Automated Scientific Modeling from Observed Data and its Application to Socio-Psychology

Takashi Washio and Hiroshi Motoda
I.S.I.R., Osaka University
8-1, Mihogaoka, Ibarakishi, Osaka 567, Japan
washio@sanken.osaka-u.ac.jp

Yuji Niwa
I.N.S.S., Inc.
64 Sata, Mihamacho, Mikatagun,
Fukui 919-1205, Japan

Abstract

The knowledge-based automated modeling framework such as CML can be applied only to the systems where their valid background knowledge is available. The conventional model equation discovery systems such as BACON require experimental environments to acquire their necessary data. The mathematical techniques, e.g., linear system identification and neural network fitting, presume the classes of equations to model a given observed data set. The study reported in this paper proposes a novel method to discover an admissible model equation from a given set of observed data while the equation is ensured to reflect first principles governing the objective system. The power of the proposed method comes from the use of the scale-types of the observed quantities, a mathematical property of identity and quasi-bi-variate fitting which identify the admissible solutions from the given data set. Its principles and automated algorithm are described with moderately complex examples, and its practicality is demonstrated through the real application to a socio-psychological modeling task.

Introduction

The knowledge-based automated modeling of an objective system has been a major research field of qualitative reasoning. One of the representative methods in this field was proposed by B. Falkenhainer and K. Forbus under the framework of compositional modeling (Falkenhainer and Forbus 1991). Later, a language, CML, to describe the knowledge of the model fragments and the modeling process has been provided in a reusable and shared manner for engineers (Falkenhainer *et al.* 1994). Some recent work developed collaborative environments for the CML-based modeling to enhance its usability (Iwasaki *et al.* 1997). Concurrently, a number of recent researches proposed new ontology to extend the range of domains of the knowledge-based automated modeling. The representatives are the hybrid representations of quantitative and qualitative models (Mosterman and Biswas 1997) and the qualitative representations of causal time (Kitamura *et al.* 1997). Moreover, a new approach to modify the generated model through the comparison with the observed behaviors have been assessed by N. Smith (Smith 1998). An advantage of the knowledge-based automated modeling framework is its capability to construct a model of an objective system based on the domain background knowledge even when any observation of the system behavior is not available. Another advantage is its capability to develop the model reflecting the first principles underlying the objective system, if the associated background knowledge is valid. However, the applicability of this framework is limited to the systems, e.g., physical systems, where their valid background knowledge is available.

Another framework of the automated modeling is the approach driven by experimental data, and this is explored through the research field of scientific discovery in AI context. The most well known pioneering system to discover scientific law equations from experimental data is BACON (Langley *et al.* 1985). It searches for a *complete equation* governing the data measured in a continuous process, where the complete equation is an equation constraining n quantities with $n-1$ degree of freedom [1]. FAHRENHEIT (Koehn and Zytkow 1986), ABACUS (Falkenhainer and Michalski 1985), etc. are the successors that basically use similar algorithms to BACON to discover a complete law equations. To reduce the high computational cost of their algorithm, some subsequent discovery systems, e.g., FAHRENHEIT, ABACUS and COPER (Kokar 1985), introduced the use of the unit dimension of physical quantities to prune the meaningless solutions. A difficulty of this approach is the narrow applicability only to the quantities whose units are clearly known. On the other hand, the most recent scientific law discovery system, SDS, has overcome the difficulties of the past systems (Washio and Motoda 1997) (Washio and Motoda 1998). It discovers scientific law equations by limiting its search space to mathematically admissible equations in terms of the constraints of *scale-type* and *identity*. These constraints come from the basic

[1] The equation ${x_1}^2 + {x_2}^2 + ... + {x_n}^2 = 0$ is not complete, since the values of all n quantities is 0, i.e., n quantities are constrained with no degree of freedom. On the other hand, $x_1 + x_2 + ... + x_n = 0$ is complete.

characteristics of the quantities' definitions and the relations necessarily standing in the objective systems. The admissible equations discovered by SDS are considered to have valid structures reflecting the relations among quantities in the fundamental mechanisms governing the objective system. The equations having such valid structures is called *first principle equations* in this paper. The detailed characterization of the first principle equations can be seen elsewhere (Washio and Motoda 1998). Since the knowledge of scale-types is widely obtained in various domains, SDS is applicable to non-physical domains including biology, sociology, economics, etc.

A major drawback of these approaches is the limited applicability to practical situations. They require the experimental environment and the interaction to control and measure the system states. The number of controllable quantities is quite limited, and even none of them are controllable due to some practical reasons in many scientific and engineering domains. For instance, the astronomical experiments to control the parameters of fusion reactions in the distant huge stars are physically impossible. The economical experiments to cause financial panics are unacceptable for our society. Under these situations where only passive observation is possible, the mathematical techniques, e.g., linear system identification (Ljung 1987) and neural network, have been traditionally applied to derive quantitative relations among observed quantities. However, the derived relations are not ensured to represent the first principle because they presume some structures of the model equations such as linear formulae and hierarchical sigmoid formulae. The discovery of the first principle equations under the passive observation will play highly important role to understand the fundamental mechanisms underlying the variety of the objective systems. To achieve this aim by the technique of the aforementioned scientific discovery, the current framework must be changed to discover the first principle equations by using only the data obtained under the passive observation.

