is a common assumption in BACON family. The former is the assumption of the original BACON systems, and is also assumed by other BACON family (*i.e.*, search made for a complete equation for *every* continuous region in the objective system).

The information required from the user besides the actual measurements is a list of the quantities and their scale-types. The rigorous definition of scale-type was given by Stevens[Stevens, 1946]. He defined the measurement process as "*the assignment of numerals to object or events according to some rules.*" He claimed that different kinds of scales and different kinds of measurement are derived if numerals can be assigned under different rules, and categorized the quantity scales based on the operation rule of the assignment. 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 operation rule is "*determination of equality of intervals or differences*", and 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 operation rule is "*determination of equality of ratios*", and its admissible unit conversion follows "*Similarity group*: $x' = kx$". Examples of the absolute scale quantities are dimensionless quantities. It follows the rule of "*determination of equality of absolute value*", and "*Identity group*: $x' = x$". Here, we should note that the scale-type is different from the dimension. For instance, we do not know what the force (ratio) divided by the acceleration (ratio) means within the knowledge of scale-types.

In the following sections, the details of the algorithm of SDS are explained. For clarification purpose, we first focus on the case where the model involves only ratio and absolute scales in the next section. The extension to interval scale is described in the latter section. SDS can handle all of the three scale-types.

Given a set of ratio scale quantities, RQ, and a set of absolute scale quantities, AQ,

(1-1) Apply bi-variate test for an admissible equation of ratio scale to every pair of quantities in RQ. Store the resultant bi-variate equations accepted by the tests into an equation set RE and the others not accepted into an equation set NRE.

(1-2) Apply triplet test to every triplet of associated bi-variate equations in RE. Derive all maximal convex sets for the accepted triplets, and compose all bi-variate equations into a multi-variate equation in each maximal convex set. Define each multi-variate equation as a term. Replace the merged quantities by the generated terms in RQ.

(2) Let AQ = AQ + RQ. Given candidate formulae set CE, repeat steps (2-1) and (2-2) until no more new term become generated.

(2-1) Apply bi-variate test of a formula in CE to every pair of the terms in AQ, and store them to AE. Merge every group of terms into a unique term respectively based on the result of the bi-variate test, if this is possible. Replace the merged terms with the generated terms of multi-variate equations in AQ.

(2-2) Apply identity constraints test to every bi-variate equation in AE. Merge every group of terms into a unique term respectively based on the result of the identity constraints test, if they are possible. Replace the merged terms with the generated terms of multi-variate equations in AQ. Go back to step (2-1).

The candidate models of the objective system are derived by composing the terms in AQ.

Figure 1: Outline of SDS algorithm

3 Equation Search Based on Ratio Scale

3.1 Bi-Variate Test

The algorithm of SDS is outlined in Figure 1. Step $(1-1)$ significantly reduces the search space of bi-variate equations by using the "*scale-type constraint.*" Two well-known theorems in the dimensional analysis provides the basis of this step [Buckingham, 1914].

Buckingham Π-theorem *If $\phi(x, y, ...) = 0$ is a complete equation, and if all of its arguments are either ratio or absolute scale-types, then the solution can be written in the form*

$$F(\Pi_1, \Pi_2, ..., \Pi_{n-r}) = 0,$$

where n is the number of arguments of ϕ, and r is the basic number of bases in $x, y, z.....$ For all i, Π_i is an absolute scale-type quantity.

Bases are such basic scaling quantities independent of the other bases in the given ϕ, for instance, as length $[L]$, mass $[M]$ and time $[T]$ of physical dimension. The relation of each Π_i to the arguments of ϕ is given by the following theorem [Bridgman, 1922].

Product Theorem *Assuming primary quantities, $x, y, z, ...$ are ratio scale-type, the function ρ relating a secondary quantity Π to $x, y, z, ...$ has the form:*

$$\Pi = \rho(x, y, z, ...) = \Gamma x^{\alpha} y^{\beta} z^{\gamma} ...,$$

where $\Gamma, \alpha, \beta, \gamma, ...$ are constants.

These theorems state that any meaningful complete equation consisting only of the arguments of ratio and absolute scale-types can be decomposed into an equation of absolute scale-type quantities having an arbitrary form and equations of ratio scale-type quantities having products form. The former $F(\Pi_1, \Pi_2, ..., \Pi_{n-r}) = 0$ is called an "*ensemble*" and the latter $\Pi = \rho(x, y, z, ...) = \Gamma x^{\alpha} y^{\beta} z^{\gamma} ...$ "*regime*"s.

