SINBAD AUTOMATION OF SCIENTIFIC DISCOVERY: FROM FACTOR ANALYSIS TO THEORY SYNTHESIS

Olcay Kurşun kursun@cs.ucf.edu
Oleg V. Favorov favorov@bme.unc.edu

School of Computer Science
University of Central Florida
Orlando, FL 32816, USA

Abstract

Modern science is turning to progressively more complex and data-rich subjects, which challenge the existing methods of data analysis and interpretation. Consequently, there is a pressing need for development of ever more powerful methods of extracting order from complex data and for automation of all steps of the scientific process. *Virtual Scientist* is a set of computational procedures that automate the method of inductive inference to derive a theory from observational data dominated by nonlinear regularities. The procedures utilize SINBAD – a novel computational method of nonlinear factor analysis that is based on the principle of maximization of mutual information among non-overlapping sources, yielding higher-order features of the data that reveal hidden causal factors controlling the observed phenomena. The procedures build a theory of the studied subject by finding inferentially useful hidden factors, learning interdependencies among its variables, reconstructing its functional organization, and describing it by a concise graph of inferential relations among its variables. The graph is a quantitative model of the studied subject, capable of performing elaborate deductive inferences and explaining behaviors of the observed variables by behaviors of other such variables and discovered hidden factors. The set of *Virtual Scientist* procedures is a powerful analytical and theory-building tool designed to be used in research of complex scientific problems characterized by multivariate and nonlinear relations.

Keywords: blind source separation, causal relations, concept acquisition, IMAX, knowledge representation, nonlinear factor analysis, Virtual Scientist

Introduction

Modern science turns to progressively more complex and challenging subjects across many fields – medicine, neuroscience, genomics and related fields, ecology, economics, climatology, cosmology, etc. This expansion of scientific inquiry into until recently inaccessible territories is brought about by ever growing advances in computer and sensor technologies, which enable collection of large amounts of groundbreaking novel experimental and observational data. On the other hand, the new subjects also address more complex phenomena that reflect causal relations among large numbers of relevant factors with only limited, if any, opportunities for experimental control and manipulation. The growing size and

complexity of collected data demand progressively more sophisticated analytical and theory-building methods, methods that can process large amounts of raw data and extract intricate, deeply hidden order (Mjolsness and DeCoste 2001).

This paper describes a set of computational procedures that were developed to automate analysis and theory-building process for particularly difficult research problems that: (1) involve complex – multivariate and prominently nonlinear – interrelations among the measured entities; but (2) can be approached only, or mostly, through observation, without benefits of experimental manipulation of conditions; and (3) observations can be made only of spatial, but not temporal, patterns of events. Figure 1 describes a prototypical example of such a research problem. This example will be used throughout the paper to illustrate practical implementation of the proposed set of theory-building procedures.

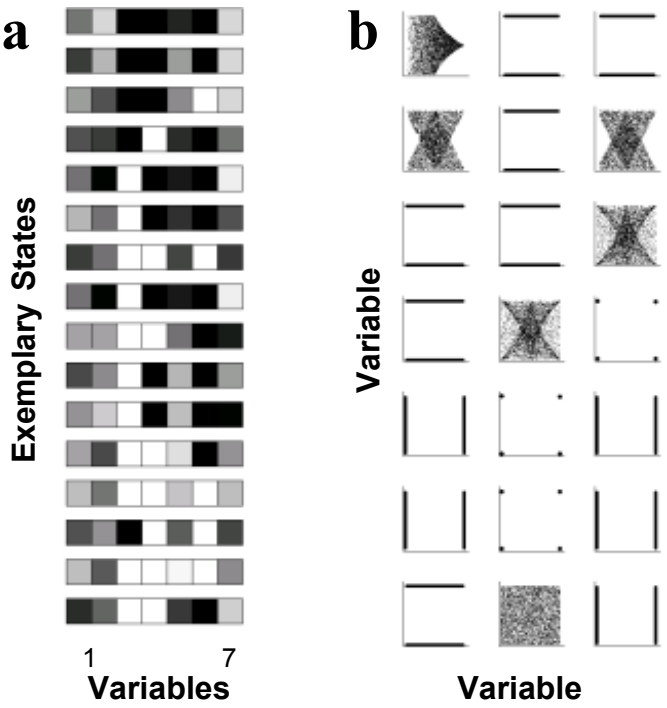

Figure 1. A research project to build a theory solely from snapshot observations of an unknown dynamical system. The dynamical system used in this example is a prototypical mathematical model of a well-known class of physical systems, but its identity will be revealed only at the end of the exercise, so as to avoid using the prior knowledge of the system in interpreting the observational data. In this exercise, designed to demonstrate the operation of the paper's theory-building procedures, the studied system's state is evaluated by taking seven different measurements, referred to as the "observed variables" V_1 … V_7. The seven measurements taken simultaneously at any given moment constitute a single "observation"; and the theory is to be built from large numbers of such observations. The observations are not continuous, but are collected randomly without any temporal order. **(a)** 16 exemplary observations of the system's state. The value of each observed variable is grayscale-coded (from white = 0 to black = 1). **(b)** Plots of pair-wise relations among the observed variables. Each variable is plotted as a function of every other variable. Some variables are binary, others are continuous. These plots reveal no clear order among the variables, suggesting that the variables might be only weakly interdependent, or the order might be hidden in relations that are multivariate and possibly nonlinear.

Our approach to theory building is based on consideration of the importance of hidden causal factors (Clark and Thornton 1997; Favorov and Ryder 2004; Ryder 2004). That is, the behaviors of the observed entities, or *variables*, might be controlled by some unknown factors (i.e., they are not among the variables that are observed in the study). Such hidden factors can vary greatly in the extent of their impact on the studied dynamical system, from factors that impact behaviors of just one or a few observed variables to factors that impact behaviors of most or even all of the observed variables. Hidden factors that are reflected in the behaviors of sufficiently large numbers of the observed variables can, in principle, be extracted from them through some computation. And they should be extracted: their reflectance in behaviors of multiple observed variables implies that these factors play prominent and central roles in the functioning of the studied dynamical system. This makes the knowledge of such hidden, but extractable factors crucial to theory building.

