# Discovering Admissible Simultaneous Equations of Large Scale Systems

**Takashi Washio and Hiroshi Motoda**

Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka,
Ibaraki, Osaka, 567, Japan
washio@sanken.osaka-u.ac.jp

## Abstract

SSF is a system to discover the structure of simultaneous equations governing an objective process through experiments. SSF combined with another system SDS to discover a quantitative formula of a complete equation derives the quantitative model consisting of simultaneous equations reflecting the first principles underlying in the objective process. The power of SSF comes from the use of the complete subset structure in a set of simultaneous equations which can be experimentally identified. The theoretical foundations of the structure identification and the algorithm of SSF are described, and its efficiency and practicality are demonstrated and discussed with large scale working examples. This work is to promote the research of scientific discovery to a novel and promising direction, since the conventional equation discovery systems could not handle such a simultaneous equation process.

## Introduction

A challenging task to find regularities in the data is discovering quantitative formulae of scientific laws from experimental measurements. Langley and others' BACON systems (P.W. Langley & Zytkow 1987) are the most well known as a pioneering work. They founded the succeeding BACON family. FAHRENHEIT (Koehn & Zytkow 1986), ABACUS (Falkenhainer & Michalski 1986) and IDS (Nordhausen & Langley 1990) and etc. are such successors that use basically similar algorithms to BACON in searching for a complete equation governing the measured data in a continuous process. However, one of the drawbacks of the BACON family is their complexity in the search of equation formulae. Another drawback is the considerable amount of ambiguity in their results for noisy data even for the relations among small number of quantities (Schaffer 1990; Huang & Zytkow 1996).

To alleviate these difficulties, some systems, e.g. ABACUS and COPER (Kokar 1986), utilize the information of the unit dimension of quantities to prune the meaningless terms. However, their applicability is limited only to the case where the quantity dimension is known. SDS is a quantitative model discovery system developed based on some novel principles(Washio & Motoda 1997). It utilizes the constraints of *scale-type* and *identity* both of which highly constrain the generation of candidate terms. Since the knowledge of scale-types is widely obtained in various domains, SDS is applicable to non-physics domains including psychophysics, sociology and etc. In addition, an extra strong mathematical constraint named *triplet checking* is introduced to check the validity of those bi-variate equations. By these constraints, the complexity of the algorithm remains quite low, and the high robustness against the noise in the measurements is provided.

In spite of these efforts, many of the practical and large scale processes have not been covered yet. This is because such processes consist of multiple mechanisms, and are represented by multiple equations in terms of given quantities. Some past studies have partially addressed this issue. The aforementioned FAHRENHEIT and ABACUS identify each operation mode of the objective process and transition conditions among those modes, and they derives an equation to represent each mode. For example, they can discover state equations of water for solid, liquid and gas phases respectively from experimental data. However, many processes such as large scale electric circuits are represented by simultaneous equations. The model representation in form of simultaneous equations is essential to grasp the dependency structure among the multiple mechanisms in the processes (Iwasaki & Simon 1986; Murota 1987). An effort to develop a system called LAGRANGE has been made to automatically discover dynamical models represented by simultaneous equations(Dzeroski & Todorovski 1994). It enumerates candidate models of an objective process based on a set of observations by using inductive logic programming technique. However, many redundant representations of the process are derived in high computational complexity, while the soundness of the solutions is not guaranteed.

The primary objective of this study is to establish a

method to discover admissible simultaneous equations governing a large scale process while maintaining the advantage of the recent scientific discovery approaches such as SDS. We set two assumptions on the feature of the objective process to be analyzed. One is that the objective process can be represented by a set of quantitative, continuous, complete and under-constrained simultaneous equations for the quantity ranges of our interest. Another is that all of the quantities in every equation can be measured, and all of the quantities except one dependent quantity can be controlled in every equation to their arbitrary values in the range under experiments while satisfying the constraints of the other equations. These assumptions are common in the past BACON family except the features associated with the simultaneous equations. The following studies have been conducted under these assumptions.

(1) Characterization of under-constrained simultaneous equations in terms of invariant structure of dependency among quantities.

(2) Algorithm to derive the invariant structure through experiments to control the objective process.

(3) Principle and algorithm to apply conventional scientific discovery approaches to separately derive each equation.

(4) Performance evaluation and demonstration of our proposing framework through various examples.

Based on the algorithm and the theory obtained in the studies from (1) to (3), we developed a tool program named "*Simultaneous Structure Finder (SSF)*". In the evaluation and demonstration of the study (4), SSF is combined with the aforementioned SDS, since SDS has an excellent feature to discover each complex equation appearing in large scale processes.

## Basic Principle to Discover Simultaneous Equations

### Insights Through an Example

First, we show an analysis of a simple process represented by under-constrained simultaneous equations to provide some important insights on the basic principle proposed in this research. Figure 1 depicts an electric circuit consisting of two parallel resistances and a battery. It can be modeled by the following equations.

$$V_1 = I_1R_1\ [1], V_2 = I_2R_2\ [2],$$
$$V_e = V_1\ [3] \text{ and } V_e = V_2\ [4], \quad (1)$$

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.