Because we know that any pair of ratio scale quantities in a given complete equation has a product relation if both belong to an identical regime, SDS searches bivariate relations having the following product form in RQ, which is the unique admissible equation that hold in such a regime.

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

The value of the constant a must be independent of any other quantities according to Product Theorem, while the constant b is dependent on the other quantities in the regime. SDS applies the least square fitting of Eq. 1 to the bi-variate experimental data of x and y that are measured while holding the other quantities constant, and determines the values of a, its expected standard error da, and b. For ease of linear fitting, the logarithmic form of Eq.1, $a \log x + \log y = \log b$, is used instead of Eq. 1 itself. The judgment is made whether this equation fits the data well enough by the following two types of statistical tests.

(1) F-test of the ratio between variances of regressive component $S_R = (\sigma_{xy}{}^2/\sigma_{yy})^2$ and residual error component $S_e = \sigma_{ee}{}^2$,

(2) test if da is larger than the absolute value of a itself.

The test (1) is to check if the equation accurately fits to the given data in terms of the power (variance) of residual component. The test (2) is to simply check if the value of the constant a is meaningful. When any of the tests fail, x and y are judged not to have the product relation. For identical pair of ratio scale quantities, this procedure is repeated $k = 10$ times to check the independence of the constant a while holding the other quantities at randomly chosen different values. Then the following test is applied to the set of values of a and da to check the independence.

(3) χ^2-test of the ratio between variance of the values of a and the average of da over the k data set.

If all these tests are passed, the pair of x and y is judged to have the admissible product relation. Then the bi-variate equation together with the average of a and da,*i.e.*, $\overline{a}$ and $\overline{da}$ is stored to RE. If any of the tests failed, the bi-variate equation, $\overline{a}$ and $\overline{da}$ are stored to NRE.

The procedure in step $(1-1)$ is now demonstrated by an example of a complex system depicted in Figure 2. 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 17 quantities.

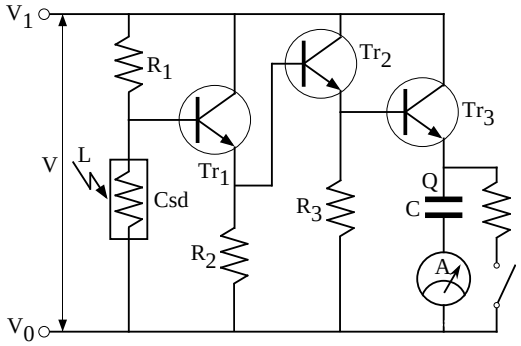

Figure 2: A circuit of photo-meter

$$\left(\frac{R_3 h_{fe_2}}{R_3 h_{fe_2} + h_{ie_2}} \frac{R_2 h_{fe_1}}{R_2 h_{fe_1} + h_{ie_1}} \frac{rL^2}{rL^2 + R_1}\right) V - \frac{Q}{C} - \frac{Kh_{ie_3} X}{Bh_{fe_3}} = 0. \quad (2)$$

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 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 (See Figure 2). Only h_{fe_i}s are absolute scale, and the rest are 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 it is told that 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 prepared 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 RQ as $\{V, L, r, R_1, R_2, R_3, h_{ie_1}, h_{ie_2}, h_{ie_3}, Q, C, X, K, B\}$ and 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 product form among the quantities in RQ, and applied the statistical tests of (1)-(3). The following shows the values of F for F-test, the power constant a and its std. errors da resulted in the bi-variate test for $x = Q$ and $y = X$ under $k = 10$ combinations of different values of the other quantities.

```
 1:  F=25.93  a= 0.6682  da=0.0100
 2:  F=1.986  a= 0.6339  da=0.0346
 3:  F=0.748  a= 0.4840  da=0.1086
 4:  F=27.08  a= 0.6789  da=0.0100
 5:  F=1.421  a= 0.5833  da=0.0640
 6:  F=0.405  a= 0.3902  da=0.1539
 7:  F=0.860  a= 0.2351  da=0.6268
 8:  F=37.09  a= 0.7655  da=0.0100
 9:  F=1.843  a= 0.6226  da=0.0424
10:  F=6.324  a=-0.0494  da=0.0557
```