Thus, the theory-building process should first analyze the behaviors of the observed variables to infer the presence of as many hidden factors as possible and learn how to compute them from the observed variables. In the next step, the theory-building process should learn orderly relations among all the observed variables and inferred factors. What will emerge out of this process is a *theory* – a detailed quantitative tracing of connections among the components of the studied dynamical system. This is a traditional approach to theory building (with hidden-factor learning recognized as conceptualization); its penetration of a subject depends on the ability of the employed analytical methods to discover hidden causal factors. Recently we have developed SINBAD – a novel computational method of nonlinear factor analysis designed specifically for finding deeply hidden factors (Ryder and Favorov 2001; Kursun and Favorov 2002; Favorov et al. 2003; Kursun and Favorov 2003; Joshi et al. 2003; Favorov and Ryder 2004). This method enabled us to design a set of computational procedures for building theories of particularly difficult research subjects characterized by multivariate and nonlinear relations. Reflecting on the fact that these procedures perform a quintessential work of a scientist, we named this set a "Virtual Scientist".

SINBAD Method for Finding Hidden Factors

SINBAD belongs to the class of unsupervised learning algorithms that are based on the principle of maximization of mutual information among disjoint sources of information, developed by Becker as *Imax* (1995, 1996, 1999; Becker and Hinton 1992). According to Becker and Hinton (1992), hidden factors can be discovered through a search for different, but nevertheless highly correlated functions of any kind over non-overlapping subsets of the available variables. Such *correlated functions* must have a reason for their statistical interdependence, a causal source in the domain of the data. Therefore, the correlated functions over different subsets of variables express a previously unrecognized feature (a hidden factor) that is responsible for the correlation (Becker and Hinton 1992; Phillips and Singer 1997; Ryder and Favorov 2001; Favorov and Ryder 2004). For examples of applications of this approach, such as discovery of surfaces in stereograms, recognition of moving objects, speaker-independent vowel recognition, see Becker and Hinton (1992), Becker (1995, 1996, 1999), Stone (1996).

Becker's Imax method works by maximizing Shannon's mutual information measure among the outputs of learning modules receiving different subsets of input variables. Unfortunately, the method is computationally complex and requires making various restrictive assumptions about the output distributions of the modules to get a tractable expression for the mutual information between two continuous signals (Becker 1996). SINBAD is a simpler approach that works by minimizing the mean-square-error among the outputs of the learning modules. SINBAD method avoids trivial minimization of this error by forcing the output signals to have nonzero variance (otherwise, if all modules give the same

constant output, the error would be trivially minimized).

SINBAD architecture is illustrated in Figure 2. It contains two (or more) learning modules, each in a form of a backprop net (an error backpropagation network of Rumelhart et al. 1986). The backprop nets receive different, non-overlapping sets of input channels, while their outputs are added together to produce a final output. This final output is used as a teaching signal for each of the backprop nets, which means that the nets are set up to teach each other to produce maximally *correlated* outputs in response to their *different* inputs. As a result, the nets tune to the causal source (a hidden factor) responsible for the correlation.

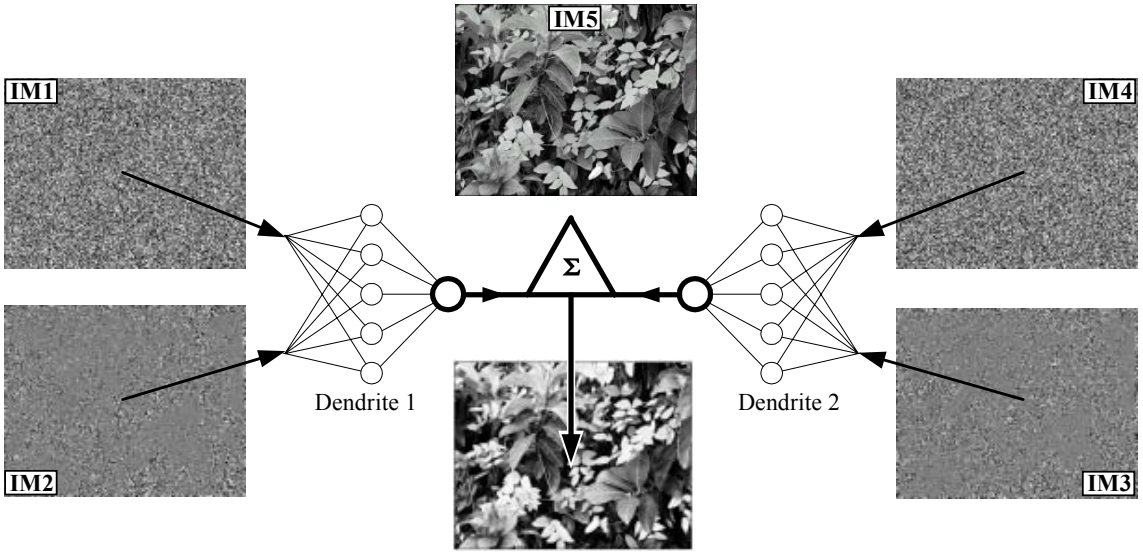

Figure 2. The SINBAD *cell* with two *dendrites*. Each dendrite is an error backpropagation network with one output unit and a single layer of hidden units. SINBAD cell is a powerful discoverer of hidden factors. For a demonstration we created a dataset of 4 images. Images 1 and 2 share mutual information with images 3 and 4 in a form of a hidden image 5. That is, each pixel in image 2 is a nonlinear function of identically located pixels in images 1 and 5 (scaled between 0 and 1): $IM2_{xy} = IM1_{xy} + IM5_{xy} - 2 \cdot IM1_{xy} \cdot IM5_{xy}$. The same holds for image 3 – each pixel in image 3 is the analogous function of identically located pixels in images 4 and, again, 5. Thus, image 5 is a hidden factor determining the contents of images 2 and 3. To discover this hidden factor, SINBAD cell is trained on the values of identically located pixels in images 1, 2, 3, 4. Pixels from images 1 and 2 are given to Dendrite 1, pixels from images 3 and 4 are given to Dendrite 2. SINBAD cell gradually learns – without any guidance – to output the value of the identically located pixel in the hidden image 5 (compare IM5 with the image below the cell). In conclusion, although image 5 was so well hidden in images 2 and 3 as to be invisible there, SINBAD cell nevertheless was able to easily detect its presence and extract its content. In contrast, linear factor analysis methods, including PCA and ICA (Hyvarinen et al. 2001), will not work here because of nonlinearity of the underlying relationship. And even nonlinear factor extraction methods – Nonlinear PCA (Kramer 1991) and Nonlinear ICA (Lappalainen and Honkela 2000; Valpola et al. 2001) – could not accomplish this task, reasons for which will be discussed later.