Another model of this circuit can be given.

$$I_1R_1 = I_2R_2\ [1], V_2 = I_2R_2\ [2],$$
$$V_e = V_1\ [3] \text{ and } V_e = V_2\ [4]. \quad (2)$$

Both representations give correct behaviors of the circuit. However, the former seems more natural and

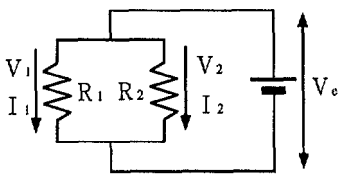


Figure 1: An circuit of parallel resistances.

comprehensive than the latter in spite of their quantitative equivalence. This may be due to the different configuration of quantities in each equation system. The configuration of the quantities in a set of simultaneous equations is represented by an "*incidence matrix*" $T$ where its rows correspond to the mutually independent equations and its columns to the quantities. If the $j$-th quantity appears in the $i$-th equation, then the $(i, j)$ element of $T$, i.e., $T_{ij}$, is 1, and otherwise $T_{ij}$ is 0 (Murota 1987). The following two expressions represent the incidence matrices $T_1$ for Eqs.(1) and $T_2$ for Eqs.(2) respectively.

$$T_1 = \begin{array}{c} \begin{matrix} V_e & V_1 & V_2 & I_1 & I_2 & R_1 & R_2 \end{matrix} \\ \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \quad (3)$$

$$T_2 = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (4)$$

More strictly speaking, when a subset consisting of $n$ independent equations containing $n$ undetermined quantities are obtained by exogenously specifying the values of some extra quantities in the under-constrained simultaneous equations, the values of those $n$ quantities are determined by solving the equations in the subset. In terms of an incidence matrix, exogenous specification of a quantity value corresponds to eliminating the column of the quantity. Accordingly, the partial solvability of the under-constrained simultaneous equations can be restated as when each of $n$ columns come to contain nonzero element in some of $n$ rows by eliminating some extra columns in an incidence matrix, the quantities corresponding to the $n$ columns are determined. Such an equation subset corresponding to the $n$ rows is called a "*complete subset*" of order $n$ in this paper. In the former model of the electric circuit, if we exogenously specify the values of $V_e$ and $R_1$, the first, the third and the forth rows of $T_1$ come to contain the three nonzero columns of $V_1, V_2$ and $I_1$. Thus these equations form a complete subset of order 3, and the three quantities are determined while the others, $I_2$ and $R_2$, are not. On the other hand, if the identical specification on $V_e$ and $R_1$ is made in the latter model, no complete subset of order 3 is obtained, since every combination of three rows in $T_2$ contains

more than three nonzero columns. In the real electric circuit, the validity of the consequence derived by the former model is clear. In fact, the former model gives correct answers for any combinations of quantities exogenously specified, while the latter becomes erroneous in some cases. The model having the incidence matrix which always derives valid interpretations of the determination of quantities of an objective process is named "*structural form*" in this paper.

## Characterizing Simultaneous Equations

Now, more strict formalization and characterization of under-constrained simultaneous equation processes are given. For the basis of the formalization, some fundamental definitions are introduced first.

**Definition 1 (incidence matrix)** *Given a set of mutually independent simultaneous equations which is a model of an objective process, $E = \{eq_i | i = 1, ...M\}$, containing a set of quantities, $Q = \{q_j | j = 1, ..., N\}$, a matrix $T$ is called an "incidence matrix" for $E$ and $Q$, where $T_{ij} = 1$ if $q_j (\in Q)$ appears in $eq_i (\in E)$, otherwise $T_{ij} = 0$. Here, $T_{ij}$ is an $(i, j)$ element of $T$.*

**Definition 2 (complete subset)** *Given an incidence matrix $T$, after applying elimination of a set of columns, $RQ (\subset Q)$, let a set of nonzero columns of $T[CE, Q - RQ]$ be $NQ (\subseteq Q - RQ)$, where $CE \subseteq E$, and $T[CE, Q - RQ]$ is a sub-incidence matrix for equations in $CE$ and quantities in $Q - RQ$. $CE$ is called a "complete subset" of order $n$, if $|CE| = |NQ| = n$. Here, $| \bullet |$ stands for the cardinality of a set.*

Based on these definitions, some characteristics of a complete subset are derived.