If $F < 5.317$ then the test (1) fails, and if $da > |a|$ then the test (2) fails. Many iterations failed either one of the tests (1) and (2). For the test (3), $\chi^2 = 39.54$ was obtained where it was larger than the threshold value 16.92. Thus, this test also failed. The resultant RE of the bi-variate equations that were passed the tests was as follows.

$$\begin{aligned} RE \;=\; & \{L^{(1.999\pm0.010)}r = b_1, L^{(-1.999\pm0.010)}R_1 = b_2, \\ & r^{(-1.000\pm0.010)}R_1 = b_3, R_2^{(-1.000\pm0.010)}h_{ie_1} = b_4, \\ & R_3^{(-1.000\pm0.010)}h_{ie_2} = b_5, Q^{(-1.000\pm0.010)}C = b_6, \\ & h_{ie_3}^{(1.000\pm0.010)}X = b_7, h_{ie_3}^{(1.000\pm0.010)}K = b_8, \\ & h_{ie_3}^{(-1.000\pm0.010)}B = b_9, X^{(1.000\pm0.010)}K = b_{10}, \\ & X^{(-0.999\pm0.010)}B = b_{11}, K^{(-1.000\pm0.010)}B = b_{12}\} \end{aligned}$$

All pairwise product forms that should hold among the quantities in RQ have been correctly enumerated.

3.2 Triplet Test

In the next step $(1-2)$, triplet consistency tests are applied to every triplet of equations in RE. Given a triplet of the power form equations in RE:

$$x^{\overline{a}_{xy}}y = b_{xy}, y^{\overline{a}_{yz}}z = b_{yz}, x^{\overline{a}_{xz}}z = b_{xz}, \qquad (3)$$

by substituting y in the first to y in the second, we obtain

$$x^{-\overline{a}_{yz}\overline{a}_{xy}}z = b_{xy}{}^{-\overline{a}_{yz}}b_{yz}.$$

Thus, the following condition must be met.

$$\overline{a}_{xz} = -\overline{a}_{yz}\overline{a}_{xy}. \qquad (4)$$

However, if any of the three equations are not correct due to the noise and error of data fitting, this relation may not hold. The following test judges if the three of the equations are mutually consistent in terms of $\overline{a}$s.

(4) Given $Err_a = \overline{a}_{xz} + \overline{a}_{yz}\overline{a}_{xy}$ and its expectation $Exp_{da} = \{\overline{da}_{xz}^2 + (\overline{da}_{yz}\overline{a}_{xy})^2 + (\overline{a}_{yz}\overline{da}_{xy})^2\}^{1/2}$, perform normal distribution-test of Err_a based on its expectation Exp_{da}.

SDS applies this test to every triplet of equations in RE, and search every maximal convex set MCS where each triplet of equations among the quantities in this set has passed the test (4). In addition, every pair of quantities in an equation in RE which does not belong to any triplet such as Eq.3 is also regarded as a tiny MCS, because the equation may be a regime. When actual regimes in the objective system are mutually independent, each MCS will correspond to a regime. However, an MCS may be different from the set of quantities in a real regime stated in Buckingham Π-theorem in the following cases.

(A) Product of two regimes in an ensemble
If two real regimes $\Pi_1 = {x_1}^{a_{x1}}{y_1}^{a_{y1}}...$ and $\Pi_2 = {x_2}^{a_{x2}}{y_2}^{a_{y2}}...$ have a relation of product in their ensemble as $F({\Pi_1}^{a_{\Pi_1}}{\Pi_2}^{a_{\Pi_2}}, ..., \Pi_{n-r}) = 0$, then MCS will be a superset of the quantities of the two real regimes.

(B) Common terms between two regimes
If two real regimes $\Pi_1 = {x_1}^{a_{x1}}{y_1}^{a_{y1}}...S^{a_{s1}}T^{a_{t1}}...$ and $\Pi_2 = {x_2}^{a_{x2}}{y_2}^{a_{y2}}...S^{a_{s2}}T^{a_{t2}}...$ share some common terms $S, T, ...$, then the partition of the set of quantities in each regime $\{x_1, y_1, ...\}, \{x_2, y_2, ...\}, \{p|p \in S\}, \{q|q \in T\}, ...$ will become MCSs.