SINBAD was originally developed as a model of a single neuron in the cerebral cortex (Ryder and Favorov 2001; Favorov et al. 2003; Favorov and Ryder 2004). SINBAD is an acronym for *Set of INteracting BAckpropagating Dendrites*. According to SINBAD hypothesis, the basic function of a cortical neuron is to discover and represent one of the hidden factors in its sensory environment. This task is proposed to be accomplished by endowing each of several dendrites that originate from a neuron's body with functional capabilities comparable to those of a backprop net. Although the neurobiological origins of SINBAD algorithm are not important for the present subject, in this paper we continue to call the entire setup a "SINBAD cell" and each backprop net a "dendrite".

The detailed formulation of the SINBAD cell can be found in (Ryder and Favorov 2001). As a brief description of its implementation in this paper, the activity of a hidden unit h in dendrite d is computed as a sigmoid function of the activities of its input sources:

$$H_{d,h} = \tanh(\sum_i w_{d,i,h} \cdot A_{d,i}), \quad (1)$$

where $A_{d,i}$ is the activity of input source d,i and $w_{d,i,h}$ is the weight of its connection onto the hidden unit h of dendrite d. The activity of the output unit, i.e. the output of dendrite d, is:

$$D_d = \sum_h w_{d,h} \cdot H_{d,h}, \quad (2)$$

where $w_{d,h}$ is the weight of the connection from the hidden unit d,h to the output unit. The outputs of the two dendrites are summated to produce the cell's output:

$$A = D_1 + D_2. \quad (3)$$

The cell's output A is the principal contributor to the training signal T; it is used to adjust the weights of connections on the three dendrites. Additional factors contributing to the training signal are: (1) the average output activity of the cell, $\bar{A}$, driving the cell to have $\bar{A} = 0$; and (2) deviation of the current output activity from the average, $A - \bar{A}$, designed to expand the dynamic range of output values. Thus,

$$T = A - \alpha \cdot \overline{A} + \beta \cdot (A - \overline{A}), \quad (4)$$

where α and β are scaling coefficients. Coefficient β is determined by the variability of the output activity: smaller the variability, greater the value of β. It is computed as:

$$\beta = \left[\beta_{\max} - \gamma \cdot \overline{|A - \overline{A}|}\right]^+, \quad (5)$$

where $\beta_{\max}$ and γ are controlling parameters, and $[\cdot]^+$ indicates that if the quantity is negative, the value is to be taken as zero. The connections of the hidden units are adjusted according to the error backpropagation algorithm of Rumelhart et al. (1986). Specifically, the error signals δ_d is first computed for the two dendrites as:

$$\delta_d = T - 2 \cdot D_d. \quad (6)$$

For the hidden units, δ_d is backpropagated as:

$$\delta_{d,h} = \delta_d \cdot w_{d,h} \cdot (1 - H_{d,h}^2). \quad (7)$$

Connection weights are adjusted by:

$$\Delta w_{d,i,h} = \mu_i \cdot A_{d,i} \cdot \delta_{d,h} \quad \text{and} \quad \Delta w_{d,h} = \mu_h \cdot H_{d,h} \cdot \delta_d, \quad (8)$$

where μ_i and μ_h are learning rate constants for the input and hidden unit connections, respectively.

Finding Minimal Sets of Related Variables

For a SINBAD cell to be able to find a hidden factor, its dendrites must be given *different but related* groups of input variables; i.e., groups with no input variable(s) in common, but with *implicit* information about the same hidden factor. In some cases such related groups might be obvious or easy to guess. For example, in binocular vision, receptors in the two eyes obviously make up two different but related groups of sensory variables with mutual information about the third visual dimension (Becker and Hinton 1992). In general, however, such knowledge is not available and the related groups of variables can only be found by trial and error. In the absence of any heuristic, such a search is exponential – if we have a total of N observed variables, then we can separate them into $O(2^N)$ possible pairs of groups. Each such pair will have to be tested for whether a hidden factor can be extracted from it by a SINBAD cell. Overwhelming majority of such blindly tried partitions of the observed variables into pairs of groups will produce unrelated or insufficiently related groups, which will fail to yield any hidden factors (i.e., SINBAD cell's dendrites will fail to learn to produce correlated outputs).

Because it is exponential, an exhaustive search among even small (e.g., 20) numbers of variables is prohibitively expensive and must be minimized as much as possible. To explain our approach to such minimization, suppose that among the entire set of N observed variables, $V_1 \ldots V_N$, variable V_t (a "target" variable) can be computed, with minimal error ε, from a subset of variables $V_a \ldots V_k$ and a hidden factor H: $V_t = f(V_a ... V_k, H) + \varepsilon$. Suppose that this relationship is fully or at least mostly reversible, allowing H to be computed, with some minor error ε^*, from V_t and $V_a \ldots V_k$: $H = f^*(V_t, V_a ... V_k) + \varepsilon^*$. Suppose also that H can be computed from the observed variables $V_l \ldots V_p$: $H = h(V_l ... V_p)$.

This means that if we set up a backprop net with inputs from all the observed variables $V_1 \ldots V_N$, except V_t, we might be able to train that net to output target variable V_t with an error close to ε (since the input variables include $V_a \ldots V_k$ and $V_l \ldots V_p$). This also means that the number of inputs to the backprop net can be reduced without any loss of its performance, *as long as the discarded variables do not include* $V_a \ldots V_k$, $V_l \ldots V_p$. In our terminology, the set of the observed variables $V_a \ldots V_k$, $V_l \ldots V_p$ is the "Predictive Set" of the "target" variable V_t. A target variable together with its Predictive Set make up a "Minimal Set of Related Variables."

The Minimal Set defined for target variable V_t contains all the observed variables needed for finding hidden factor H, and no other variables: it is the smallest possible set for finding H. If we test all possible two-group partitions of this set, one of the partitions will be $\{V_t, V_a \ldots V_k\}$ vs. $\{V_l \ldots V_p\}$. With one dendrite of a SINBAD cell receiving V_t, $V_a \ldots V_k$ and the other dendrite receiving $V_l \ldots V_p$, the two dendrites will learn to produce correlated outputs by computing

$$f^*(V_t, V_a \ldots V_k) \approx h(V_l \ldots V_p) = H. \quad (9)$$

The great benefit of confining an exhaustive search of variable partitions to a Minimal Set, rather than doing it over all the observed variables, is in drastic reduction of the number of partitions that will require testing for hidden factors.