**Theorem 1 (symmetry theorem)** *Given a complete subset $CE$ of order $n$ under the elimination of a set of columns $RQ$ in $T$ where $|RQ| = m$, let a set of nonzero columns in $T[CE, Q]$ be $CQ$, where $T[CE, Q]$ is a sub-incidence matrix for equations in $CE$ and all quantities in $Q$. Under the elimination of any subset $RQ_i$ of $CQ$ where $|RQ_i| = m$ and $i = 1, ..., {}_{(n+m)}C_m$, $CE$ is a complete subset of order $n$.*

*Proof.* Because $NQ$ is a set of nonzero columns of $T[CE, Q - RQ]$, $CQ = RQ + NQ$. $|CQ| = m + n$, since $|RQ| = m$ and $|NQ| = n$ by definition of a complete subset. For the elimination of any $RQ_i (\subset CQ)$ where $|RQ_i| = m$ and $i = 1, ..., {}_{(n+m)}C_m$, the rest of $NQ_i = CQ - RQ_i$ has the cardinality $(m+n) - m$, i.e., $|NQ_i| = n$. Thus, $CE$ is a complete subset of order $n$ for the elimination of any $RQ_i$. ∎

**Theorem 2 (invariance theorem)** *Given a transform $f : U_E \rightarrow U_E$ where $U_E$ is the entire universe of equations. When $CE$ is a complete subset of order $n$ in $T$, $f(CE)$ is also a complete subset of order $n$, if $f(CE)$ for $CE \subset U_E$ maintains the number of equations and the nonzero column structure, i.e., $|CE| = |f(CE)|$ and $CQ = CQ_f$, where $CQ_f$ is a set of nonzero columns in $T[f(CE), Q]$.*

*Proof.* Because of $CQ = CQ_f$, an identical set of nonzero columns $NQ$ is obtained by eliminating $RQ$ in both $T[CE, Q]$ and $T[f(CE), Q]$. When $CE$ is a complete subset of order $n$, $|CE| = n$, and thus if $|CQ| = |CQ_f| = m + n$, then $|NQ|$ can be $n$ by choosing $RQ$ to be $|RQ| = m$. Under such a $RQ$, $|CE| = |f(CE)| = |NQ| = n$, and $f(CE)$ satisfies the condition of a complete subset of order $n$. ∎

**Remark 1** *The "symmetry theorem" indicates that given an objective simultaneous equation process, every complete subset can be identified independent of the choice of quantities to be exogenously controlled in the experiment, as far as it controls the required number m of the quantities involved in each subset. An efficient, complete and sound search algorithm can be developed based on this feature of a complete subset.*

**Remark 2** *Our assumption on the controllability of the quantities admits one dependent quantity which can not be directly controlled in an equation. The "symmetry theorem" is the theoretical basis to correctly derive every complete subset and obtain valid quantitative form of each equation through the experiment. If a dependent quantity exits in a complete subset, the other controllable quantities in the subset can be used to constrain the subset and make up identical states.*

**Remark 3** *Given a model of an objective process, various simultaneous equation formulae maintaining the equivalence of the quantitative relations and the dependency structure among quantities can be derived by limiting the equation transform $f$ of the "invariance theorem" to the quantitative manipulation such as substitution and arithmetic operation among equations.*

In the example of the aforementioned electric circuit, if the value of $V_e$ is exogenously specified in Eqs.(1), i.e., the first column of $T_1$ is eliminated, the third and forth rows of $T_1$ become to involve only two nonzero columns. Consequently, the set of equations $\{V_e = V_1[3], V_e = V_2[4]\}$ in Eqs.(1) is known to be a complete subset of order 2. These equations can be transformed by the linear algebra as follows while keeping their quantitative equivalence.

$$V_e = 2V_1 - V_2\ [3], V_e = -V_1 + 2V_2\ [4]. \tag{5}$$

For this third model, the following incidence matrix is obtained.

$$T_3 = \begin{array}{c} \begin{matrix} V_e & V_1 & V_2 & I_1 & I_2 & R_1 & R_2 \end{matrix} \\ \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \tag{6}$$

By applying the elimination of the first column for $V_e$ similarly to the case of $T_1$, the third and forth rows of $T_3$ become to involve only two nonzero columns, and thus the complete subset of order 2 consisting of the these two rows still remains in this new model.

As the complexity of the algorithm to enumerate all forms of a complete subset admitted by the transform $f$ faces the combinatorial explosion, our approach identifies only one specific form defined bellow.

**Definition 3 (canonical form of a complete subset)** *Given a complete subset $CE$ of order $n$, the "canonical form" of $CE$ is the form where all elements of the nonzero columns $CQ$ in its incidence matrix $T[CE, Q]$ are* 1.

An example of the canonical form of a complete subset is eq.5. Because every admissible form is mathematically equivalent with the others, the identification of the canonical form is sufficient, and the others can be derived by applying appropriate $f$ to the form.

Though each complete subset represents a basic mechanism to determine the values of quantities in a given simultaneous equation process, some complete subsets are not mutually independent in many cases. For instance, the following four complete subsets can be found in the example of Eqs.(1).

$$\begin{aligned} &\{[3],[4]\}(n=2), \quad \{[1],[3],[4]\}(n=3), \\ &\{[2],[3],[4]\}(n=3), \quad \{[1],[2],[3],[4]\}(n=4) \end{aligned} \tag{7}$$

The number in [ ] indicates each equation and $n$ the order of the subset. They mutually have many overlaps, and the complete subsets having higher order represent the redundant mechanism with the lower subsets. The following theorem characterizes the dependency among complete subsets.