The past scientific discovery is for the class of the problem to discover the law equations under the experimental environment. Its algorithm basically consists of two operations. The first is called *bi-variate fitting* which identifies the relation within a pair of quantities, $P_{ij} = \{x_i, x_j\} \subseteq X$, where $X = \{x_1, x_2, ..., x_m\}$ is the set of all quantities to represent the objective system. It derives the pairwise relation within P_{ij} from the experimental data in which the values of all quantities in the rest $X - P_{ij}$ is fixed by the experimental control. This pairwise relation is noted as $f_{X-P_{ij}}(x_i, x_j) = 0$. The bi-variate fitting is required to identify the intrinsic structure of the relation within P_{ij} under the exclusion of the influence from the other quantities. The second operation is to merge the multiple pairwise relations into an equation. Through the iteration of these two operations, the complete equation $\phi(x_1, x_2, ..., x_m) = 0$ to represent the entire objective system is derived. In the new class of the problem to discover the law equations under the passive observation environment, the experimental control of the values of $X - P_{ij}$ is not allowed. Accordingly, the conventional bi-variate fitting is not applicable. In this paper, "*quasi-bi-variate fitting*", an extension of the bivariate fitting based on a polynomial approximation, is proposed to enable the application of the framework of SDS to this class of the problem.

The proposed quasi-bi-variate fitting requires some assumptions which are feasible in many practical applications. One is that the scale-types of all observed quantities are known. This does not limit the applicability of the proposed method because the scale-types of the measurement quantities are widely known based on the measurement theory as shown later. Another assumption is that the observed data are uniformly distributed over the value range that each quantity can take within the possible states of the objective system. If the observed data points are concentrated within the vicinity of a value for some quantity, the data set does not provide any meaningful information on the relation of the quantity with the others. Accordingly, the discovery of the first principle equations becomes difficult if this assumption is strongly violated. However, this requirements is not limited to our proposed approach. The lack of the uniform distribution of the data over a certain value range of a quantity implies the low observability of the quantity (Ljung 1987). It is well known that the conventional approaches such as the linear system identification and the neural network do not derive valid models of the objective systems under the low observability condition. This limitation is generic for any data-driven modeling approaches, and further discussion on this issue is out of scope of this paper.

The objectives of this paper are (i) to propose the principles and an algorithm of the quasi-bi-variate fitting under the framework of SDS, (ii) to evaluate the basic performance of the proposed approach through simulations and (iii) to demonstrate its high practicality through a real application.

Background Principles

Before proposing quasi-bi-variate fitting, some background principles are explained to facilitate the comprehension. The details of the principles are described in our papers on SDS (Washio and Motoda 1997) (Washio and Motoda 1998). Only its outline is explained in this section.

Scale-type Constraints

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 scale-types and different kinds of mea-

surement are derived if numerals can be assigned under different rules, and categorized the scale-types of quantities based on the operation rule of the assignment. Quantitative measurement quantities are mathematically characterized and categorized into three major quantitative scale-types of interval scale, ratio scale and absolute scale. 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*: $x' = x$".

Luce claimed that the basic formula of the functional relation among quantities of ratio and interval scales can be determined by their scale-type features, if the quantities are not coupled through any dimensionless quantities (Luce 1959). Under this condition, the quantities should share some common basic dimensions, and consequently the unit change of a quantity affects the value of other quantity. Suppose x and y are both ratio scale quantities, and y is defined by x through a continuous functional relation $y = u(x)$. Suppose the form of $u(x)$ is logarithmic, *i.e.*, $y = \log x$. We multiply a positive constant k to x, *i.e.*, a change of unit, without violating the group structure of the ratio scale quantity x, then this leads $u(kx) = \log k + \log x$. This fact causes the shift of the origin of y by $\log k$, and violates the group structure of y which is the ratio scale quantity. Hence, the direct functional relation from x to y must not be logarithmic. Based on the admissibility condition of the relations among ratio and interval scale quantities, we mathematically derived the following two theorems to represent the generic formulae of the relations (Washio and Motoda 1997). [2]

Theorem 1 (Extended Buckingham Π-theorem) *If $\phi(x_1, x_2, x_3, ..., x_m) = 0$ 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*

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

where m is the number of arguments of ϕ, and w is the basic number of bases in $x_1, x_2, x_3, ..., x_m$, respectively.

Bases are such basic scaling factors and origins independent of the other bases in the given ϕ, for instance, as length $[L]$, mass $[M]$ and time $[T]$ of physical unit and as temperature origin $[t_0]$ of Celsius and elevation origin $[h_0]$ of potential energy for interval scale quantities. The relation of each Π_i with the arguments of ϕ is given by the following theorem.

Theorem 2 (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 one of the forms*

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

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

where R and I can be null sets, P is a partition of I, and P_g is a partition of $I - I_g$ where $I_g \subseteq I$. All coefficients except Π are constants.

The formula in Theorem 1 is called an "*ensemble equation*" and those in Theorem 2 "*regime*"s.

Table 1 shows all admissible bi-variate relations deduced from the "*Extended Product Theorem*". The coefficients G_{ij} and H_{ij} can be dependent on the other quantities except x_i and x_j. Thus, they are represented as $G_{ij}(X - P_{ij})$ and $H_{ij}(X - P_{ij})$, while a_{ij} is independent, and remains constant. These consequences play an important role in the quasi-bi-variate fitting explained later.

Table 1: Admissible bi-variate relations within a regime

scale-types		
x_i	x_j	admissible relations
ratio	ratio	$x_j = G_{ij}\|x_i\|^{a_{ij}}$
ratio	interval	$x_j = G_{ij}\|x_i\|^{a_{ij}} + H_{ij}$
		$x_j = a_{ij}\log\|x_i\| + H_{ij}$
interval	ratio	$x_j = G_{ij}\|x_i + H_{ij}\|^{a_{ij}}$
		$x_j = G_{ij}\exp a_{ij}x_i$
interval	interval	$x_j = a_{ij}x_i + G_{ij}$

Identity Constraint

When the scale-types of quantities are absolute and/or unknown as the case of "*ensemble equation*", the scale-type constraints are not applicable. In such cases, the identity constraint is used to determine the admissible equation.

The basic principle of the identity constraints comes in by answering the question that "*what kind of relation holds 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 $\theta_i = G_{hi}(\theta_j)\theta_h + H_{hi}(\theta_j)$ and $\theta_j = G_{hj}(\theta_i)\theta_h + H_{hj}(\theta_i)$ are given, the following identity equation is obtained by

[2] The original *Buckingham Π-theorem* (Buckingham 1914) and *Product Theorem* (Bridgman 1922) represent the generic relation among only ratio scale quantities.