To summarize, for SINBAD search for hidden factors to be thorough while accomplishable in reasonable time, Minimal Sets of related variables should be identified first. Then each Minimal Set should be partitioned in all possible ways into pairs of subsets. Each such pair of subsets should be tested, using SINBAD cell method, for whether it will yield a hidden factor.

To identify Minimal Sets, we can use each observed variable in turn as a target variable. For each Minimal Set, the first step is to determine how accurately its target variable can be predicted from all the available observed variables. The next step is to determine which of the variables contribute to this prediction and which ones can be dropped without loss of prediction accuracy. A simple way to accomplish the first step is to set up a backprop net with inputs from all but the target observed variables

and train it on the target variable. However, this approach can under-perform or even fail if the relationship between the target variable and the other variables is orderly but not unique. For example, the available observed variables might only have information about how much the target variable deviates from its mean, but not in which direction (e.g., an ability to predict a line segment in an image from its context, but not whether it is darker or lighter than background). In such cases it will be necessary to identify the "regular" component of the target variable; i.e., such a derivative of the target variable that preserves maximal information about the variable, *and* that is also maximally predictable from the other available variables. Such a regular component of a variable might be the variable itself or another *unary* function of it.

The task of learning the regular component of the target variable and determining the accuracy of its prediction from the observed variables can be accomplished by using a SINBAD cell (Figure 3). For this task one dendrite is given all but the target observed variables and the other dendrite is given only the target variable. After a training period the second dendrite will learn to output the regular component of the target variable, while the first dendrite will learn to output the closest possible approximation of that component. The accuracy of this approximation can be measured by coefficient of determination (i.e., the squared coefficient of correlation of the outputs of the two dendrites).

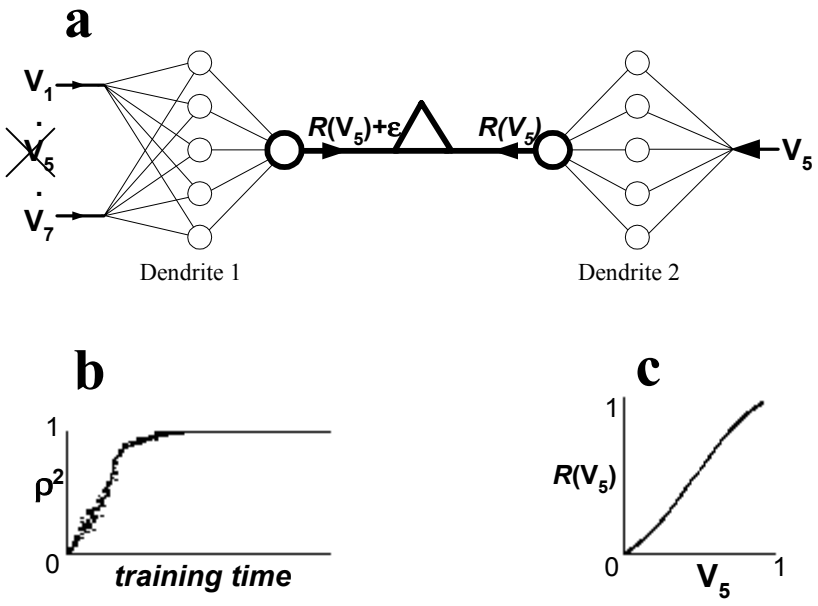

Figure 3. Learning the regular component of a target variable and the accuracy of its prediction from the observed variables. Variable V_5 from the research project described in Figure 1 is used as an example. **(a)** SINBAD cell with dendrite 2 receiving V_5 and dendrite 1 receiving all the other six observed variables. $R(V_5)$ – the regular component of V_5. After a learning period, dendrite 2 will output $R(V_5)$ and dendrite 1 will output $R(V_5)$ + a minimal error. **(b)** Time-course of learning. ρ – correlation coefficient between outputs of dendrites 1 and 2. ρ^2, coefficient of determination, is plotted as a function of training time (i.e., the number of training trials). Note that the two dendrites learned to match each other's output almost perfectly ($\rho^2 = 0.999$). **(c)** The regular component of V_5. The plot shows that the regular component of V_5 is the variable itself.

To determine which variables are really used by the first dendrite in predicting the regular component of the target variable, we use a version of *sequential backward elimination* technique (Bishop 1995). Other network pruning techniques, such as weight elimination (Hanson and Pratt 1989; Lang and Hinton 1989), optimal brain damage (LeCun et al. 1990), or optimal brain surgeon (Hassibi and Stork 1993) are possible alternatives. The technique works as follows: starting with the first dendrite connected to all but the target observed variables, we remove those variables one at a time and each time re-train the dendrite. At this stage, learning in the second dendrite must be stopped, so that the training signal for the first dendrite will remain to be the regular component of the target variable, already found by the second dendrite. In effect, the SINBAD cell is reduced here to a single backprop net trained on the regular component. If correlation between the two dendrites declines as a result of the removal of an input variable from the first dendrite, that means that the removed variable was useful for predicting the regular component of the target variable and should be restored. If dendritic correlation does not decline, it means that the removed variable was not relevant and should not be restored (Figure 4).