**Theorem 3 (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 lattice of the sets, where $CE_i \cup CE_j \in L$ and $CE_i \cap CE_j \in L, \forall CE_i, CE_j \in L$.*

*Proof.* Omitted. ∎

**Theorem 4 (modular lattice theorem)** *Given a model of an objective process consisting of equations $E$, the set of 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., $n(CE_i \cup CE_j) = n(CE_i) + n(CE_j) - n(CE_i \cap CE_j)$ where $n$ is the order of a given complete subset.*

*Proof.* The order of a complete subset is equal to its cardinality by definition. Because of the relation $|CE_i \cup CE_j| = |CE_i| + |CE_j| - |CE_i \cap CE_j|$, the relation among the order in the theorem is clear. ∎

Based on the modular lattice structure among complete subsets, the independent component and its order of each complete subset can be defined as follows.

**Definition 4 (independent component of a complete subset)** *The independent component $DE_i$ of the complete subset $CE_i$ is defined as*

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

*The set of essential quantities $DQ_i$ of $CE_i$ which do not belong to any other complete subsets but involved only in $CE_i$ is also defined as*

$$DQ_i = CQ_i - \bigcup_{\substack{\forall CE_j \subset CE_i \\ \text{and } CE_j \in L}} CQ_j,$$

*where $CQ_i$ is a set of nonzero columns of $T(CE_i)$. The order $\delta n_i$ and the freedom $\delta m_i$ of $DE_i$ are defined as*

$$\delta n_i = |DE_i| \text{ and } \delta m_i = |DQ_i| - |DE_i|.$$

**Remark 4** *An "independent component" of a complete subset represents an independent mechanism to determine the values of some quantities under a given dependency structure among quantities in a set of simultaneous equations. The values of quantities appearing only within an independent component $DE_i$ can be changed with the $\delta m_i$ degree of freedom without violating any other constraints.*

In the example of Eq.(7), the three independent components are derived.

$$\begin{aligned} DE_1 &= \{[3],[4]\} - \phi = \{[3],[4]\}, \\ \delta n_1 &= 2 - 0 = 2, \\ DE_2 &= \{[1],[3],[4]\} - \{[3],[4]\} = \{[1]\}, \\ \delta n_2 &= 3 - 2 = 1, \\ DE_3 &= \{[2],[3],[4]\} - \{[3],[4]\} = \{[2]\}, \\ \delta n_3 &= 3 - 2 = 1. \end{aligned} \tag{8}$$

Because of the monotonic structure of set inclusion in the modular lattice, a bottom up and greedy search is applicable without facing very high complexity of the algorithm to derive every independent component.

Because each independent component $DE_i$ is a subset of the complete subset $CE_i$, the nonzero column structure of $DE_i$ also follows the invariance theorem. Consequently, the subset of the canonical from of $CE_i$ is applicable to represent $DE_i$. Based on this consideration, the definition of the canonical form of the simultaneous equations representing an given objective process is introduced.

**Definition 5 (canonical form of simultaneous equations)** *The "canonical form" of a set 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 $CE_i$, where $b$ is the total number of $DE_i$.*

The incidence matrix of the model of the electric circuit can be derived in the canonical form $T_4$ based on the result of Eq.(8).

$$T_4 = \begin{array}{c} \begin{matrix} V_e & V_1 & V_2 & I_1 & I_2 & R_1 & R_2 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \tag{9}$$

If the canonical form of simultaneous equations are experimentally derived to reflect the actual dependency structure among quantities in the objective process, then the model must be a "*structural form*". Thus, the following terminology is introduced.

**Definition 6 (structural canonical form)** *If the canonical form of simultaneous equations is derived to be a "structural form", then the form is named "structural canonical form".*

The incidence matrix $T_4$ which has been obtained from a structural form $T_1$ corresponds to the structural canonical form for the example.

## Algorithm of SSF and its Implementation

Under our aforementioned assumption on the measurements and the controllability of quantities, the bottom up and greedy algorithm indicated in Fig.2 has been developed and implemented into SSF. SSF requires a list of the quantities for the modeling of the objective process and their actual measurements. Starting from the set of control quantities having small cardinality, this algorithm tests if values of any quantities become to be fully under control. If such controlled quantities are found, the collection of the control quantities and the controlled quantities are considered as a newly found complete subset $|CE_i|$. Then, based on the definition 4 , its $|DE_i|$, $|DQ_i|$, $\delta n_i$ and $\delta m_i$ are derived and stored. Once any independent components are derived, only $\delta m_i$ of the quantities in every $|DQ_i|$ and the quantities which do not belong to any $|DQ_i|$ are used for control. Though the complexity of this algorithm is NP-hard, this constraint by $|DQ_i|$ significantly reduces the computation amount. as shown in the latter section. The constraint of $|DQ_i|$ does not miss any complete subset to search due to the monotonic lattice structure among complete subsets.