Table 2: Identity constraints

bi-variate relation	general relation
$\theta_j = G_{ij}\theta_i + H_{ij}$	$\sum_{(A_i \in 2^{LQ}) \& (p \not\subseteq A_i \forall c \in LR)} {a_i} \prod_{\theta_j \in A_i} \theta_j = 0$
$\theta_j = H_{ij}\theta_i^{G_{ij}}$	$\prod_{(A_i \in 2^{PQ}) \& (p \not\subseteq A_i \forall c \in PR)} \exp(a_i \prod_{\theta_j \in A_i} \log \theta_j) = 0$

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

solving each for θ_h.

$$\theta_h \equiv \frac{1}{G_{hi}(\theta_j)}\theta_i - \frac{H_{hi}(\theta_j)}{G_{hi}(\theta_j)} \equiv \frac{1}{G_{hj}(\theta_i)}\theta_j - \frac{H_{hj}(\theta_i)}{G_{hj}(\theta_i)}$$

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

$$\begin{aligned} 1/G_{hi}(\theta_j) &= -\alpha_1\theta_j - \beta_1, \\ H_{hi}(\theta_j)/G_{hi}(\theta_j) &= \alpha_2\theta_j + \beta_2. \end{aligned}$$

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. Thus, by knowing some bi-variate linear relations among the quantities, the admissible equation formula for the whole quantities is derived.

This principle is generalized to various bi-variate relations f among multiple quantities. Table 2 shows such relations for linear relations and product relations.

Quasi-bi-variate Fitting

As noted in the first section, the conventional bi-variate fitting requires experimental control of some quantities, and is not applicable to the passive observation environments. To overcome this difficulty, we propose the "*quasi-bi-variate fitting*" procedure which extracts a bi-variate relation between two quantities under the approximated constant values of the other quantities.

Fitting for Scale-type Constraint

Figure 1 shows the outline of its principle for the admissible bi-variate relations in Table 1. Let $OBS = \{X_1, X_2, ..., X_n\}$ be a set of observations where each $X_h (h = 1, ..., n)$ is a m-dimesional vector of observed values of the m quantities in X. The fitting of a candidate bi-variate formula for a pair of two quantities $P_{ij} = \{x_i, x_j\} (\subseteq X)$ is applied to a subset of OBS. This subset OBS_{ijg} is chosen in such a way that every quantity $x_k \in (X - P_{ij})$ takes a value in the vicinity of the value of x_{kg}, where $X_g = \{x_{1g}, x_{2g}, ..., x_{mg}\} \in OBS$ is an arbitrary chosen observation vector. The vicinity of x_{kg} is defined as

$$\Delta x_k = |x_k - x_{kg}| < \epsilon_k. \quad (1)$$

ϵ_k determines the size of the vicinity. This vicinity is indicated by a rectangular cube in the upper figure of Fig. 1. Every admissible bi-variate formula indicated in Table 1 is generally represented by the form

$$F_{ij}(P_{ij}, a_{ij}, G_{ij}(X - P_{ij}), H_{ij}(X - P_{ij})) = 0. \quad (2)$$

Here, G_{ij} and H_{ij} are dependent on the quantities in $X - P_{ij}$, while a_{ij} remains constant. Given an OBS_{ijg}, if each ϵ_k is moderately small, the values of G_{ij} and H_{ij} become slightly dependent on $X - P_{ij}$, and their polynomial approximation of the order p can be applied.

$$\begin{aligned} F_{ij}(P_{ij}, a_{ij}, G^0_{ijg} + \sum_{k=1}^{m}\sum_{h=1}^{p}(G^h_{ijkg}\Delta x_k^h), \\ H^0_{ijg} + \sum_{k=1}^{m}\sum_{h=1}^{p}(H^h_{ijkg}\Delta x_k^h)) = 0, \end{aligned} \quad (3)$$

where G^h_{ijkg} and H^h_{ijkg} stand for the coefficients of the h-th order of Δx_k at X_g. The least square fitting of Eq.(3) approximately provides the functional relation within P_{ij} and the coefficient a_{ij} as depicted in the bottom figure of Fig. 1. while almost excluding the influence of the other dimensions $X - P_{ij}$.

After the least square fitting of this formula to OBS_{ijg}, the goodness of the fitting is evaluated by the following F-test.

$$\text{If } F_0 > F(1, n_{ijg} - 2, \alpha) \quad (4)$$
then the fitting is acceptable else unacceptable,

where

$$V_R = (\sigma_{ij}{}^2/\sigma_{ii})^2, V_e = \sigma_{ee}{}^2/(n_{ijg} - 2), F_0 = V_R/V_e.$$

Here, $\sigma_{ij}{}^2, \sigma_{ii}{}^2$ and $\sigma_{ee}{}^2$ are the correlation of x_i and x_j, the squared summation of x_i and the squared summation of fitting error respectively. n_{ijg} is the total number of data points in OBS_{ijg} and $F(1, n_{ijg} - 2, \alpha)$ the lower bound of F value under the degree of freedom $(1, n_{ijg} - 2)$ and a risk rate α. The value of n_{ijg} is subject to the size of the vicinity ϵ_ks. α is set to be 0.05 throughout this paper. The quasi-bi-variate fitting and the F-test are repeated for multiple OBS_{ijg}s defined by q different X_gs. This repetition is to confirm the stability of the F-test consequences. q is set to be 10 which is sufficient enough to check the stability of the consequences of the F-test. After these trials, the following χ^2-test over the q trials is conducted to check if a_{ij} in Eq.(3) is identified as constant.