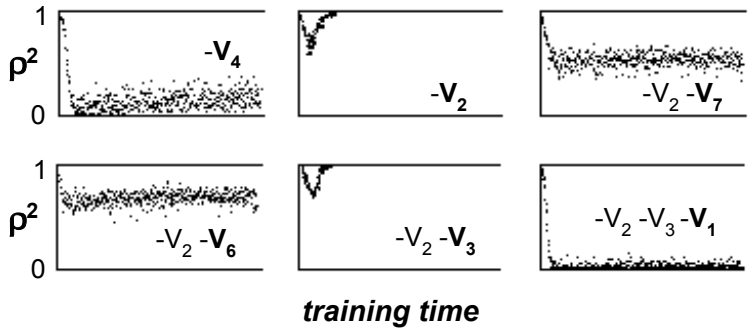

Figure 4. Finding a Minimal Set of related variables, with V_5 as a target variable. Dendrite 1 of the SINBAD cell in Figure 3 is trained on $R(V_5)$. Each plot shows the time-course of dendrite 1 learning after removal of some of the input variables. For a benchmark, the horizontal line at the top of each plot shows the magnitude of the coefficient of determination (ρ^2) between $R(V_5)$ and the output of dendrite 1 when it had all six input variables, $V_1 - V_4$, V_6, V_7. **Panels from left to right:** Variable V_4 was removed from dendrite 1 first. Correlation between $R(V_5)$ and dendrite 1 output ("dendritic correlation") dropped almost to 0 and never recovered, indicating that V_4 is crucial for predicting $R(V_5)$ and should be kept as an input to dendrite 1. V_2 was removed next (next panel). Dendritic correlation declined only transiently and made a quick and complete recovery to the original level, indicating that V_2 is not needed for predicting $R(V_5)$. Next, V_2 and V_7 were removed, and since dendritic correlation dropped permanently, V_7 was judged to be necessary for $R(V_5)$ prediction. The same conclusion was reached for V_6. However, removing V_2 and V_3 did not reduce dendritic correlation, indicating that V_3 can be discarded. Finally, V_1 was found to be needed. Based on this series of tests, $R(V_5)$ can be best predicted from the observed variables V_1, V_4, V_6, V_7. Therefore, this set of variables is a predictive set of V_5 and together they make up a Minimal Set $\{V_1\ V_4\ V_5\ V_6\ V_7\}$.

To be more precise, when we test whether a variable contributes to the prediction of a target variable, we compare the accuracy of the prediction with and without that variable. If the difference is less than a predetermined *threshold*, then we conclude that the variable is not needed and remove it from the dendrite. The variables that remain still connected to the first dendrite after testing all of them constitute the *predictive set* of the target variable. The threshold, T, determines which and how many variables are chosen for the predictive set of the target variable. Let P_T denote the predictive set of the target variable found by the sequential backward selection algorithm with threshold T, where $|P_T|$ is the size (cardinality) of the predictive set, and ρ^2_T is the coefficient of determination between the regular component of the target variable and its prediction by the set P_T. We want to use the threshold that yields the predictive set of the smallest size, but with the maximal prediction of the target variable. Formally, we are looking for a T value such that $|P_{T+\xi}| \leq |P_T| << |P_{T-\xi}|$, but $\rho^2_{T+\xi} << \rho^2_T \approx \rho^2_{T-\xi} \approx \rho^2_0$, where ξ is a small number and ρ^2_0 is the coefficient of determination using all variables.

A given target variable and its predictive set constitute one Minimal Set of related variables. We identify multiple such Minimal Sets, using every observed variable as a target variable (Table 1). It is a common occurrence that Minimal Sets identified using different variables as targets will have identical compositions. For example, in Table 1, sets 1, 4, and 5 are identical, and so are sets 2 and 3, and sets 6 and 7. As a result of such rediscoveries of the same Minimal Sets, the number of distinct Minimal Sets identified in a given study is likely to be smaller than the number of the observed variables.

Set #	**Target Variable**	**Minimal Predictive Set**	ρ^2
1	V1	V4 V5 V6 V7	.998
2	V2	V3 V4 V5 V6 V7	.921
3	V3	V2 V4 V5 V6 V7	.941
4	V4	V1 V5 V6 V7	.994
5	V5	V1 V4 V6 V7	.999
6	V6	V1 V2 V3 V7	.995
7	V7	V1 V2 V3 V6	.999

Table 1. List of Minimal Sets of related variables extracted from the observed variables $V_1 - V_7$.

Partitioning Minimal Sets to Find Hidden Variables

Once the Minimal Sets of related variables are identified, each of them should be partitioned in all possible ways and each partition should be tested for hidden variables. As stated above, if in a given predictive set, the role of a subset of variables is to contribute information about a hidden factor relevant to the prediction, then a SINBAD cell will learn this factor if one of its dendrites is given this subset of variables and the other dendrite is given the rest of the predictive set together with the target variable (eq. 9). Any other partitioning of the predictive set across the two dendrites will only reduce their mutual information, and will therefore reduce correlation of the dendrites' outputs (Figure 5).

