Discovering Admissible Simultaneous Equation Models from Observed Data

Takashi Washio[1], Hiroshi Motoda[1], and Yuji Niwa[2]

[1] Institute for Scientific and Industrial Research, Osaka University,
8-1, Mihogaoka, Ibarakishi, Osaka, 567-0047, Japan
{washio, motoda}@sanken.osaka-u.ac.jp
[2] Institute of Nuclear Safety System, Inc.,
64 Sada, Mihamacho, Mikatagun, Fukui, 919-1205, Japan
niwa@inss.co.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 paper, SSF is extended to discover the structure of a simultaneous equation model from passively observed data, and is combined with the extended SDS to discover a quantitative simultaneous equation model reflecting the first principle.

1 Introduction

Langley and others' BACON [6] is the most well known pioneering work to discover a complete equation representing scientific laws governing an objective process under experimental observations. FAHRENHEIT [4], 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 to discover the equations representing the first principle underlying the objective process, is reported. To alleviate the drawback, some systems, *e.g.*, ABACUS and COPER [5], utilize the information of the unit dimensions of quantities to prune the meaningless terms. However, many of these conventional scientific equation discovery systems have the following limitations.

(1) The information of the unit dimension of each quantity in the data is needed to discover the first principle based equation.
(2) 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.
(3) 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.

L. De Raedt and P. Flach (Eds.): ECML 2001, LNAI 2167, pp. 539–551, 2001.

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 [10]. 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*" [12] for the application 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 [11]. SSF identifies the number of equations needed to represent the objective process, and further identifies the sets of quantities to appear in the respective equations of the model while excluding 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.

One of the important unexplored studies of the scientific law equation discovery is to propose a practical framework to discover a simultaneous equation model reflecting the first principles from passively observed data. This study tries to address all aforementioned limitations at once. If SSF can be extended to be applicable to the passively observed data, the second and the third limitations in the discovery of the structure of the simultaneous equation model are removed. Once the sets of quantities appearing in respective equations are derived, the aforementioned extended SDS which addresses the first and the second limitations is applicable to figure out the equation formula governing each quantity set. Accordingly, the extension of the applicability of SSF to the passive observations is the main issue in this study. The objectives of this paper are (i) to propose a practical principle to discover the first principle based simultaneous equation structure from passively observed data, (ii) to provide an algorithm of the "*extended SSF*" based on the principle, (iii) to evaluate the basic performance of the combination of the extended SSF and the extended SDS through simulations and (iv) to demonstrate its practicality through a real application.

The main technical contribution of this study is to propose a principle named "*quasi-experiment on dependency*" which checks the dependency among quantities in the passively observed data without performing actual experiment. The quasi-experiment probes the influence propagation from a quantity to some other quantities while virtually fixing the values of some extra quantities by a data sampling technique. The repetitions of this probing figure out the entire dependency structure among quantities in form of a simultaneous equation model. The quasi-experiment on dependency is different from the quasi-bi-variate fitting used in the extended SDS since the latter assumes that the quantities under consideration are governed by a complete equation, and focuses only on the binary relation between every pair of quantities. The approach to combine the extended SSF and the extended SDS requires three assumptions, which are

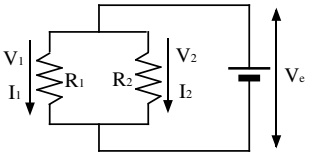

Fig. 1. An circuit of parallel resistances.

allowable in many practical applications. One is that the scale-types of all observed quantities are known. The scale-types of the measurement quantities are widely known based on the measurement theory [10]. The second assumption is that the observed data are uniformly distributed over the possible states of the objective system [12]. The lack of the uniform distribution of the data over a certain value range of a quantity implies the low observability of the quantity [7]. 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 paper. The third assumption is that the simultaneous equation model under consideration is not over-constrained where the number of the equation is not more than the number of quantities in the model. This assumption always holds for the models in scientific and engineering domains, since the over-constrained state does not exist in any real world process.

2 Structure of a Simultaneous Equation Model

The principle to discover the simultaneous equation structure from passively observed data is based on some fundamental and generic characteristics of simultaneous equation models presented in the past work [11]. These characteristics are briefly explained though an example electric circuit depicted in Fig. 1. This can be represented by the following simultaneous equation model.