The conventional systems to discover a complete equation can not directly accept the knowledge of the structural canonical form for the discovery. The problem to derive quantitative knowledge of the simultaneous equations must be decomposed into sub-problems to derive each equation individually. Accordingly, an algorithm to decompose the entire problem into such small problems is also implemented into SSF. As previously stated in Definition 4 , the values of the quantities within an independent component of each complete subset are mutually constrained in the order $\delta n_i$ degree. Accordingly, the constraints within the independent component disable the bi-variate tests among the quantities of an equation in the structural canonical form, if the order $\delta n_i$ is more than one. However, this difficulty is removed if the $(\delta n_i - 1)$ quantities are eliminated by the substitution of the other $(\delta n_i - 1)$ equations within the independent component. The reduction of the number of quantities by $(\delta n_i - 1)$ in each equation enables to control each quantities as if it is in

*(S1)* *Let $Q = \{q_k | k = 1, ..., N\}$ be a set of quantities to appear in the model of an objective process. Set $X = \{x_k | x_k = q_k$, for all but directly controllable $q_k \in Q\}$, $DE = \phi$, $DQ = \phi$, $N = \phi$, $M = \phi$, $h = 1$ and $i = 1$.*

*(S2)* *Choose $C_j \subset DQ_j \in DQ$ for some $DQ_j$ and also $C_x \subseteq X$, and take their union $C_{hi} = ... \cup C_j \cup ... \cup C_x$, while maintaining $|C_j| \leq \delta m_j$ and $|C_{hi}| = h$. Control all $x_k \in C_{hi}, k = 1, ..., |C_{hi}|$ in an experiment.*

*(S3)* *Let a set of all quantities which values are determined be $D_{hi} \subseteq (Q - C_{hi})$. Set $DE_{hi} = C_{hi} + D_{hi}$, $DQ_{hi} = DE_{hi} - \bigcup_{\substack{\forall DE_{h'i'} \subset DE_{hi} \\ DE_{h'i'} \in DE}} DE_{h'i'}$, $\delta n_{hi} = |D_{hi}| - \sum_{\substack{\forall DE_{h'i'} \subset DE_{hi} \\ DE_{h'i'} \in DE}} \delta n_{h'i'}$, and $\delta m_{hi} = |DQ_{hi}| - \delta n_{hi}$. If $\delta n_{hi} > 0$, then add $DE_{hi}$ to the list DE, $DQ_{hi}$ to the list DQ, $\delta n_{hi}$ to the list N, $\delta m_{hi}$ to the list M and $X = X - DQ_{hi}$.*

*(S4)* *If all quantities are determined, i.e., $D_{hi} = Q - C_{hi}$, then go to (S5), else if any more $C_{hi}$ where $|C_{hi}| = h$ does not exist, $h = h + 1, i = 1$ and go to (S2), else $i = i + 1$ and go to (S2).*

*(S5)* *The contents of the lists DE, DQ and N represent the sets of quantities involved in independent components, the sets of essential quantities and their orders respectively.*

Figure 2: Algorithm for structural canonical form

a complete equation. This elimination of quantities is essential to enable the application of the equation discovery system based on the bi-variate test. The reduction of quantities in equations provides further advantage, since the computation amount required in the equation search strongly depends on the number of quantities. In addition, the less degree of freedom of the objective equation in the search introduces more robustness against the noise in the data and the numerical error in the data fitting. The algorithm for the problem decomposition of SSF minimizes the number of quantities involved in each equation based on the admissible equation transform stated in the invariance theorem. Once quantitative form of each equation is obtained by the equation discovery system, then those forms can be transformed again into the different forms requested by the users.

Figure 3 indicates the algorithm to transform a structural canonical form to minimize the number of quantities in each equation. This algorithm uses the list of the complete subsets and their order resulted in the algorithm of Fig.2. The quantities involved in each equation are eliminated by the equations in the other complete subset in (S2). In the next (S3), the quantities involved in each equation are eliminated by the other equation within the same complete subset, if the order of the subset is more than one. The quantities to eliminate in (S2) and (S3) are selected by lexicographical order in the current SSF. This selection can be more

*(S1) Let DE, DQ and N be the lists obtained in the algorithm of Fig.2.*
*(S2) For $i = 1$ to $|DE|$ {*
*For $j = 1$ to $|DE|$ where $j \neq i$ {*
*If $DE_i \supset DE_j$ where $DE_i, DE_j \in DE$ {*
*$DE_i = DE_i - DQ'_j$,*
*where $DQ'_j$ is arbitrally, and*
*$DQ'_j \subset DQ_j \in DQ$ and $|DQ'_j| = N_j$.}}}*
*(S3) For $i = 1$ to $|DE|$ {*
*For $j = 1$ to $N_i$ {*
*$DE_{ij} = DE_i - DQ_{ij}$,*
*where $DQ_{ij}$ is arbitrally, and*
*$DQ_{ij} \subset DQ_i \in DQ$ and $|DQ_{ij}| = N_i - 1$.}}*
*(S4) Every $DE_{ij}$ shows the list of quantities contained in a transformed equation.*

Figure 3: Algorithm for minimization.

tuned up based on the information of the sensitivity to noise and error of each quantities in the future.