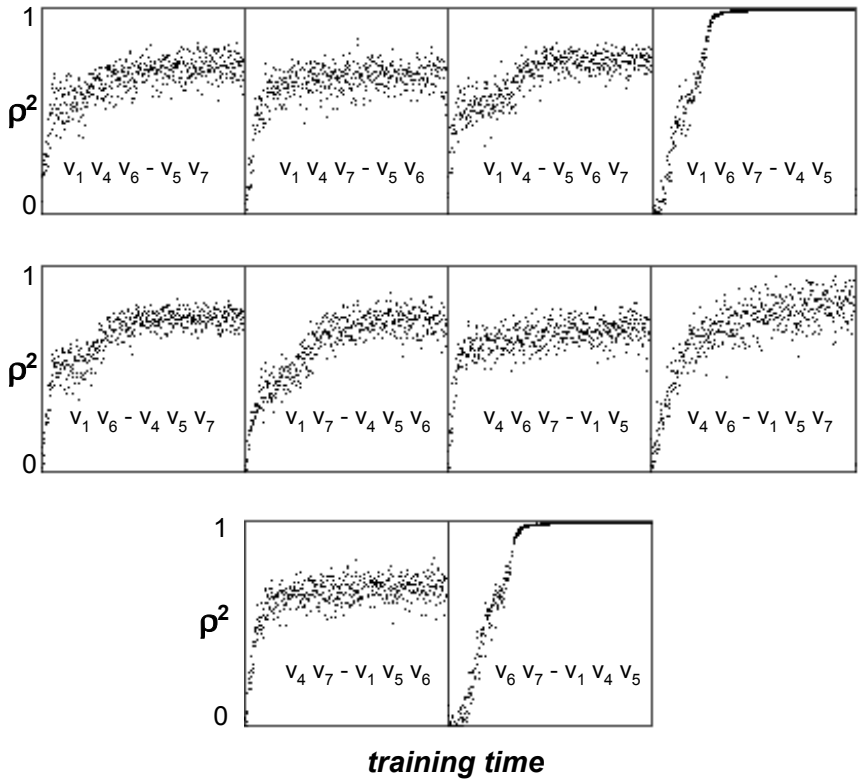

Figure 5. Extracting hidden factors from a Minimal Set of related variables. The analysis is performed on the Minimal Set $\{V_1\ V_4\ V_5\ V_6\ V_7\}$, identified in Figure 4 using V_5 as the target variable. This set of five variables can be partitioned in 10 different ways (with at least two variables on each side of a partition). A SINBAD cell was trained on each of these partitions, with time-course of learning shown in one of the panels. For example, in the top-left panel, dendrite 1 was given variables V_1, V_4, V_6 and dendrite 2 was given V_5 and V_7. The plot of dendritic correlation (ρ^2, coefficient of determination of the two dendrites' outputs) shows that the dendrites were unable to correlate their outputs as well as in some other panels. The best correlation was achieved in two panels, for partitions $V_1\ V_6\ V_7$ vs. $V_4\ V_5$ and $V_6\ V_7$ vs. $V_1\ V_4\ V_5$. For these partitions, $\rho^2 = 0.999$ – the same level as was achieved by the entire predictive set of V_5 against V_5 (see Figure 4). All other partitions produced clearly much lower dendritic correlations, indicating that they are unnatural. Therefore we take the two best partitions as apparently revealing two candidate hidden factors. These factors are new variables derived from the observed variables. They expand the list of variables characterizing the studied dynamical system; we label them as V_8 and V_9: $V_8 = f_8(V_1, V_6, V_7)$ and $V_9 = f_9(V_6, V_7)$. These derived variables are computed by the dendrite that does not receive the target variable (see eq. 9); the dendrite that receives the target variable computes only an approximation of the derived variable.

SINBAD testing of Minimal Set partitions identifies *candidates* for hidden factors. To make certain that these candidates are true hidden factors, they should be tested for their ability to substitute fully for the observed variables from which they were derived (Figure 6). If a candidate cannot predict the target variable as well as the variables from which it is computed, then this candidate should be discarded.

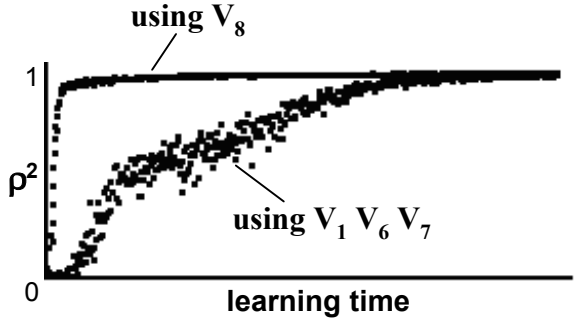

Figure 6. Testing a candidate hidden factor. According to Figure 4, $V_5 = f_5(V_1, V_4, V_6, V_7)$. In Figure 5, this Minimal Set was successfully partitioned into $\{V_1\ V_6\ V_7\}$ vs. $\{V_4\ V_5\}$, suggesting that $V_5 = f(V_4, V_8)$, where V_8 is taken as the output of the SINBAD dendrite in Figure 5 with inputs from V_1, V_6, V_7. In other words, $V_8 = f_8(V_1, V_6, V_7)$ might be a true hidden factor. To verify this suggestion, a backprop net (dendrite 1 used in Figure 3) is trained on V_5, receiving its input either from V_1, V_4, V_6, V_7 or, alternatively, from V_4, V_8. The plot shows the time-courses of learning under these two conditions, reaching maximal $\rho^2 = 0.999$. The plot reveals not only that V_5 prediction from V_8 is just as good as from V_1, V_6, V_7, but also that the net learns much faster with V_8, indicating that V_8 makes learning task much easier. Conclusion: V_8 is a true hidden factor.

Hidden factors with effects on multiple observed variables are likely to be discovered again and again while searching different Minimal Sets of related variables. For this reason, hidden factors derived from different Minimal Sets should be compared with each other (for example, compute cross-correlations among all the discovered factors) to identify any clones. Overall, a single Minimal Set can yield multiple hidden factors and different Minimal Sets can yield the same hidden factor (Table 2).

Minimal Set	Successful Partitions	Hidden Factor
V1 V4 V5 V6 V7	V4 V5 – V1 V6 V7 V6 V7 – V1 V4 V5	V8 V9
V2 V3 V4 V5 V6 V7	V4 V5 – V2 V3 V6 V7 V6 V7 – V2 V3 V4 V5 V2 V3 – V4 V5 V6 V7	V8 V9 V10
V1 V2 V3 V6 V7	V6 V7 – V1 V2 V3 V2 V3 – V1 V6 V7	V9 V10

Table 2. Hidden factors derived from all the Minimal Sets listed in Table 1.

Learning Orderly Relations

Hidden factors extracted from all the Minimal Sets extend, as derived variables, the list of variables with which we can characterize the studied dynamical system. The next task of theory building is to identify, quantify and express in a comprehensive way, the interdependencies among all the available – observed and derived – variables. To accomplish this goal, we need to identify the most direct relations among variables; these are the relations that involve minimal numbers of variables. Once such relations are known in sufficient numbers to form a more or less complete web, then interdependencies among all the available variables can be traced through chains of these direct relations.

In the process of identifying the Minimal Sets and the hidden factors, we have already learned a large number of relations, which computed one variable from a set of other variables. Together, they form the initial set of relations (Table 3). These relations give us a starting point for finding the most direct relations; i.e., relations involving minimal numbers of variables.

Relation #	Computed Variable	Predictive Set	ρ^2
1	1	4, 5, 6, 7	.998
2	2	3, 4, 5, 6, 7	.921
3	3	2, 4, 5, 6, 7	.941
4	4	1, 5, 6, 7	.994
5	5	1, 4, 6, 7	.999
6	6	1, 2, 3, 7	.995
7	7	1, 2, 3, 6	.999
8	8	4, 5	1.00
9	8	1, 6, 7	.999
10	8	2, 3, 6, 7	.998
11	9	6, 7	1.00
12	9	1, 4, 5	.999
13	9	1, 2, 3	.999
14	9	2, 3, 4, 5	.997
15	10	2, 3	1.00
16	10	4, 5, 6, 7	.998
17	10	1, 6, 7	.999

Table 3. List (compiled from all the relations listed in Tables 1 and 2) of all the ways by which each variable was learned, in the course of finding Minimal Sets and hidden factors, to be computed from other variables.