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

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. We consider a thought 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.(1), the values of the other quantities, V_1, V_2 and I_1, that are involved in the first, the third and the forth 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 $RQ(\subset Q)$ for external specification, when the values of all quantities in $NQ = CQ - RQ$ are determined where CQ ($RQ \subset CQ \subset Q$) 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. 1 can be represented by the following different simultaneous equation formula.

$$I_1R_1 = I_2R_2 \ \#1',\ V_2 = I_2R_2 \ \#2,\ V_e = V_1 \ \#3 \text{ and } V_e = V_2 \ \#4. \tag{2}$$

If the same specification on V_e and R_1 is made in Eq.(2), 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 observation on the electric circuit, the value of I_1 is physically determined, and this fact contradicts the consequence derived by the analysis on Eq.(2). In contrast, the model of Eq.(1) 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 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 [11].

Theorem 1 (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.(1).

$$\{\#3, \#4\}(n = 2),\ \{\#1, \#3, \#4\}(n = 3),$$
$$\{\#2, \#3, \#4\}(n = 3),\ \{\#1, \#2, \#3, \#4\}(n = 4). \tag{3}$$

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*

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

where L is the set of all complete subsets of the model. The set of essential quantities DQ_i of CE_i which do not belong to any other complete subsets but are involved only in CE_i is also defined as

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

where CQ_i is the set of all quantities in CE_i. The order δn_i and the freedom δm_i of DE_i are defined as

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

For instance, the following independent components can be found for Eq.(1).

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

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 $\{V_1 = I_1R_1\#1, V_e = V_1\#3, V_e = V_2\#4\}$ in Eq.(1) which is a complete subset of order 3 can be transformed by the linear transformation as follows.

$$\begin{aligned} &2V_e + V_1 + V_2 = 4I_1R_1 \ \#1, \ 2V_e = 2V_1 - V_2 + I_1R_1 \ \#3, \\ &\text{and } 3V_e = -V_1 + 2V_2 + 2I_1R_1 \ \#4. \end{aligned} \tag{5}$$

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 [11]. Accordingly, only the following formulae 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.(5). 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 a simultaneous equation 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 a simultaneous equation is derived to be a "structural form", then the form is named "structural canonical form".*

The structural form of Eq.(1) is shown as follows.

$$DE_1 = \{f_{11}(V_e, V_1, V_2) = 0 \ \#3, \ \ f_{12}(V_e, V_1, V_2) = 0 \ \#4\},$$
$$DE_2 = \{f_2(V_e, V_1, V_2, I_1, R_1) = 0 \ \#1\}, \ DE_3 = \{f_3(V_e, V_1, V_2, I_2, R_2) = 0 \ \#2\}, \quad (6)$$

where $f(\bullet) = 0$ is an arbitrary formula to represent a quantitative relation. Because Eq.(1) is a structural form, Eq.(6) is the structural canonical form.

3 Principle and Algorithm

3.1 Quasi-Experiment on Dependency

If actual experiments are applicable, the algorithm of SSF can search the complete subsets in which quantities are mutually constrained through the control to fix the values of the other quantities. However, when only the passively observed data are available, a novel principle, "*quasi-experiment on dependency*," proposed in this study is needed to enable virtual experiments under the aforementioned assumption that the observed data are uniformly distributed over the possible states of the objective process. Given a set of quantities representing the objective process, $Q = \{q_1, ..., q_w\}$, and the set of their passively observed data, $OBS = \{X_1, ..., X_t\}$ where X_i is the i-th observation of Q, we consider to virtually control each quantity q_k in a subset $Q_c (\subset Q)$. As depicted in Fig. 2, a datum $X_g (\in OBS)$ is chosen, and the data of OBS involved in the vicinity of X_g in the subspace defined by Q_c are sampled as OBS_{cg}. The vicinity is defined as follows for every $q_k (\in Q_c)$.

$$\Delta q_k = |q_k - q_{kg}| < \epsilon_k, \quad (7)$$

where ϵ_k is a parameter to define the size of the vicinity. ϵ_k is determined as 5% of the total value range of q_k upon an extensive parameter survey in this paper. This vicinity is shown as a rectangular parallelepiped in Fig. 2 (a). This operation is called "*quasi-control*" and Q_c "*quasi-controlled quantity set*".