$$\text{If } \chi_0^2 < \chi^2(q - 1, \alpha) \quad (5)$$
then a_{ij} is constant else not constant,

where $\sigma_{0ij}{}^2 = (\sum_{g=1}^{q} \delta a_{ijg}{}^2)$,

$$\sigma_{ij}{}^2 = \sum_{g=1}^{q} a_{ijg}^2 - (\sum_{g=1}^{q} a_{ijg})^2/q, \chi_0^2 = \sigma_{ij}{}^2/\sigma_{0ij}{}^2.$$

Here, δa_{ijg} is the standard error of a_{ijg} estimated from the residual error of the quasi-bi-variate fitting. $\chi^2(q - 1, \alpha)$ stands for the upper bound of χ^2 value

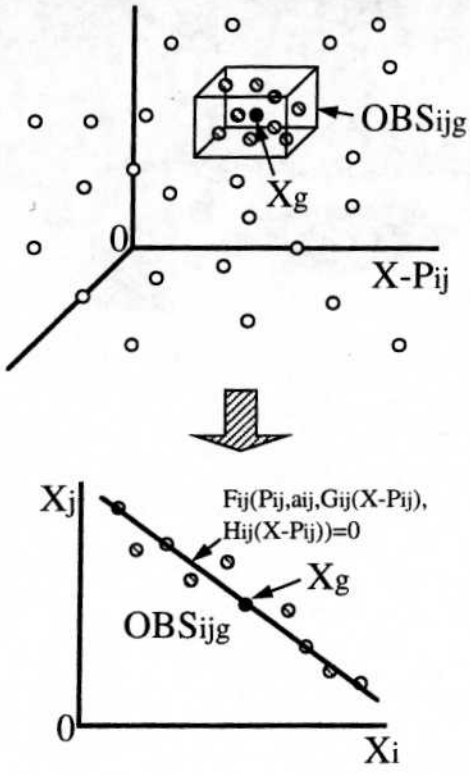

Figure 1: Outline of quasi-bi-variate fitting

under the degree of freedom $(q-1)$ and the risk level α. The approximated bi-variate formula Eq.(3) that passed these tests is considered to be a part of the admissible model equation indicated in Theorem 1 and 2. The expectation value of a_{ij} is estimated as $\bar{a}_{ij} = (\sum_{g=1}^{q} a_{ijg})/p$.

Fitting for Identity Constraint

For the bi-variate relations of the identity constraints indicated in the first column of Table 2, the similar scheme of the quasi-bi-variate fitting is applied. Let $OBS = \{\Theta_1, \Theta_2, ..., \Theta_n\}$ be a set of observations, where each $\Theta_h (h = 1, ..., n)$ is a vector of observed values of the m' quantities in $\Theta = \{\theta_1, \theta_2, ..., \theta_{m'}\}$. The fitting of a candidate bi-variate formula for a pair of two quantities $P_{ij} = \{\theta_i, \theta_j\} (\subseteq \Theta)$ is applied to OBS_{ijg}. OBS_{ijg} is a subset of OBS in the vicinity of the value of θ_{kg}, where $\Theta_g = \{\theta_{1g}, \theta_{2g}, ..., \theta_{m'g}\}$ is arbitrary chosen in OBS. Every bi-variate formula indicated in Table 2 is generally represented by the form

$$F_{ij}(P_{ij}, G_{ij}(\Theta - P_{ij}), H_{ij}(\Theta - P_{ij})) = 0. \qquad (6)$$

Again, the terms of $G_{ij}(\Theta - P_{ij})$ and $H_{ij}(\Theta - P_{ij})$ are approximated by their polynomials. After the least square fitting of the approximated formula to OBS_{ijg}, the goodness of the fitting is evaluated by the F-test in the similar manner.

In the quasi-bi-variate fitting, the number of data included in OBS_{ijg} increases by relaxing the size of the vicinity of X_g and Θ_g. This has an effect of reducing the statistical error of the quasi-bi-variate fitting. On the other hand, if the size of the vicinity is too large, the higher order approximation is required to absorb the influence of the values of $X - P_{ij}$ and $\Theta - P_{ij}$. However, excessively high order approximation may introduce some systematic error due to the over-fitting to the data. Accordingly, some appropriate values of ϵ_ks and p must be used for the given data.

Algorithm

As the details of the algorithm to discover a complete model equation in the frame work of SDS are represented in our previous paper, only its essential contents related to the quasi-bi-variate-fitting are explained in this section (Washio and Motoda 1997). Initially, a set of ratio scale quantities RQ, a set of interval scale quantities IQ and a set of absolute scale quantities AQ which are required to express the objective model equation are given together with a set of observed data OBS of these quantities.

Step (1-1) The quasi-bi-variate fitting for scale-type constraints is applied to the bottom formula in Table 1 for pairwise interval scale quantities. The least-square fitting of the formula using the approximation of Eq.(3), F-test to check the goodness of fitting to the data of each subset OBS_{ijg} and χ^2-test to check the constant value of a_{ij} are conducted. Subsequently, the expectation value $\bar{a}_{ij}$ is estimated, and the formulae together with the values of $\bar{a}_{ij}$ are stored into an equation set IE.

This step is now demonstrated by an example of a moderately complex system depicted in Figure 2. This is an electric circuit where the model of this system based on the first principle is represented by the following equation involving eight quantities.

$$\frac{R_{BE}}{h_{FE}} = \frac{1}{R_1/R_2 + 1}(V_1 - V_2)I^{-1} - R_3, \qquad (7)$$

where R_{BE} is the resistance between the base and the emitter and h_{FE} the ratio between the base current and the collector current, respectively. V_1 and V_2 are interval scale, and h_{FE} is absolute scale. The rests are ratio scale. The observed data set is obtained by a numerical simulator. The values of parameter quantities are set to be $R_2 = 1000\Omega$, $R_{BE} = 10^6\Omega$ and $h_{FE} = 100$. The value ranges of the variable quantities are taken to be $0\Omega < R_1 < 1000\Omega, 0\Omega < R_3 < 1000\Omega$ and $0V < V_2 < V_1 < 30V$. The values of these variables are generated by using uniform random numbers over their value ranges in the simulation. Only the five variable quantities R_1, R_3, V_1, V_2 and I are assumed observable in this demonstration. Thus, $IQ = \{V_1, V_2\}$, $RQ = \{R_1, R_3, I\}$ and $AQ = \phi$. The values of the parameter quantities are implicitly assumed to be constant. The total number of data points provided in OBS is 500, and no observation noise is added here. Our proposed method has been implemented in a prototype program. The size of each vicinity ϵ_k, has been