The basic idea is to substitute a group of variables in a known relation with a single variable. For example, we notice, in Table 3, that in relation #4, V_4 is computed from V_1, V_5, V_6, V_7, but that V_1, V_6, V_7 are also used to compute V_8 (relation #9). This observation raises a possibility that the role of V_1, V_6, V_7 in relation #4 is to compute V_8, suggesting that relation #4 can be simplified to that of $V_4 = f(V_5, V_8) + \varepsilon$. An alternative possibility, however, is suggested by relation #17: in this relation V_1, V_6, V_7 are used to compute V_{10}, not V_8. Therefore, a possible role of V_1, V_6, V_7 in relation #4 might instead be to compute V_{10}, allowing relation #4 to be simplified to that of $V_4 = f(V_5, V_{10}) + \varepsilon$. We can test these alternative hypotheses by training a backprop net on inputs either from V_5, V_8 or from V_5, V_{10}, with V_4 as the training

signal. If the backprop net learns to perform on one of these inputs just as well as when learning on V_1, V_5, V_6, V_7, then we succeeded in simplifying relation #4 and finding a new and *more direct* relation. We will add this new relation to our list of known relations and look for other such opportunities.

The following algorithm automates the process of generating such substitutions and finding all the most direct relations. To give a definition of this recursive algorithm, Set I is a set of inferential relations of the type $S_i \rightarrow V_x$, where the state of variable V_x is inferred with a maximal degree of accuracy by performing an optimized computation (carried out by a backprop net) on the states of a set of variables S_i. At the start, Set I comprises all the relations learned by SINBAD dendrites in the course of finding Minimal Sets and hidden factors. This set is expanded by applying Rule 1 to it.

Rule 1: if I contains two relations $S_i \rightarrow V_x$ and $S_j \rightarrow V_y$ such that S_j is a subset of S_i, and if the prediction accuracy of V_x by $[\{V_y\} \cup (S_i - S_j)] \rightarrow V_x$ is not significantly worse than prediction accuracy of V_x by $S_i \rightarrow V_x$, then relation $[\{V_y\} \cup (S_i - S_j)] \rightarrow V_x$ is added to Set I.

Rule 1 is applied iteratively, expanding the size of Set I. The search for new relations continues until no new pair of relations satisfying the rule's conditions can be found anymore in I.

Next, among all the relations in I, those that were not simplified by Rule 1 at any time during the search are taken to be *the most direct relations*. They form a new set, D. A new rule, Rule 2, is applied iteratively to relations in Set D. This rule does not simplify relations; instead, it defines new relations by trading one variable in the already present relation for a different variable.

Rule 2: if D contains two relations $[\{V_x\} \cup S_i] \rightarrow V_y$ and $[\{V_y\} \cup S_j] \rightarrow V_z$, and if S_i is a subset of S_j or is the same as S_j, then relation $[\{V_x\} \cup S_j] \rightarrow V_z$ is added to Set D.

In the last part of the algorithm, when after its iterative applications, Rule 2 fails to find any new pairs of relations satisfying the rule's conditions, new relations are defined by reversing every relation in Set D and testing them for their predictive powers. The reason is that if we know a relation in which V_a and V_b successfully predict V_c, then we can expect that V_a can also be predicted, more or less accurately, from V_b and V_c, and so can V_b from V_a and V_c. The predictive accuracies of the reversed relations are evaluated by training backprop nets to implement them.

The product of this algorithm is a set of the most direct relations among the observed variables and the discovered factors, each relation executable by a backprop net. This is a very efficient searching algorithm. Applied, for example, to the set of 17 original relations in Table 3, it identified and tested on backprop nets 36 potential relations, 32 of which were accepted. The algorithm found 21 most direct relations, listed in Table 4. In comparison, if we searched for the most direct relations by testing, using backprop net training, all the possible relations among the 10 available variables, we would have to perform 5020 such tests.

Relation #	Computed Variable	Predictive Set	Relation #	Computed Variable	Predictive Set
1	1	8, 9	12	8	9, 10
2	1	8, 10	13	8	1, 10
3	1	9, 10	14	9	6, 7
4	2	3, 10	15	9	1, 8
5	3	2, 10	16	9	1, 10
6	4	5, 8	17	9	8, 10
7	5	4, 8	18	10	2, 3
8	6	7, 9	19	10	1, 9
9	7	6, 9	20	10	8, 9
10	8	4, 5	21	10	1, 8
11	8	1, 9			

Table 4. The most direct inferential relations among the seven observed and the three derived (hidden) variables, listing the sets of variables that are predictive of each of the 10 variables.

Formulating a Theory of the Studied Subject

The set of direct relations that Virtual Scientist procedures extract from the observed variables describe most explicitly the order discernable in the observed dynamical system (given that the only source of information about the system are the observed variables). For a full appreciation of the interdependencies among the variables, this set of relations should be considered as a unified graph, rather than as a disjointed list of separate relations. Different relations are linked by variables they have in common (such as, for example, when the same variable is predicted by one relation and is used for prediction by another relation). As a result of such overlaps, all the different direct relations are linked together into a single functional entity, a *web of inferential relations* (Figure 7).