## Outline of SDS

The information required by SDS besides the knowledge given by SSF and the actual measurements is the scale-type of each quantity (Washio & Motoda 1997). The scale-types of measured quantities reflect the rules of the assignment of numerals to objects in the measurement process. The representative scale-types of the quantitative measurements are interval scale, ratio scale and absolute scale. Examples of the interval scale quantities are temperature in Celsius and musical tone where the origins of their scales are not absolute. Examples of the ratio scale quantities are physical mass and absolute temperature where each has an absolute zero point. The absolute scale quantities are dimensionless quantities.

The two important theorems called *"extended Buckingham $\Pi$-theorem"* and *"extended product theorem"* provide the basis of the equation search of SDS. Former states that any meaningful complete equation $\phi(x_1, x_2, x_3, ....) = 0$ consisting of the arguments of interval, ratio and absolute scale-types can be decomposed into an equation $F(\Pi_1, \Pi_2, ..., \Pi_{n-w}) = 0$ called an *"ensemble equation,"* where $n$ is the number of arguments of $\phi$, $w$ is the basic number of bases in $x_1, x_2, x_3....$, respectively. For all $i$, $\Pi_i$ is an absolute scale-type quantity. Latter presents the following multiple formulae named *"regimes"* to represent $\Pi$s by interval and ratio scale-type quantities.

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

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

where $R$ and $I$ are sets of ratio and interval scale type quantities, respectively. *all* coefficients except $\Pi$ are constants and $I_k \cap I_g = \phi$. SDS initially seeks the relations of regimes represented in the extended product theorem through the bi-variate data fitting. Because *"scale-type constraint"* admits only these formulae as valid relations, the relations discovered by this approach have high possibility to represent first principle governing the objective process.

After all regime formulae are derived, SDS starts to seek the relation of ensemble equation by using *"identity constraint"*. The basic principle of the identity constraints comes by answering the question that *"what is the relation among $\Theta_h$, $\Theta_i$ and $\Theta_j$, if $a(\Theta_j)\Theta_h + \Theta_i = b(\Theta_j)$ and $a(\Theta_i)\Theta_h + \Theta_j = b(\Theta_i)$ are known?"* The following answer is easily proven.

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

This principle is generalized to various relations among multiple terms.

Once a triplet of bi-variate relations is identified for a set of three quantities by using the aforementioned constraints, a certain consistency checking among the three relations called *"triplet test"* is applied to remove invalid relations due to the noise and error of data fitting.

The superior abilities of SDS have been confirmed of its low complexity, high robustness, high scalability and wide applicability (Washio & Motoda 1997). This is because of the introduction of these new types of mathematical constraints and tests.

## Evaluation of SSF combined with SDS

The program of SSF has been developed in the environment of a numerical processing shell named MATLAB (Mat 1992). The knowledge of an equation in the structural canonical form discovered by SSF is transferred to SDS, and SDS executes its experiments based on the transferred knowledge and the knowledge of the scale-types of quantities. This process is iterated for each equation. The objective processes are provided by simulation in this research.

The performance of SSF has been evaluated in terms of the validity of its results and the computational complexity through some examples including quite large scale processes. In addition, the performance of the SDS combined with SSF is also checked through the same examples. The examples we applied are summarized as follows.

**(1) Two parallel resistances and a battery**
This is depicted in Fig.1, and has been already explained in the previous sections. Its model consists of 4 equations and 7 quantities as shown in Eqs.1.

**(2) Heat conduction at walls of holes**
Given a large solid material having two vertical holes, gas goes into those holes, and condensed to its liquid phase during the flow in the holes by providing its heat energy to the walls of the holes. In these holes, the heat conduction process are represented by the following 8 equations involving 17 quantities(Kalagnanam,

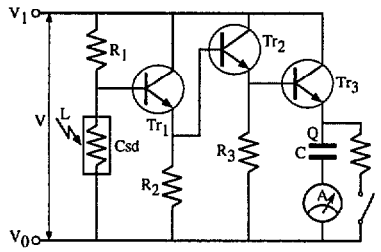


Figure 4: A circuit of photo meter.