Furthermore, for a quantity q_m in $Q - Q_c$, the following correlation coefficient between q_m and each $q_d (\in Q - Q_c - \{q_m\})$ is calculated within the data of OBS_{cg}.

$$r_{md} = \frac{S_{md}}{\sqrt{S_{mm}}\sqrt{S_{dd}}}, \quad (8)$$

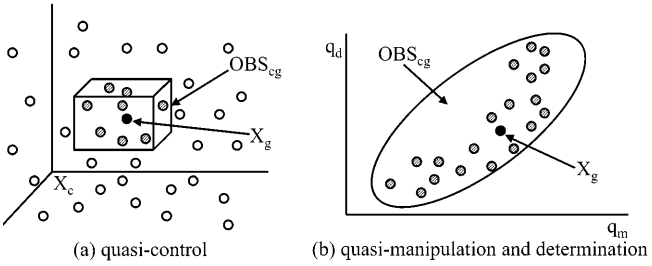

Fig. 2. Quasi-experiment on dependency

where

$$S_{md} = \sum_{OBS_{cg}} q_m q_d - (\sum_{OBS_{cg}} q_m \sum_{OBS_{cg}} q_d)/|OBS_{cg}|,$$

$$S_{mm} = \sum_{OBS_{cg}} q_m^2 - (\sum_{OBS_{cg}} q_m)^2/|OBS_{cg}|, \quad S_{dd} = \sum_{OBS_{cg}} q_d^2 - (\sum_{OBS_{cg}} q_d)^2/|OBS_{cg}|.$$

If r_{md} shows significant correlation as depicted in Fig. 2 (b), q_m and q_d may be mutually constrained under the quasi-control of Q_c in the equation structure embedded in the given data. This process is considered to virtually manipulate the value of q_m by using the scatter of q_m's values in OBS_{cg} and to check if the value of q_d is determined. The operation on q_m is called "*quasi-manipulation*" and q_m "*quasi-manipulated quantity*". The determination of q_d is called "*quasi-determination*" and q_d "*quasi-determined quantity*".

The significance of r_{md} must be tested by the following index and criterion based on the one-sided interval of t-distribution, $t(|OBS_{cg}| - 2, \alpha)$, under the freedom $|OBS_{cg}| - 2$ and a significance level $\alpha(= 0.05)$,

$$r(|OBS_{cg}| - 2, \alpha) = \frac{t(|OBS_{cg}| - 2, \alpha)}{\sqrt{|OBS_{cg}| - 2 + \{t(|OBS_{cg}| - 2, \alpha)\}^2}},$$

$$|r_{md}| \geq r(|OBS_{cg}| - 2, \alpha) \Rightarrow r_{md} \text{ is significant.} \tag{9}$$

The test may fail when the relation between q_m and q_d has strong non-monotonicity. However, this possibility is not very problematic, since the first principle law equations do not contain very strong nonlinearity in most of cases [6,10]. Furthermore, this test is repeated for multiple OBS_{cg}s defined by p different X_gs to confirm the stability of the t-test consequences. p is set to be 10 in this work which is sufficient to check the stability. Let s be the number of the test satisfying the condition Eq.(9). Because s follows the binomial distribution $B(p, 1 - \alpha)$, the following condition should be met, if q_m and q_d is mutually constrained.

$$s/p \geq (1 - \alpha) - 2\sqrt{\alpha(1 - \alpha)/p} \; (\simeq 0.8). \tag{10}$$

The principle of the quasi-experiment on dependency seems similar to the quasi-bi-variate test of the extended SDS. However, the latter assumes that

(S1) *Let* $Q = \{q_k | k = 1, ..., w\}$ *be a set of quantities to appear in the model of an objective process, and let* OBS *be a given data set for* Q. *Set* $DEQ = \phi, \quad DQ = \phi, \quad N = \phi, \quad M = \phi, \quad h = 1$ *and* $i = 1$.

(S2) *Choose* $C_j \subset DQ_j \in DQ$ *for some* DQ_j *and also* $C_q \subseteq Q$, *and take their union* $C_{hi} = ... \cup C_j \cup ... \cup C_q$, *while maintaining* $|C_j| \leq \delta m_j$ *and* $|C_{hi}| = h$. *For every* $q_m \in C_{hi}, m = 1, ..., |C_{hi}|$, *let* $Q_c = C_{hi} - \{q_m\}$, *and apply quasi-experiment on dependency under the quasi-controlled quantity set* Q_c, *quasi-manipulated quantity* q_m *and* OBS. *If* $q_d (\in Q - C_{hi})$ *has significant correlation with any* q_m, q_d *is quasi-determined.*

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

(S4) *If all quantities are quasi-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* DEQ, DQ, N *and* M *represent the sets of quantities involved in independent components, the sets of essential quantities of independent components, their orders and their degrees of freedom respectively.* SCF *is a list of all* $f_{jr}(q_k \in DEQ_j) = 0$ $(r = 1, ..., \delta n_j)$ *for every* $DEQ_j (\in DEQ)$ *and its order* $\delta n_j (\in N)$.