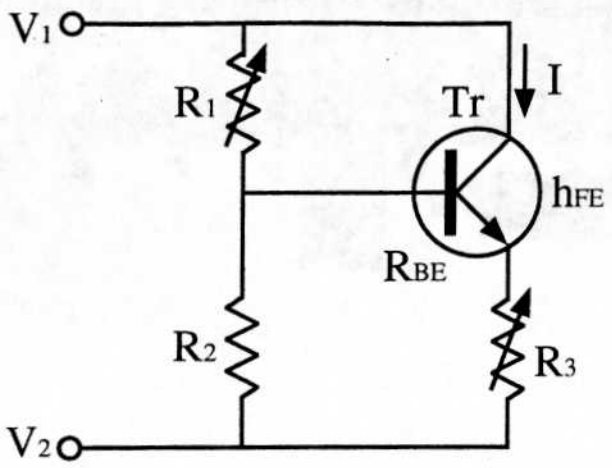

Figure 2: An electric circuit

set at 15% of the difference between the maximum and the minimum values of x_k. The 0th order approximation is used because the observation is not distorted by any noise in this case. As the interval scale quantities are limited to V_1 and V_2, the bottom linear equation in Table 1 is immediately applied to this pair, and the relation $0.98V_1 - V_2 = G_{V_1V_2}$ is identified. Thus $IE = \{0.98V_1 - V_2 = G_{V_1V_2}\}$.

Step (1-2) This step first applies the following triplet-tests. For a triplet of the linear formulae among $\{x_i, x_j, x_h\}$ in IE,

$$x_i = \bar{a}_{hi}x_h + G_{hi}, \quad x_j = \bar{a}_{ij}x_i + G_{ij}, \quad x_h = \bar{a}_{jh}x_j + G_{jh},$$

if they are mutually consistent in terms of $\bar{a}$s, the following condition should be met.

$$\bar{a}_{ij}\bar{a}_{jh}\bar{a}_{hj} = 1.$$

Because of the existence of the noise and the fitting error, this condition does not hold in exact manner, even if the three formulae are consistent. Thus, the following normal distribution test judges if the l.h.s. and the r.h.s. of the above expression are equal.

$$\text{If } N_0 < N(0, \sigma^2, \alpha/2) \text{then } \bar{a}_{ij}, \bar{a}_{jh}, \text{ and }, \bar{a}_{hj} \text{ are mutually consistent else inconsistent,} \quad (8)$$

where

$$N_0 = |1 - \bar{a}_{ij}\bar{a}_{jh}\bar{a}_{hj}|,$$
$$\sigma^2 = (\bar{a}_{jh}\bar{a}_{hj}\delta\bar{a}_{ij})^2 + (\bar{a}_{ij}\bar{a}_{hj}\delta\bar{a}_{jh})^2 + (\bar{a}_{ij}\bar{a}_{jh}\delta\bar{a}_{hj})^2.$$

Here, $\delta\bar{a}_{ij} = \sqrt{(\sum_{g=1}^{p}\delta a_{ijg}^2)/p}$, and $\delta\bar{a}_{jh}$ and $\delta\bar{a}_{hj}$ are similarly defined. $N(0, \sigma^2, \alpha/2)$ stands for the upper bound of the error under the normal distribution and the risk level α. This test is applied to every triplet of equations in IE, and every maximal convex set MCS is searched. A convex set is a set where each triplet of equations among the quantities in this set has passed the test Eq.(8). And, the maximal convex set MCS is a convex set where any superset of the set is not a convex set. In addition, every formula in IE which does not belong to any consistent triplet is also regarded as a tiny MCS. Once all MCSs are found, the formulae are merged into the following form in every MCS.

$$\Gamma = \sum_{x_s \in MCS} a_s x_s,$$

where Γ is an intermediate quantity which appears in the reasoning process. Before the final value of a_s is determined, the following integer-test is applied.

$$\text{If } |a_s - [a_s]| < 2\delta a_s \text{then } a_s = [a_s], \quad (9)$$
$$\text{where } [a_s] \text{ is the nearest integer of } a_s$$
$$\text{and } \delta a_s \text{ std. error of } a_s.$$

This is based on the observation that the majority of the first principle based equations have integer power coefficients and integer linear coefficients for interval scale quantities.

In the current circuit example, an MCS is uniquely determined because $IE = \{0.98V_1 - V_2 = G_{V_1V_2}\}$ contains only one formula, and thus the triplet test is not required. The example of the triplet test is shown in section 6. The above integer-test set a_{V_1}, the coefficient of V_1, to be 1 because $2\delta a_{V_1} = 0.092$. Thus, we obtain $IE = \{V_1 - V_2 = G_{V_1V_2}\}$. Furthermore, V_1 and V_2 in IQ is merged into $G_{V_1V_2}$, and $G_{V_1V_2}$ is stored into a quantity set TQ as $TQ = \{G_{V_1V_2}\}$. $G_{V_1V_2}$ is a new ratio scale quantity by the mutual cancellation of the basic origins of V_1 and V_2. Finally, $TQ = TQ + RQ$ becomes $\{R_1, R_3, I, G_{V_1V_2}\}$.