In case of (B), $S^{a_{s1}}T^{a_{t1}}$ and $S^{a_{s2}}T^{a_{t2}}$ can be p^2q and pq^2 respectively for instance, where $S \equiv p$ and $T \equiv q$. Then $\{p\}$ and $\{q\}$ are MCSs. In another case, if $S \equiv p_1p_2p_3$ and $T \equiv q_1{q_2}^2$, then $\{p_1, p_2, p_3\}$ and $\{q_1, q_2\}$ are MCSs. These facts also hold for more than two regimes. These consideration indicates that every MCS does not have any intersection with others in any case. If any MCSs mutually sharing some quantities are obtained, those MCSs may not be valid due to the noise and error of the data fitting in step (1-1). It means some pairwise product forms among the elements of those MCSs have been missed in the tests. Accordingly, the following test and operation are applied to the resulted MCSs.

(5) Given a set of MCSs $S = \{M_1, M_2, ...\}$ where each M_i shares some quantities with the other elements in S, obtain the merged MCSs,*i.e.*, $M_S = \cup_{M_i \in S}M_i$, if $p \leq p_{th}$, by assuming that the M_is in S have been obtained because of missing p pairwise product forms among the elements in M_S. Then move the p pairwise product forms from NRE to RE.

The valid number p is always given by the following expression.

$$p = f(M_s) + \Sigma_{A \in 2^S}(-1)^{|A|}f(\cap_{M_i \in A}M_i), \qquad (5)$$

where $f(M) = \frac{|M|(|M|-1)}{2}$ is the number of the pairwise links in a set M, 2^S is the power set of S, and $|A|$ is the cardinality of A. p_{th} is empirically set to be 3 in SDS. The calculation of Eq.5 is limited to $|S| \leq 3$, because p always exceeds 3 for $|S| > 3$. For example, when $S = \{\{x_1, x_2, y_1\}, \{x_1, x_2, y_2\}, \{x_1, x_2, y_3\}\}$ has been derived, we once assume $M_s = \{x_1, x_2, y_1, y_2, y_3\}$. The number of missing pairwise product forms is calculated as

$$\begin{aligned} p \;=\; & f(\{x_1, x_2, y_1, y_2, y_3\}) - f(\{x_1, x_2, y_1\}) \\ & -f(\{x_1, x_2, y_2\}) - f(\{x_1, x_2, y_3\}) + f(\{x_1, x_2\}) \\ & +f(\{x_1, x_2\}) + f(\{x_1, x_2\}) - f(\{x_1, x_2\}) = 3, \end{aligned}$$

where p is equal to p_{th}. Thus, three pairs of quantities in M_s, $\{y_1, y_2\}, \{y_2, y_3\}$ and $\{y_3, y_1\}$, which do not belong to any of $\{x_1, x_2, y_1\}, \{x_1, x_2, y_2\}$ and $\{x_1, x_2, y_3\}$, are moved from NRE to RE. Once all MCSs are found, the data-driven regimes are given by the following form.

$$\Pi_i = \prod_{x_j \in MCS_i} {x_j}^{a_j}. \qquad (6)$$

a_js and their std. errors da_js are evaluated by the average of $\overline{a}$ and $\overline{da}$ of the equations in RE. Before the final value of a_j is determined, the following test is applied.

(6) normal distribution-test to check if a_j is close to an integer under the error da_j. If a_j is judged to be an integer, it is set to the integer value.

This test is based on the observation that the majority of the first principle based equations have integer power coefficients. The product form given by Eq.6 is named as a *pseudo-regime* to distinguish it from the real regime. As we see, the Πs given by pseudo-regimes are not guaranteed to be dimensionless (absolute scale), and also the pseudo-regimes do not share any quantities mutually, even when the original regimes share some quantities. Finally, the merged quantities are replaced by the term of each equation of the derived pseudo-regime in RQ.

In the example in Figure 2, after performing the triplet test of (4) for the RE, the resultant MCSs did not mutually share any quantities, and thus, they are combined in the form of Eq.6 skipping the test (5). Subsequently, their power coefficients were evaluated by the test (6), and they were known to be integer values. The final forms of pseudo-regimes replaced the merged terms in RQ in this step as follows.
$RQ = \{\Pi_1 = R_1 r^{-1.0} L^{-2.0}, \Pi_2 = h_{ie_1} R_2^{-1.0}, \Pi_3 = h_{ie_2} R_3^{-1.0}, \Pi_4 = h_{ie_3} XKB^{-1.0}, \Pi_5 = QC^{-1.0}, \Pi_6 = V\}$