Fig. 3. Algorithm for structural canonical form

the set of quantities $Q = \{q_1, ..., q_w\}$ are governed by a complete equation $f(q_1, ..., q_w) = 0$, and searches feasible binary quantitative relations on some pairs of quantities $\{q_i, q_j\} \in Q$ while quasi-controling the rest of the quantities, i.e., $Q_c = Q - \{q_i, q_j\}$, for each pair. On the other hand, the quasi-experiment on dependency searches a simultaneous equation structure not limiting to a complete equation. The quasi-controlled quantity set Q_c can be an arbitrary subset of Q not limiting to $Q - \{q_i, q_j\}$, and multiple pairs of quantities $\{q_m, q_d\} \in Q - Q_c$ can be found under the unique Q_c.

3.2 Algorithm for Structural Canonical Form

Based on the theory described in section 2 and the quasi-experiment on dependency in subsection 3.1, we propose a novel algorithm shown in Fig. 3 to discover the structure of a simultaneous equation from passively observed data. The notations in Fig. 3 follows Definition 2. It takes a list of quantities Q and their observed data OBS, and outputs the structural canonical form SCF. Starting from the small set C_{hi} which is a union of Q_c and $\{q_m\}$, q_d for the q_m is searched. Though the quasi-experiment on dependency for a q_m can derive all elements of D_{hi} in principle, the experiment is repeated while selecting every quantity in C_{hi} as q_m. This is because some quantity q_m may not have enough sensitivity to change the value of its q_d, and does not show the significant correlation even if they are mutually constrained. The resultant D_{hi} together with $C_{hi} = Q_c + \{q_m\}$ forms a set DEQ_{hi} of quantities belonging to the independent component of a complete subset. Then its set of the essential quantities DQ_{hi}, its order δn_{hi} and its freedom δm_{hi} are derived. Based on the modular lattice structure of a simultaneous equation, the $|C_{hi}| (= h)$ quantities for quasi-control and quasi-manipulation are taken from the union of δm_j quantities in

each independent component DQ_j and the quantities not included in any independent components. The constraint of DQ_j does not miss any complete subset in the search due to the monotonic lattice structure among complete subsets. By repeating this procedure, all independent components are found and stored in the list DEQ. Though the complexity of this algorithm is non-polynomial in the worst case, the search space is significantly reduced by DQ_j. At the final step, the formulae where each indicates the quantities to appear in an equation included in the structural canonical form are listed in SCF.

The law discovery systems such as the extended SDS [12] for passively observed data cannot directly accept SCF. The values of the quantities within an independent component are simultaneously constrained in the order δn_i, and the constraints disable the quasi-bi-variate fitting, if the order δn_i is larger than one. To remove this difficulty, the $(\delta n_i - 1)$ quantities are eliminated by the substitution of the other $(\delta n_i - 1)$ equations within the independent component, and the "*maximally eliminated structural canonical form*" $MESCF$ is derived. The algorithm to obtain $MESCF$ has already been reported [11]. Using the resultant $MESCF$, the extended SDS determines the quantitative formula of each equation reflecting the first principle underlying the objective process.

4 Performance Evaluation

The extended SSF has been developed and combined with the extended SDS on a numerical processing shell named MATLAB in a PC of PentiumIII 666MHz and 128MB RAM. The performance has been evaluated through the following artificial examples for certain combinations of data sizes and noise levels.
1) **Two parallel resistances and a battery**: This has been explained in Fig. 1. Its model consists of 4 equations and 7 quantities as shown in Eq.(1).
2) **Two parallel resistances and a battery**: The objective process is identical with the first example except that an extra equation $R_1 = R_2$ is added. Its model consists of 5 equations and 7 quantities.
3) **Heat transfer at walls of holes**: A large solid material having two vertical holes is considered. Gas goes into these holes, and condenses to its liquid phase by providing its heat energy to the walls while flowing in the holes. The heat transfer process is represented by the 8 equations involving 17 quantities.
4) **A circuit of photometer**: An electric circuit of photometer to measure the rate of increase of photo intensity within a certain time period is considered. It consists of 3 transistors, 3 resistance, 1 light Csd sensor, 1 capacitor and 1 current meter. This system is represented by 14 equations involving 22 quantities.