Henrion, & Subrahmanian 1994).

$$\omega = 0.9423\left(\frac{vsk^3}{L\mu}\right)^{1/4}, \quad \dot{H} = \dot{H}_1 + \dot{H}_2$$
$$\Delta T_1 = T_f - T_{w1}, \quad \Delta T_2 = T_f - T_{w2}$$
$$h_1 = {\Delta T_1}^{-1/4}\omega, \quad h_2 = {\Delta T_2}^{-1/4}\omega \qquad (10)$$
$$\dot{H}_1 = 2\pi\gamma L h_1 \Delta T_1, \quad \dot{H}_2 = 2\pi\gamma L h_1 \Delta T_2$$

Here, $v$, $s$, $k$, $\mu$ are the latent heat per volume, density, heat conductance, viscosity in liquid phase of the fluid. $L$ is the length of the holes, $T_f$ temperature of the fluid, $T_{w1}$ and $T_{w2}$ temperature of walls of two holes. $\dot{H}$ is the total rate of heat conduction from the fluid to the wall material.

**(3) A circuit of photo-meter**

Figure 4 depicts a circuit of photo-meter to measure the rate of increase of photo intensity within a time period. The model of this system is represented by 14 equations involving 22 quantities. The contents of equations are not shown due to the limited space.

**(4) Reactor core of power plant**

A model of nuclear fission reaction process, heat removal of nuclear fuel, and heat and mass balance of reactor coolant is tested. This model involves 24 equations and 60 quantities.

Table 1 is the summary of the specifications of each problem size, complexity and robustness against noise. $T_{scf}$ is to derive the structural canonical form and $T_{min}$ to minimize the number of quantities appearing in each equation. $T_{tl}$ and $T_{av}$ are the total and the average time per equation required by SDS. $T_{scf}$ shows strong dependency to the parameter $m$ and $n$, i.e., the size of the problem. This is natural, since the algorithm to derive structural canonical forms is NP-hard to the size. $T_{scf}$ also moderately depends on the difference between the numbers of quantities and equations in the model, i.e. $n - m$. This seems reasonable because the large number of $n - m$ represents the high degree of freedom of the objective simultaneous equations. This also exponentially increases the search space. In contrast, $T_{min}$ shows very slight dependency to the size of the problem, and the absolute value of the required time is negligible. This observation is also highly consistent with the theoretical view that its complexity should be

Table 1: Statistics on complexity and robustness

| Ex. | m | n | av | $T_{scf}$ | $T_{min}$ | $T_{tl}$ | $T_{av}$ | $NL$ |
|---|---|---|---|---|---|---|---|---|
| (1) | 4 | 7 | 2.5 | 3 | 0.00 | 206 | 52 | 35 |
| (2) | 8 | 17 | 3.9 | 1035 | 0.05 | 725 | 91 | 29 |
| (3) | 14 | 22 | 2.6 | 1201 | 0.05 | 773 | 55 | 31 |
| (4) | 26 | 60 | 4.0 | 42395 | 0.11 | 3315 | 128 | 26 |

m: number of equation, n: number of quantities, av: average number of quantities/equation, $T_{scf}$: CPU time (sec) to derive structural canonical form, $T_{min}$: CPU time to derive minimum quantities form, $T_{tl}$: CPU time to derive all equations by SDS, $T_{av}$: average CPU time per equation by SDS, $NL$: limitation of % noise level of SDS.

only $O(n^2)$. The total time $T_{tl}$ required by SDS seems not to strongly depend on the size of the problem. This consequence is also very natural, because SDS handles each equations separately. The required time of SDS should be proportional to the number of equations in the model. Instead, the efficiency of the SDS more sensitively depends on the average number of quantities involved in an equation. This tendency becomes clearer by comparing $T_{av}$ with $av$. In the past study, the complexity of SDS is known to be around $O(n^2)$. The relation between $T_{av}$ and $av$ roughly follows this claim. Thus, $T_{tl}$ may vary almost in $O(mn^2)$. In short summary, the complexity of SSF shown in the result of $T_{scf}$ seems to be crucial for a large scale problem, However, the performance shown in the Table 1 may be sufficient for numbers of engineering problems.

The last column of Table 1 shows the influence of the noise to the result of SSF+SDS, where Gaussian noise is artificially introduced to the measurements. The noise does not affect the computation time in principle. The result showed that 25-35% of relative noise amplitude to the absolute value of each quantity was acceptable at the maximum under which 8 times per 10 trials of SSF+SDS successfully give the correct structure and coefficients of all equations with statistically acceptable errors. The noise sensitivity dose not increase significantly, because SSF focuses on a complete subset which is a small part of the entire system. Similar discussion holds for SDS. The robustness of SDS combined with SSF against the noise is sufficient for practical application.