4 Searching Ensemble Equations

4.1 Generation of Terms based on Bi-Variate Test

Once all pseudo-regimes are identified, new terms are generated in step (2-1) by merging these pseudo-regimes in preparation to compose the ensemble equation. First, RQ is added to AQ. Subsequently, SDS searches bivariate relations having one of the formulae specified in the equation set CE. The repertoire in CE governs the ability of the equation formulae search in SDS. Currently, only the following two simple formulae are given in CE. Nevertheless, SDS performs very well in search for the ensemble equation.

$$x^a y = b, \text{ (product form)} \quad (7)$$

$$ax + y = b, \text{ (linear form)} \quad (8)$$

First, SDS adopts the least square fitting of Eq.7 as in step(1-1). Then, the statistical tests (1) and (2) mentioned earlier are applied. This process is repeated $k = 10$ times for randomly chosen different combinations of the values for the other quantities in AQ. If all these tests are passed, the bi-variate equation is stored to AE, and the test (3) is conducted to check the independence of a. Note that this test is not used to reject the relation here because x and y may be absolute scale, and thus a can depend on the other quantities in AQ. SDS marks the relation having the independent a in AE. After all pairwise relations in AQ are examined, SDS searches every maximal convex set MCS as in step(1-2) for the relations marked as the independent a, and the quantities in an MCS are merged into the following term.

$$\Theta_i = \prod_{x_j \in MCS_i} {x_j}^{a_j}. \quad (9)$$

Similar procedure is applied to Eq.8, in which case the merged term of an MCS is:

$$\Theta_i = \sum_{x_j \in MCS_i} a_j x_j. \quad (10)$$

This procedure is repeated in couple for both Eqs.7 and 8 until no new term becomes possible. If all terms in AQ is merged into one, the equation of the final term is the ensemble equation.

In the example of the circuit, Eq.7 was applied first, and three MCSs were found. 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, Eq.8 was tested, then one MCS was found.

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

AQ became as $\{\Theta_1, \Theta_2, \Theta_3, \Theta_4, \Pi_6, \}$. Again, by applying Eq.7, another MCS was newly generated.

$$\Theta_5 = \Pi_6 \Theta_4^{-1.0} = V(h_{ie_3} XKB^{-1.0} + QC^{-1.0})^{-1.0}$$

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

4.2 Generation of Terms based on Identity Constraints

In step (2-2), the identity constraints are applied for further merging 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 identity equation is obtained by solving each for Θ_h.

$$\Theta_h \equiv -\frac{\Theta_i}{a(\Theta_j)} + \frac{b(\Theta_j)}{a(\Theta_j)} \equiv -\frac{\Theta_j}{a(\Theta_i)} + \frac{b(\Theta_i)}{a(\Theta_i)}$$

Because the third expression is linear with Θ_j for any Θ_i, the second must be so. Accordingly, the following must hold.

$$\begin{aligned} 1/a(\Theta_j) &= \alpha_1 \Theta_j + \beta_1, \\ b(\Theta_j)/a(\Theta_j) &= -\alpha_2 \Theta_j - \beta_2. \end{aligned}$$

By substituting these to the second expression,

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

is obtained. This principle is generalized to various relations among multiple terms. Table 1 shows such relations for multiple linear relations and multiple product relations. SDS checks every bi-variate equation in AE derived in step (2-1). If a bi-variate linear equation has a that depends on other terms, it is stored in a set L, and if a bi-variate product relation has such a, it is stored in P. Then the bi-variate least square fitting of the general relations indicated in Table 1 is applied to AQ. For every bi-variate fitting and their coefficients, the test (1), (2) and (3) are also conducted. If all the

Table 1: Identity constraints

bi-variate relation	general relation
$ax + y = b$	$\sum_{(A_i \in 2^{LQ}) \& (p \not\subseteq A_i \forall p \in L)} {}^{a_i} \prod_{x_j \in A_i} {}^{x_j} = 0$
$x^a y = b$	$\prod_{(A_i \in 2^{PQ}) \& (p \not\subseteq A_i \forall p \in P)} {}^{\exp(a_i \prod_{x_j \in A_i} {}^{\log x_j}) = 0}$

L is a set of pairwise terms having a bi-variate linear relation and $LQ = \cup_{p \in L} p$. P is a set of pairwise terms having a bi-variate product relation and $PQ = \cup_{p \in P} p$.

coefficients except one are independent in a relation, the relation is solved for the unique dependent coefficient, and the coefficient is set to be the merged term of the relation. If all coefficients are independent in a relation, the relation is the ensemble equation. If such ensemble equation is not found, SDS goes back to the step (2-1) for further search.