Table 1 is the summary of problem size, required computation time and error rate for each example for the given OBS consisting of 1000 observed data which contain 5% Gaussian noise relative to the absolute value of each quantity in Q. T_{ssf} is the time to derive the $MESCF$. T_{sds} and T_{av} are the total time and the average time per equation required by the extended SDS. T_{ssf} shows strong dependency on the parameter m and n, *i.e.*, the size of the problem, since the algorithm to derive a structural canonical form requires non-polynomial time

Table 1. Computation time and failure rate

Ex.	m	n	av	T_{ssf}	T_{sds}	T_{av}	FR
1)	4	7	2.5	18	37	9	0.0
2)	5	7	2.4	15	46	9	0.0
3)	8	17	3.9	2,936	147	18	0.18
4)	14	22	2.6	13,992	142	10	0.16

m, n: numbers of equations and quantities, av: average number of quantities per equation, T_{ssf}, T_{sds}: CPU time (sec) required by the extended SSF and the extended SDS, T_{av}: average CPU time (sec) per equation required by the extended SDS, FR: failure rate in the discovery of correct simultaneous equation in 100 trials.

Table 2. Failure rate for noise levels

	failure rate (FR)			
Ex.	0%	5%	10%	20%
1)	0.00	0.00	0.00	0.00
2)	0.00	0.05	0.10	0.83
3)	0.13	0.18	0.23	1.00
4)	0.11	0.16	0.24	1.00

Table 3. Computation time and failure rate

Data Num.	T_{ssf}	T_{sds}	FR
100	2,583	79	0.96
1000	2,936	147	0.18
10000	3,288	207	0.11

to the size. T_{ssf} also moderately depends on $n - m$, because the large number of $n - m$ represents the high degree of freedom of the objective simultaneous equation model which exponentially increases the search space. T_{sds} does not seem to very strongly depend on the size of the problem. Because the extended SDS handles each equation independently in $MESCF$, the required time is proportional to the number of equations in the model. The complexity of the extended SDS is known to be around $O(av^2)$ [12]. This is almost consistent with the relation between T_{av} and av. Thus, T_{sds} may vary approximately in $O(mav^2)$. Except the examples 1) and 2), we observed certain level of failure rates FRs in the discovery. Especially when m and/or av are large, the extended SSF tends to become erroneous. This is because the coupling of many quantities though the equations increases the dependency among the quantities in the observation data, and the required assumption that the observed data are uniformly distributed over the possible states of the objective process becomes no more valid.

Table 2 shows FRs of each example under $0\% - 20\%$ relative noise levels and OBS consisting of 1000 data. When the coupling of quantities is stronger, the larger FR is observed. This tendency is significant in the difference of FRs between the examples 1) and 2). In the example 2), the coupling effect of the extra equation $R_1 = R_2$ significantly increases FR. In addition, the examples containing tight coupling show high sensitivity to the increase of the noise level. Table 3 shows the required computation time and the failure rates of the example

3) for OBS of $100 - 10000$ data and 5% relative noise level. T_{ssf} and T_{sds} seem to be almost $O(\log |OBS|)$. This is because only a limited number of the data sampled by Eq.(7) are used for the discovery even for the large amount of data. FR shows the significant increase for small data size, because the statistical stability is not ensured. In short summary, the computational complexity of the extended SSF seems to be crucial for a very large-scale problem but not for the large data size. The upper limit of the noise contained in the data is considered to be 10% for the extended SSF. The performance of the combined use of the extended SSF and SDS seems to work well for numbers of engineering problems.