Finally, the validity of the results are checked. In the example (1), SSF derived the expected structural canonical form shown in Eq.9. Then SSF gave the following form of minimum number of quantities to SDS. Here, each equation is represented by a set of quantities involved in the equation.

$$\{V_e, R_1, I_1\},\ \{V_e, R_2, I_2\},\ \{V_e, V_1\},\ \{V_e, V_2\} \qquad (11)$$

As a result, SDS derived the following answer.

$$V_e = I_1 R_1\ [1], V_e = I_2 R_2\ [2],$$
$$V_e = V_1\ [3] \text{ and } V_e = V_2\ [4], \qquad (12)$$

This is equivalent with Eq.1 not only in the sense of the invariance theorem but also the quantitativeness. In the example (2), SSF derived the following structural canonical form.

$$\begin{aligned}&\{\dot{H}, \dot{H}_1, \dot{H}_2\},\ \{\omega, v, s, k, L, \mu\},\\&\{\Delta T_1, T_f, T_{w1}\},\ \{\Delta T_2, T_f, T_{w2}\},\\&\{h_1, \Delta T_1, \omega\},\ \{h_2, \Delta T_2, \omega\},\\&\{\dot{H}_1, \Delta T_1, \gamma, L, h_1, \omega\},\ \{\dot{H}_2, \Delta T_2, \gamma, L, h_2, \omega\}.\end{aligned} \quad (13)$$

Then, by the elimination of $\omega$ in the last two equations by substituting the fifth and the sixth equations, SSF gave the form of minimum number of quantities which configuration is identical with the original. Then, SDS successfully reconstructed the equations in Eqs.10. Similarly almost original equations could be reconstructed in the other examples, and they have been confirmed to be equivalent to the original in the sense of the invariant theorem and quantitativeness.

## Discussion and Related Work

The form of the process models in which the appearance of quantities are minimized resulted by SSF is quite close to the configuration of our familiar models in many cases. This might be because the less connection links among equations through quantities clarify the process represented by each equations, and hence the models obtained by SSF and SDS can provide comprehensive knowledge of the objective processes.

As mentioned in the introduction, the conventional equation discovery systems can derive only one or a few complete equation(s) with high computational complexity. SSF and its background theory, not only to overcome this limitation, provide generic tool and measure which can be combined with any conventional equation discovery systems. Moreover, the background theory can be used in more generic manner to identify various simultaneous structures embedded in real systems. It can be applied to some discrete systems as far as the systems have structures to propagate states through simultaneous constraints.

The basic theory of complete subsets in simultaneous equations can be seen as an extension of the causal ordering theory (Iwasaki & Simon 1986). A complete subset involves many candidates of self-contained subsets. A part of a complete subset becomes a self-contained subset once the exogenous specification of the values of some quantities is given. The structural form introduced in this research is also an extension of structural equations (Iwasaki & Simon 1986). Our theory gives more precise definition and characterization of structural equations.

## Conclusion

The research presented here characterized under-constrained simultaneous equations in terms of complete subsets, and provided an algorithm to derive the structure through experiments. In addition, an algorithm to apply conventional scientific discovery to simultaneous equations are established. These are implemented into a generic tool named SSF, and its significant performance under the combination with a discovery system SDS have been readily confirmed.

A remained but important problem is to establish more efficient algorithm of SSF.

## References


Dzeroski, S., and Todorovski, L. 1994. Discovering Dynamics: From Inductive Logic Programing to Machine Discovery. *Journal of Intelligent Information Systems* 3:1–20.

Falkenhainer, B., and Michalski, R. 1986. Integrating Quantitative and Qualitative Discovery: The ABACUS System. *Machine Learning* 367–401.

Huang, K., and Zytkow, J. 1996. Robotic discovery: the dilemmas of empirical equations. In *Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery.*

Iwasaki, Y., and Simon, H. 1986. Causality in Device Behavior. *Artificial Intelligence* 3–32.

Kalagnanam, J.; Henrion, M.; and Subrahmanian, E. 1994. The Scope of Dimensional Analysis in Qualitative Reasoning. *Computational Intelligence* 10(2):117–133.

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

Kokar, M. 1986. Determining Arguments of Invariant Functional Descriptions. *Machine Learning* 403–422.

The Math Works, Inc. 1992. *MATLAB Reference Guide.*

Murota, K. 1987. Systems Analysis by Graphs and Matroids - Structural Solvability and Controllability. *Algorithms and Combinatorics* 3.

Nordhausen, B., and Langley, P. 1990. An Integrated Approach to Empirical Discovery. In *Computational Models of Scientific Discovery and Theory Formation.* San Mateo, California: Morgan Kaufman Publishers.

P.W. Langley, H.A. Simon, G. B., and Zytkow, J. 1987. *Scientific Discovery; Computational Explorations of the Creative Process.* Cambridge, Massachusetts: MIT Press.

Schaffer, C. 1990. A Proven Domain-Independent Scientific Function-Finding Algorithm. In *Proceedings Eighth National Conference on Artificial Intelligence.* AAAI Press/The MIT Press.

Washio, T., and Motoda, H. 1997. Discovering Admissible Models of Complex Systems Based on Scale-Types and Identity Constraints. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence.*