In the example of the circuit, SDS found a set of the bi-variate linear relations in AE. These were on the combinations of $\{\Theta_1, \Theta_5\}, \{\Theta_2, \Theta_5\}$ and $\{\Theta_3, \Theta_5\}$. By applying the bi-variate fitting of the general linear equation in Table 1, 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.2 is easily checked by substituting the intermediate terms to this ensemble equation.

5 Equation Search Based on Interval Scale

The conventional Buckingham Π-theorem and Product Theorem do not consider the equation involving interval scale quantities. We have extended these theorems to include interval scales[Washio and Motoda, 1997].

Extended Buckingham Π-theorem *If* $\phi(x_1, x_2, x_3$ $....) = 0$ *is a complete equation, and if each argument is one of interval, ratio and absolute scale-types, then the solution can be written in the form*

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

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

Extended Product Theorem *Assuming primary quantities in a set* R *are ratio scale-type, and those in another set* I *are interval scale-type, the function* ρ *relating a secondary quantity* Π *to* $x_i \in R \cup I$ *has the forms:*

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

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

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 ensemble having an arbitrary form and regimes of interval and ratio scale-type quantities in products and logarithmic form. In each regime, every interval scale-type quantities appears in linear relation with some other interval scale-type quantities. Therefore, specific tasks in the equation search associated with interval scale quantities are to seek linear forms among interval scale-type quantities and to seek the logarithmic relation between a linear form and the others. For these tasks, the steps indicated in Figure 3 are inserted in the original algorithm of SDS.

Additionally given a set of interval scale quantities, IQ*,*

(0-1) Apply bi-variate test for an admissible linear equation of interval scale to every pair of quantities in IQ*. Store the resultant bi-variate equations accepted by the tests into an equation set* IE *and the others not accepted into an equation set* NIE*.*

(0-2) Apply triplet test to every triplet of associated bi-variate equations in IE*. Derive all maximal convex sets* MCS*s for the accepted triplets, and compose all bi-variate equations into a multi-variate equation in each* MCS*. Define each multi-variate equation as a term. Replace the merged terms by the generated terms of the multi-variate equations in* IQ*. Let* $RQ = RQ + IQ$*.*

(1-3) Apply bi-variate test for an admissible logarithmic equation between the linear forms of interval scale-type quantities and the other terms in RQ*. Replace the terms in the resultant bi-variate equations accepted in the tests by the generated terms in* RQ*.*

Figure 3: Extended part of algorithm

The step (0-1) and (0-2) are almost identical with the steps (1-1) and (1-2) except that the following admissible relation is used at the bi-variate data fitting in IQ.

$$ax + y = b \quad (11)$$

Once a multi-variate linear form is obtained after the triplet test, the form is dealt with a term in the regime formulae based on the extended Product Theorem, and the term is stored into RQ by IQ. In step (1-3), the

following bi-variate logarithmic relations are sought between the linear forms of interval scale-type y and the other terms x in RQ.

$$a \log x + y \quad = \quad b \tag{12}$$

The triplet test is not applied at this step because Eq.12 is asymetric and essentially a bi-variate relation.

In case of the aforementioned example, the circuit does not involve any interval scale-type quantities. However, if we look the electric voltage not to be a voltage difference V but two voltage levels V_0 and V_1, they become interval scale-type. Hence, the system is represented by the following 18 quantities.

$$\begin{aligned} IQ &= \{V_0, V_1\}, \\ RQ &= \{L, r, R_1, R_2, R_3, h_{ie_1}, h_{ie_2}, h_{ie_3}, Q, C, X, K, B\}, \\ AQ &= \{h_{fe_1}, h_{fe_2}, h_{fe_3}\}. \end{aligned}$$

SDS applied the step(0) to the experimental data, and figured out a term $\Theta_0 = V_1 - V_0$ quickly. The rest of the reasoning was identical with the description in the previous sections.