Step (2-1) Similarly to step $(1-1)$, the quasi-bivariate fitting, F-test and χ^2-test are performed on the quantities included in TQ. Then, the discovered equations are stored in an equation set RE. The unique difference from step $(1-1)$ is to apply the formulae except the bottom one in Table 1 in the quasi-bi-variate fitting. In the circuit example, only the pair of I and $G_{V_1V_2}$ is found to satisfy the first formula in Table 1 as $I = G_{IG_{V_1V_2}}G_{V_1V_2}^{1.003}$. Thus, $RE = \{I = G_{IG_{V_1V_2}}G_{V_1V_2}^{1.003}\}$.

Step (2-2) The triplet-test among the formulae in RE is conducted. The basic procedure is identical with the step (1-2). In the example, as RE contains only one formula again, a unique MCS becomes $\{I, G_{V_1V_2}\}$, and they are merged into a term $G_{IG_{V_1V_2}} = I/G_{V_1V_2}^{1.003}$. The value $a_{G_{V_1V_2}} = 1.003$ is modified to 1 in the integer test since $2\delta a_{G_{V_1V_2}} = 0.014$. Thus, $G_{IG_{V_1V_2}} = I/G_{V_1V_2}$. Then, TQ becomes $\{R_1, R_3, G_{IG_{V_1V_2}}\}$, and finally $TQ = TQ + AQ = \{R_1, R_3, G_{IG_{V_1V_2}}\}$.

Step (3) This is the step to apply the quasi-bivariate fitting for identity constraints. A formula is arbitrary selected from the first column of Table 2. In our current program, a linear formula $\theta_j = G_{ij}\theta_i + H_{ij}$ has the first priority in the selection and a power formula $\theta_j = H_{ij}\theta_i^{G_{ij}}$ the next priority. The quasi-bi-variate fitting of the formula and F-test are applied to each pair of quantities in TQ. If some pairs of the quantities are judged to well fit to the bi-variate formula, the identity constraints is applied. In the example of

$TQ = \{R_1, R_3, G_{IG_{V_1V_2}}\}$, the bi-variate linear relations in the pairs of $\{R_1, G_{IG_{V_1V_2}}\}$ and $\{R_3, G_{IG_{V_1V_2}}\}$ are accepted through the F-test, and thus the following multi-linear formula obtained from the principle of the identity constraints is applied to the entire data set.

$$0 = a_0 + a_1 G_{IG_{V_1V_2}} + a_2 R_1 + a_3 R_3 + R_1 R_3$$

Its least square fitting is accepted by F-test, and thus $AE = \{0 = a_0 + a_1 G_{IG_{V_1V_2}} + a_2 R_1 + a_3 R_3 + R_1 R_3\}$. The values $a_0 = 10^7, a_1 = -10^3, a_2 = 10^4$ and $a_3 = 10^3$ are obtained by the integer-test. By substituting the formulae in IE and RE to $G_{IG_{V_1V_2}}$, the final solution of the admissible model equation of

$$0 = a_0 + a_1(V_1 - V_2)/I + a_2 R_1 + a_3 R_3 + R_1 R_3 \qquad (10)$$

is resulted. As th values of $a_0 - a_3$ correspond to the relations $a_0 = R_{BE}R_2/h_{FE}, a_1 = -R_2, a_2 = R_{BE}/h_{FE}$ and $a_3 = R_2$, this equation is known to be equivalent to Eq.(7).

Step (4) Finally, when multiple candidate model equations remain, a parsimony criterion is applied to prioritize the candidates. Though MDL principle is a representative criterion, AIC, which is widely used in statistics to determine an appropriate numerical model equation, is applied in our program (Akaike 1978). The index of AIC is calculated through the expression

$$AIC = n \ln V_e + 2M, \qquad (11)$$

where $n = |OBS|$, V_e the residual error variance of the model equation and M the number of the coefficients included in the model. The model equation having less value of AIC is preferred in the sense of the parsimony criterion. This is not used in the example of the circuit, since the unique solution Eq.(10) is obtained. Its example is given in latter section 6.

Evaluation through Simulation

Table 3 indicates the required computation time for various examples. "Ideal Gas" is the simulation of the state equation of the ideal gas. "Coulomb", "Stoke's" and "Momentum" are the simulations of Coulomb force law, Stoke's equation and the momentum balance equation. "Circuit*1" is the case of the aforementioned electric circuit where R_2, R_{BE} and h_{FE} are hidden parameters, and "Circuit*2" is the case of the identical circuit where all quantities are observable. They are represented by various number of quantities. The computation time of the proposed algorithm has been evaluated for various numbers of the data points for each example. The computation time does not change significantly with the increase of the data size. This is because the 10 vicinities selected in the quasi-bi-variate fitting cover only a limited portion of the given data when the data size is large. Thus the required computation time increases very slowly. In contrast, the computation time is sensitive to the size of the objective system. The increase is almost order of $O(m^2)$ where m is the number of the quantities in the data. This

Table 3: Required computation time

Example	Number of quantities	CPU time (sec) 50 data	500 data	5000 data
Ideal Gas	4	25.9	46.1	67.9
Coulomb	5	40.5	77.6	112.9
Stoke's	5	46.3	82.6	119.8
Circuit*1	5	43.8	81.6	115.8
Momentum	8	151.4	271.0	385.3
Circuit*2	8	135.2	255.7	371.7

Table 4: Relative Error of Coefficients

Num. of data	Relative noise	Order=0	Order=1	Order=2
50	0%	62%	46%	77%
	0.5%	66%	45%	47%
	5%	133%	65%	77%
500	0%	0%	0%	0%
	0.5%	2.8%	0.9%	0.70%
	5%	3.2%	2.8%	5.7%
5000	0%	0%	0%	0%
	0.5%	0.8%	0.4%	0.5%
	5%	1.3%	1.6%	1.5%

is due to the quasi-bi-variate fitting having the complexity $O(m^2)$. Though the most complex process is the triplet-test which is $O(m^3)$, this test is very simple compared with the data fitting to many data points.