5 Application to Practical Problem

The proposed method has been applied to a real world problem to discover a simultaneous equation model consisting of generic law formulae governing the mental preference of people for social infrastructures based on their subjective impressions. We designed a questionnaire sheet to ask the subjective evaluation on the five infrastructures of aviation transport facilities, waste disposal facilities, nuclear power plants, automobile transport facilities and oil power plants from the viewpoints of affinity q_1, unsafety q_2, scale of facility q_3, frequency of daily contacts q_4, benefit q_5, availability of alternative measure q_6 and genetic influence q_7. The last viewpoint may be meaningful only for the infrastructures producing radio-active and/or chemical wastes, and may be evaluated as negligible for the others. The former four viewpoints are asked in form of pair wise comparisons, and the obtained categorical data are transformed to ratio scale quantities by using the constant-sum method which is widely used in the experimental psychology [9]. The latter three are asked in form of the choice from categorical degrees, and the data are transformed to interval scale quantities by following the method of successive categories which is also widely used [1]. We distributed this questionnaire sheet to 482 persons living in a district of a country where the aforementioned facilities are located within a certain distance, and all of the answer sheets have been collected back. Hence, $OBS = \{X_1, X_2, ..., X_{482}\}$ was obtained where $X_i = [q_{1i}, q_{2i}, ..., q_{7i}]$.

The extended SSF was applied to OBS, and the following structural canonical form SCF and maximally eliminated structural form $MESCF$ were obtained.

$$SCF = \{f_{11}(q_1, q_4, q_5, q_6) = 0,\ f_{12}(q_1, q_4, q_5, q_6) = 0,\ f_{13}(q_1, q_4, q_5, q_6) = 0,\\ f_{21}(q_2, q_3, q_7) = 0,\ f_{21}(q_2, q_3, q_7) = 0\} \tag{11}$$

$$MESCF = \{f'_{11}(q_1, q_6) = 0,\ f'_{12}(q_4, q_6) = 0,\ f'_{13}(q_5, q_6) = 0,\\ f'_{21}(q_2, q_7) = 0,\ f'_{21}(q_3, q_7) = 0\} \tag{12}$$

Subsequently, the extended SDS was applied to OBS based on $MESCF$ and the scale-type information, and it derived the following model.

$$q_6 = -0.59 \log q_1 - 1.09,\ q_6 = 1.04 \log q_4 + 1.34,\ q_6 = 0.69 q_5 + 0.57,\\ q_7 = -0.90 \log q_2 - 1.00, q_7 = -0.47 \log q_3 - 1.00. \tag{13}$$

The statistical tests on the goodness of fitting [12] indicated the sufficient accuracy of this model. The former three equations relate affinity, frequency of daily contacts, benefit and availability of alternative measure, and can be interpreted to represent a psychological mechanism developing the affinity on a social facility based on its benefit and necessity in people's daily life. The latter two relate unsafety, scale of facility and genetic influence. They seem to represent another psychological mechanism developing the sense of danger on a social facility.

6 Discussion and Conclusion

Dzeroski and Todorovski developed LAGRANGE which discovers simultaneous equation models from observed data [2]. However, the mathematical admissibility is not considered in the discovery process, and many redundant representations of simultaneous equations can be derived at an expense of high computational complexity. They recently extended it to LAGRAMGE which allows the user to explicitly define the space of possible equations [8]. But, it does not provide definitions to efficiently prune the search space within the admissible equation formulae. In contrast, COPER, which also discovers simultaneous equations, uses very strong mathematical constraints based on the unit dimensions to prune the meaningless terms [5]. However, it essentially requires the unit information which is not frequently obtained in non-physical domains. The major advantages of our proposing method 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.

In this paper, we proposed the principle and the algorithm of a practical method to discover the first principle based simultaneous equations from passively observed data. The satisfactory performance of the method has been confirmed through simulations. Moreover, its high practicality has been demonstrated through a real application in socio-psychology. The application of the proposed method to a real-world research project collaborated with socio-psychologists is currently underway.

References

[1] Comrey, A. L.: A proposed method for absolute ratio scaling. In *Psychometrika*, Vol.15 (1950) 317–325

[2] 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

[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] 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

[5] Kokar, M. M.: Determining Arguments of Invariant Functional Descriptions. In *Machine Learning*, Boston, Kluwer Academic Publishers (1986) 403–422
[6] 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)
[7] Ljung, L.: *System Identification*, P T R Prentice-Hall (1987)
[8] 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
[9] Torgerson, W. S.: In *Theory and Methods of Scaling*, N.Y.: J. Wiley (1958)
[10] 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
[11] Washio T. and Motoda, H.: Discovering Admissible Simultaneous Equations of Large Scale Systems, In *Proceedings of AAAI'98*, Madison (1998) 189–196
[12] 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