6 Discussion and Related Work

Main features of the discovery system SDS are its low complexity, robustness, scalability and wide applicability. The basic algorithm of SDS consists of two types of procedures. One is the bi-variate test for each pair of quantities and terms in steps (0-1), (1-1), (1-3) and (2-1). The complexity of this type of procedure is $O(n^2mk)$ where n, m, k are the number of quantities to represent the objective system, the number of experimental data used for a data fitting and the number of iteration of the data fitting in a bi-variate test, respectively. Another is the triplet test for each triplet of quantities and terms in steps (0-2), (1-2) and (2-2), where its complexity is $O(n^3)$. m and k usually do not affect the performance of SDS as they are almost independent of the complexity of objective system structure. Moreover, the computational cost required in the bi-variate test is much larger than the triplet test because the former involves multiple experiments, data sampling, data fitting and some statistical tests, whereas the latter involves the triplet consistency checking among the given coefficients only. Thus, the practical complexity is almost proportional to the second order of n. Table 2 shows the performance of SDS to discover various physical law equations. The relative CPU time of SDS normalized by the first case shows that its complexity is nearly proportional to n^2. For reference, the relative CPU time of ABACUS is indicated for the same cases except for the circuit examples of this paper[Falkenhainer and Michalski, 1985]. Though ABACUS applies various heuristics including the information of dimension, its complexity is still NP-hard. As this feature is shared by BACON family, they can hardly derive the model of the electric circuit of this complexity.

The robustness of SDS against the noisy experimental environment has been also evaluated. The upper limitation of the noise level to obtain the correct result in the cases of more than 80% of 10 trials was investigated for each physical law, and they are indicated in the last column of Table 2. The noise levels shown here are the std. of Gaussian noise relative to the real values of quantities, and were added to both controlled (input) quantities and measured (output) quantities at the same time. Thus actual noise level is higher than these levels. The results show the significant robustness of SDS. SDS can provide appropriate results under any practical noise condition.

Table 2: Statistics on complexity and robustness

Example	n	TC(S)	TC(A)	NL(S)
Ideal Gas	4	1.00	1.00	±40%
Momentum	8	6.14	22.7	±35%
Coulomb	5	1.63	24.7	±35%
Stoke's	5	1.59	16.3	±35%
Kinetic Energy	8	6.19	285.	±30%
Circuit*1	17	21.6	-	±20%
Circuit*2	18	21.9	-	±20%

n: Number of Quantities, TC(S): Total CPU time of SDS, TC(A): Total CPU Time of ABACUS, NL(S): Limitation of Noise Level of SDS, *1: Case that electronic voltage is represented by a ratio scale V, *2: Case that electronic voltage is represented by two interval scale V_0 and V_1.

The low complexity and the high robustness shown here ensure the significant scalability of SDS to engineering problems. Many systems in BACON family adopt generate and test in the search. In contrast, the low complexity of SDS comes from its straightforward algorithm to apply only product and linear forms in polynomial time order in concert with the highly restrictive but domain independent constraints. By adding some more basic functional equations to CE, the search of SDS will become more powerful. The robustness of SDS comes from the bi-variate direct fitting to data and the structure of the triplet test. The systems in BACON family repeat formulae fitting to coefficients resulted from the other fitting if it is necessary. This method accumulates the error of data fitting, and derives erroneous results. On the other hand, SDS uses only the bi-variate and direct fitting to the given data, and efficiently composes the result in statistically accurate manner. The multiple statistical tests provide quite conservative judgment on the selection of equations, which contributes to reducing the ambiguity of reasoning. But it also requires following up of missed equations. This is done by reconstructing MCSs in the triplet test by assuming some missed equations in the derived MCS.

The wide applicability is another advantage of SDS, as it does not require any information on dimensions of quantities. For example, the following equation is known to be the law of spaciousness of a room in psychophysics[Kan*et al.*, 1972].

$$S_p = c \sum_{i=1}^{n} R L_i^{0.3} W_i^{0.3},$$

where S_p, R, L_i and W_i are average spaciousness of a room, room capacity, light intensity and solid angle of window at the location i in the room. Though the dimension of S_p is unclear, its scale-type is known to be ratio scale based on its definition. L and R are ratio scale, and W is absolute scale. We applied SDS to this system for the case of $n = 3$, and easily obtained the above expression. The dimension based approach such as COPER may not be applicable to this case.