Table 4 shows the average of the relative error of the coefficients under several conditions of the data size, the relative noise level and the order of the approximation of the quasi-bi-variate fitting in case of the aforementioned electric circuit. The size of the vicinity ϵ_k is kept at 15%. When the amount of the data is very limited, the error rate increases significantly. This is because the data points covered by a vicinity is so small that the sufficient statistic accuracy is not maintained. The accuracy of the coefficients is also influenced by the approximation order. In general, the 1st order approximation shows the good performance. This tendency becomes significant, when the number of the data is very limited, and/or the noise level is high. The 0th order approximation does not effectively reduce the influence of $(X - P_{ij})$, when the data points are sparse in the state space. Besides, the 2nd order approximation also becomes erroneous due to the over fitting effect, when the data is sparse and/or the noise level is high.

Automated Modeling in Socio-Psychology

The proposed method has been applied to a real world problem. The objective of the application is to discover

a model formula representing a generic law to govern the mental preference of people on their houses. We assume that a generic law governs the mental preference subject to the cost for buying the house and the social risk at the place of the house. The validity of this assumption is assessed through this application, and the model formula is derived. We designed a questionnaire sheet to ask the preference of the house in the trade off between the frequency of huge earthquakes x_1 (earthquake/year) and the price x_2 ($) with the other conditions being equal. In the questionnaire, 9 cases of the combinations of the price and the earthquake frequency are presented, and each person chooses its preference level from the 7 grades for each combination. We distributed this questionnaire sheet to the people owning their houses in the suburb area of Tokyo, and totally 400 answer sheets are collected back. The answer data has been processed by following the method of successive categories which is widely used in the experimental psychology to compose an interval scale preference index y (Torgerson 1958). Through these research process, $OBS = \{X_1, X_2, ..., X_{400}\}$ where $X_i = [x_{i1}, x_{i2}, y_i]$ is obtained.

The proposed method has been applied to figure out a first-principle-based model equation $y = f(x_1, x_2)$, where x_1 and x_2 are ratio scale quantities. Hence, $RQ = \{x_1, x_2\}, IQ = \{y\}$ and $AQ = \phi$. Because IQ contains only one quantity, steps $(1-1)$ and $(1-2)$ are skipped, and the quasi-bi-variate fitting of 1st order approximation is applied to $RQ = RQ + IQ = \{x_1, x_2, y\}$ in step $(2-1)$. First, the top formula in Table 1 is tested for the relation between x_1 and x_2, and

$$x_1 = a(y)x_2^{-0.25}$$

is obtained. Next, the second and third formulae are tested for x_1 and y. Then,

$$\begin{aligned} y &= a(x_2)x_1^{-0.23} + b(x_2) \\ y &= 0.62 \log x_1 + b(x_2) \end{aligned}$$

have been identified, respectively. Both of them were accepted by the F-test. Similar search has been made for x_2 and y, and

$$\begin{aligned} y &= a(x_1)x_2^{0.026} + b(x_1) \\ y &= 0.34 \log x_2 + b(x_1) \end{aligned}$$

are derived. In step $(2-2)$, the triplet-test among $\{x_1, x_2, y\}$ is conducted. As the admissible formulae in Theorem 2 for these quantities are limited to $y = bx_1^{a_1}x_2^{a_2} + c$ and $y = a_1 \log x_1 + a_2 \log x_2 + c$, the consistency among the coefficients obtained in step $(2-1)$ are checked by following these formulae. As a result, the consistency has been confirmed for both. Consequently, we obtained the following two candidates.

$$\begin{aligned} y &= 0.62 \log x_1 + 0.34 \log x_2 - 2.9 \\ &\quad (AIC = [-1537, -1326, -1121]) \qquad (12) \\ y &= -0.61x_1^{-0.23}x_2^{0.026} + 3.2 \\ &\quad (AIC = [-810, -599, -394]) \qquad (13) \end{aligned}$$

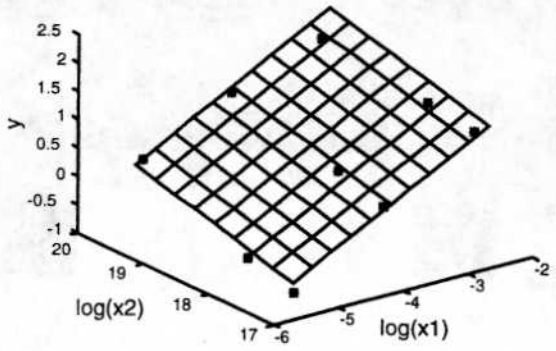

Figure 3: Plot of Eq.(12)

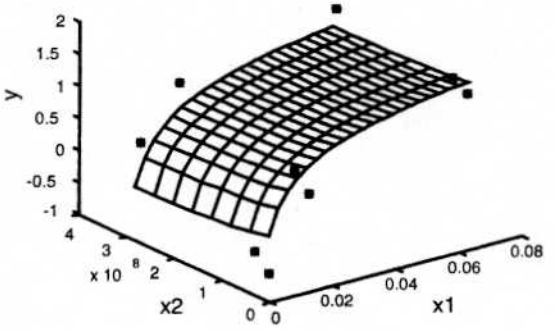

Figure 4: Plot of Eq.(13)

Step (3) is skipped, since the other quantities to be merged do not exist. In step (4), the value and its uncertainty range of AIC are evaluated for each candidate. The expression $AIC = [L, M, U]$ represents the lower bound L, the expected value M and the upper bound U of the AIC. Since the former Eq.(12) has the smaller AIC value, the former is preferred. Moreover, because the upper bound of Eq.(12) is smaller than the lower bound of Eq.(13), Eq.(12) is uniquely chosen to be a model equation $y = f(x_1, x_2)$. Figures 2 and 3 shows the plots of the two equation curves together with the average values ■ of the 400 answered preference level y for the 9 cases of the price and the earthquake frequency. The high accuracy of the Eq.(12) is clearly observed in those figures. Eq.(12) can evaluate the subjective preference in the accuracy of almost ± 1 levels of the questionnaire from the values of x_1 and x_2.

Discussion and Related Work

A scientific discovery system called LAGRANGE (Dzeroski and Todorovski 1994) is also applicable to the condition of the passive observation. It uses the principles of ILP and generate/test. Though no equation classes are presumed in this approach, many spurious solutions can be derived due to the weakness of the

search heuristics. Also, it indicates high computational complexity. TETRAD (Glymour 1995) is another system to identify the models of the objective system from the passive observation. Its basic framework takes the bottom up modeling approach. However, the class of the model formulae is presumed such as linear expressions. In contrast, the method proposed in this paper has a strong mathematical background to characterize first principle equations. Moreover, it has a high applicability to the passive observation data, while maintaining the flexibility of the bottom up modeling approach taken by the conventional scientific discovery systems.

The source of the advantage of our proposed method is the systematic use of the constraints of scale-types and identity with the approximation in quasi-bi-variate fitting. This is considered to be a typical example that Ginsberg claimed (Ginsberg and Geddis 1991). He claimed that any domain-dependent control rules can be replaced with a domain-independent control rules and modal sentences describing the structure of the search space. The knowledge of the scale-types and the quasi-bi-variate approximation have been implicitly used by scientists as domain-dependent control rules of their reasoning. In our work, these rules have been replaced as Ginsberg claimed. The constraints and the approximation are formalized as generic domain-independent control rules applicable to any objective system represented by numerical quantities. The modal knowledge required to control the reasoning by these generic rules is concentrated on the scale-type information of each quantity and the empirical quasi-bi-variate relation. On the other hand, Minton argued that in many cases, domain-dependent control rules cannot, in a practical sense, be derived due to the complexity of the reasoning that would be required (Minton 1996). Since the concepts such as the scale-types and the quasi-bi-variate approximation have been established based on massive experience of the scientists for hundreds of years, his argument also holds.

Conclusion

In this paper, quasi-bi-variate fitting, an extension of the bi-variate fitting based on an polynomial approximation, has been proposed. This extension enables to handle the new class of the problem to discover the law equations under the passive observation environment. Its basic performances in terms of computation time and noise robustness have been evaluated through simulations. The evaluation indicates the satisfactory performance to discover the model equation based on the first principle of objective system of moderately large size under practical noise levels. Finally, a real application to discover a law equation in socio-psychology was demonstrated, and its practicality has been readily confirmed.

References

B. Falkenhainer and K. Forbus. Compositional modeling: finding the right model for the job. *Artificial Intelligence*, Vol.51, pages 1–3, 1991.

B. Falkenhainer et al. CML: A Compositional Modeling Language. *Technical report KSL-94-16*, Knowledge Systems Laboratory, Stanford University, 1994.

Y. Iwasaki et al. A Web-Based Compositional Modeling System for Sharing of Physical Knowledge. *Proceedings of IJCAI'97: Fifteenth International Joint Conference on Artificial Intelligence*, Vol.1, pages 494–500, 1997.

P.J. Mosterman and G. Biswas. Formal Specifications for Hybrid Dynamical Systems. *Proceedings of IJCAI'97: Fifteenth International Joint Conference on Artificial Intelligence*, Vol.1, pages 568–573, 1997.

Y. Kitamura, M. Ikada and R. Mizoguchi. A Causal Time Ontology for Qualitative Reasoning. *Proceedings of IJCAI'97: Fifteenth International Joint Conference on Artificial Intelligence*, Vol.1, pages 501–506, 1997.

N. Smith. A New Architecture for Automated Modeling. *Proceedings of AAAI'98: Fifteenth National Conference on Artificial Intelligence*, pages 225–231, 1998.

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.

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

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.

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

T. Washio and H. Motoda. Discovering Admissible Models of Complex Systems Based on Scale-Types and Identity Constraints, In *Proceedings of IJCAI-97: Fifteenth International Joint Conference on Artificial Intelligence*, Vol.2, pages 810-817, Nagoya, 1997.

T. Washio and H. Motoda. Discovering Admissible Simultaneous Equations of Large Scale Systems, In *Proceedings of AAAI'98: Fifteenth National Conference on Artificial Intelligence*, pages 189-196, Madison, 1998.

T. Washio and H. Motoda. Compositional Law Discovery Based on Scale Cognition of Feature Quanti-

ties, In *Cognitive Science* (In Japanese), Vol.5, No.2, pages 80-94, Tokyo, 1998.

L. Ljung. *System Identification*, P T R Prentice-Hall, 1987.

S.S. Stevens. On the Theory of Scales of Measurement, In *Science*, Vol.103, No.2684, pages 677-680, 1946.

R.D. Luce. On the Possible Psychological Laws, In *The Psychological Review*, Vol.66, No.2, pages 81-95, 1959.

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

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

H. Akaike. A new look at the Bayes procedure. In *Biometrika*, Vol.65, pages 53–59, 1978.

W.S. Torgerson. In *Theory and Methods of Scaling*, N.Y.: J. Wiley, 1958.

S. Dzeroski and L. Todorovski. Discovering Dynamics: From Inductive Logic Programing to Machine Discovery. In *Journal of Intelligent Information Systems*, pages 1–20, Boston, Kluwer Academic Publishers, 1994.

C. Glymour. Available Technology for Discovering Causal Models, Building Bayes Nets, and Selecting Predictors: The TREAD II Program, In *Proc. of the First International Conference on Knowledge Discovery & Data Mining*, pages 130–135, Montreal Quebec, Canada, 1995.

M.L. Ginsberg and D.F. Geddis. Is there any Need for Domain-Dependent Control Information?, In *Proc. of Ninth National Conference on Artificail Intelligence*, pages 452–457, 1991.

S. Minton. Is there any Need for Domain-Dependent Control Information? A Reply, In *Proc. of Thirteenth National Conference on Artificail Intelligence*, pages 855–862, 1996.