The weakness of the approach of SDS is some limits on the class of formulae to be discovered. First, the regimes and ensemble formulae must be *read-once formulae*, where each quantity appears at most once in it. Second, the relations among quantities must be *arithmetic*, where the operators are limited to addition, subtraction, multiplication, division, exponentiation and logarithm because of the limited contents of CE. Third, the formula of every pair of quantities searched in the bivariate test is limited to the relation of a simple *binary operator*. These restrictions should be relaxed, even though the majorities of the first principle formulae fall into this class. Bshouty *et al.* proposed an approach to find three unary arithmetic functions $g(x)$, $h(y)$ and $f(\cdot)$ related by a binary arithmetic operator, e.g., $f(g(x) + h(y))$ for a given arithmetic relation $F(x, y)$. It is based on an invariance principle of this structure under the linear conversion of $g(x)$ and $h(y)$[Bshouty, 1994]. Their approach may not be very adequate for the data-driven discovery, because it assumes an initially given precise relation of $F(x, y)$ and its derivatives. However, this invariance principle on the binary relation has a possibility to provide an efficient remedy to the third limitation. The second limitation can be relaxed by increasing the variety of the contents of CE. The first is also a challenging issue, and some invariance or identity principle can be used for the relaxation. All of these issues are left for the future work.

7 Conclusion

SDS implements newly introduced constraints of scale-type and identity in the algorithm of bi-variate and triplet equation test. This architecture has shown to have low complexity, high robustness, promising scalability and wide applicability. It is true that the most of the scientific discoveries have been made through a large number of experiments and observations. However, the scientists have not solely relied on the data but some admissible conditions such as invariance of light speed, symmetry for time inverse and continuity of relations. The constraints of scale-type and identity are two of such conditions having wide applicability. Our future plan is to extend this work to further larger systems and also to seek new laws in non-physical domains.

References

[Langley *et al.*, 1985] P.W. Langley, H.A. Simon, G. Bradshaw and J.M. Zytkow. *Scientific Discovery; Computational Explorations of the Creative Process.* MIT Press, Cambridge, Massachusetts, 1987.

[Koehn and Zytkow, 1986] B. Koehn and J.M. Zytkow. Experimeting and theorizing in theory formation. *Proceedings of the International Symposium on Methodologies for Intelligent Systems*, pages 296–307, 1986. ACM SIGART Press.

[Falkenhainer and Michalski, 1985] B.C. Falkenhainer and R.S. Michalski. Integrating Quantitative and Qualitative Discovery: The ABACUS System. In *Machine Learning*, pages 367–401, Boston, 1986. Kluwer Academic Publishers.

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

[Wu and Wang, 1989] Y. Wu and S. Wang. Discovering Knowledge from Observational Data. In: Piatetsky-Shapiro, G. ed. *Knowledge Discovery in Database, IJCAI-89 Workshop Proceedings*, pages 369–377, 1989. Detroit, MI.

[Schaffer, 1990] C. Schaffer. A Proven Domain-Independent Scientific Function-Finding Algorithm. *Proceedings Eighth National Conference on Artificial Intelligence*, pages 828–833, 1990. AAAI Press/The MIT Press.

[Huang and Zytkow, 1996] K.M. Huang and J.M. Zytkow. Robotic discovery: the dilemmas of empirical equations.

[Kokar, 1985] M.M. Kokar. Determining Arguments of Invariant Functional Descriptions. In *Machine Learning*, pages 403–422, Boston, 1986. Kluwer Academic Publishers.

[Buckingham, 1914] E. Buckingham. On physically similar systems; Illustrations of the use of dimensional equations. In *Physical Review*, Vol.IV, No.4, pages 345–376, 1914.

[Ljung, 1987] L. Ljung. In *System Identification Theory for the User*, 1987. PTR Prentice Hall, Englewood Cliffs, New Jersey.

[Stevens, 1946] S.S. Stevens. On the Theory of Scales of Measurement. In *Science*, pages 677–680, 1946.

[Bridgman, 1922] P.W. Bridgman. In *Dimensional Analysis*, 1922. Yale University Press, New Haven, CT.

[Kan*et al.*, 1972] M. Kan., N. Miyata and K. Watanabe. Research on Spaciousness. In *Japanese Journal of Architechture*, No.193, pages 51–57, 1972.

[Washio and Motoda, 1997] T. Washio and H. Motoda. Discovery of First Principle Based on Data-Driven Reasoning. *Proc. of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 169–182, 1997. Singapore.

[Bshouty, 1994] D. Bshouty and N.H. Bshouty. On Learning Arithmetic Read-Once Formulas with exponentiation. *Proc. of the Seventh Annual ACM Conference on Computational Learning Theory*, pages 311–317, 1994. New Brunswick, NJ.