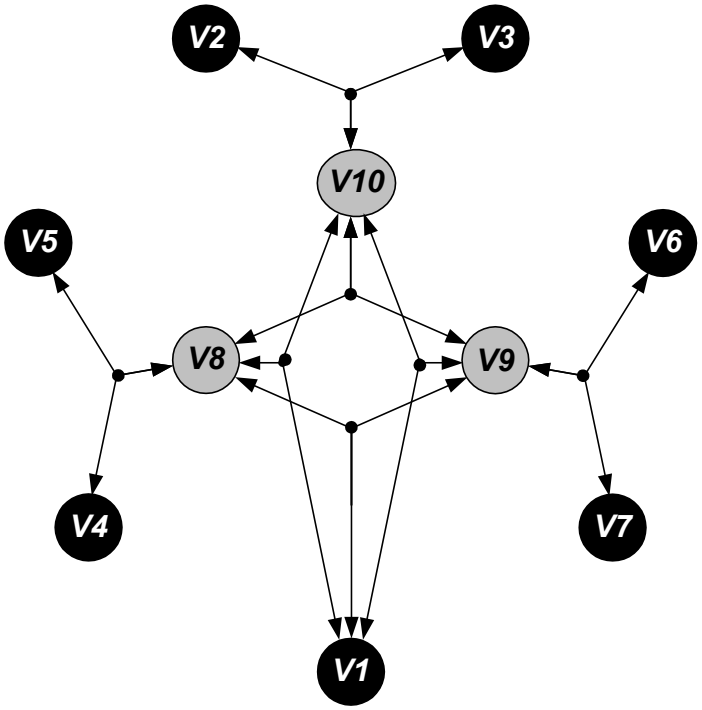

Figure 7. The web of connections among the observed variables and hidden factors discovered by Virtual Scientist in the studied dynamical system. The plot shows the seven observed variables (solid circles) and the three derived variables (shaded circles). All 21 direct relations listed in Table 4 are shown by lines connecting the variables. Note that the lines do not connect variables directly, but go through *hubs*, which tie several lines together. The presence of a hub indicates that a given variable does not have predictive significance for another variable just by itself, but in conjunction with one or more other variables joining it in the hub. Thus, each hub identifies a set of variables engaged together in a multivariate, typically nonlinear relation. An arrow emerging from a hub and pointing at a variable indicates that the state of that variable can be predicted with a significant degree of accuracy (in this plot, all have $\rho^2 > 0.9$) from a combination of the states of the other variables linked by the hub. Note also that multiple arrows converging on a variable indicate alternative sources of prediction and should *not* be viewed as additive in their effects. Several of the variables in the plot are engaged in more than one relationship and, therefore, can alternatively be predicted from more than one set of variables.

Viewed as a whole, this diagram of variables linked together by the web of connections provides a graphic representation of the order discovered by Virtual Scientist procedures in the studied dynamical system. By virtue of its power to unravel and explain behaviors of the variables in terms of other variables, this web of inferential relations offers a concise and comprehensive *theory* of the studied subject.

The web of inferential relations reveals all the pathways by which even remote interactions among distant variables can be traced through chains of variables connecting them. Furthermore, the usefulness of the web of relations extends beyond mapping functional interconnections among the variables. Each direct relation in the web has been learned – in the process of its discovery – by a separate backprop net and can, therefore, be used to make predictions. Also, with a direct relation involving, typically, only a small number of variables, such a low-dimensional relation can be visualized by plotting the involved variables against each other, so that the underlying mathematical form of this relation can then be appreciated. Finally, standard least-square approximation methods can be used to fit the data distribution with a suitable approximating function, thus expressing the relation mathematically. Fitting approximating functions to all the direct relations in the web, it might be possible to arrive at a mathematical description of the studied dynamical system in a form of a system of equations.

Whether expressed by formulae or by backprop nets, the direct relations together make up a quantitative model of the studied subject, which is the kind of *theory* that Virtual Scientist procedures extract from the observational data. Among the central contributions of such a theory are (i) derivation of explanatorily useful concepts of influential hidden factors, (ii) explanation of the behaviors of the observed variables from behaviors of other such variables and the conceptualized hidden factors, and (iii) elucidation of the functional organization of the studied dynamical system.

The web of inferential relations shown in Figure 7 is an example of such a theory, developed by Virtual Scientist procedures from the observational data described in Figure 1. These data were generated by a mathematical model of a well-known dynamical system, the Kitchen Sink. Sinks encountered at various times and in different kitchens can be viewed as a single dynamical device – the Sink – that varies its configuration (i.e., how pipes and valves are arranged) in different kitchens and varies its state (i.e., how much water is flowing and its temperature) at different times. To simplify matters, our mathematical model describes sinks that all receive water from two pipes and are controlled by the following five variables: *HC* indicates which of the two pipes carries hot/cold water (0 – the left pipe is hot and the right pipe is cold, 1 – vice versa), DIR_L and DIR_R are the directions in which the left and the right knobs should be turned to open the pipes (0 – clockwise, 1 – counterclockwise), and KP_L and KP_R are the radial positions of the two knobs. Sinks in different kitchens can have different DIR_L, DIR_R, and *HC*; KP_L and KP_R can vary in the same sink.

These variables determine the flows of water through the two pipes:

$$F_L = (1 - DIR_L - KP_L + 2 \cdot DIR_L \cdot KP_L)/2 \quad \text{and} \tag{10}$$

$$F_R = (1 - DIR_R - KP_R + 2 \cdot DIR_R \cdot KP_R)/2 . \tag{11}$$

In turn, the flows of water through the two pipes determine the total water outflow from the faucet and its temperature:

$$F_T = F_L + F_R \quad \text{and} \tag{12}$$

$$t^\circ = ((1 - HC) \cdot F_L + HC \cdot F_R)/(F_L + F_R) . \tag{13}$$

Collection of observational data was envisioned to involve observing a random sequence to "snapshots" of sinks in various configurations of pipes and valves, and with various knob positions and the resulting water outputs. The observed variables $V_1 \ldots V_7$, analyzed by Virtual Scientist procedures, had the following identities: $V_1 = F_T$, $V_2 = t°$, $V_3 = HC$, $V_4 = DIR_L$, $V_5 = KP_L$, $V_6 = DIR_R$, $V_7 = KP_R$. Importantly, two sink variables of central significance for sink functional organization were not observed, turning them into *hidden factors*. These variables are F_L and F_R, the flows of water in the left and right pipes, respectively. Although hidden, these factors were, nevertheless, discovered by SINBAD cells and labeled as V_8 (=F_L) and V_9 (=F_R). Interestingly, SINBAD cells also discovered an additional, unanticipated, hidden factor, V_{10}. Our analysis of its behavior reveals that V_{10} is a ratio: $V_{10} = F_L/F_T$. Upon some

consideration, V_{10} is, in fact, an important factor: it determines, together with *HC*, the water temperature, $t°$ (see eq. 13).

Following the discoveries of hidden factors F_L, F_R, and V_{10}, Virtual Scientist procedures learned how F_L and F_R are determined by knob positions ($V_5 = KP_L$, $V_7 = KP_R$) and directions to open the valves ($V_4 = DIR_L$, $V_6 = DIR_R$), and how F_L and F_R, in turn, determine the total water outflow F_T and its temperature $t°$ (see Figure 7). Overall, the web of inferential relations in Figure 7, generated by the Virtual Scientist, does reflect accurately and efficiently the functional organization of kitchen sinks (actually, to be precise, the organization of our mathematical model of sinks). The web is an inferential model of sinks and, as such, it can be used to predict, for example, how knobs should be positioned in a given sink in order to produce desired water flow and temperature, or to deduce the knob positions from the known faucet output, or to perform any other prediction that can be performed using equations 10-13.

In another demonstration of the close correspondence between the true mathematical description of the studied dynamical system (i.e., equations 10-13) and the Virtual Scientist model of it, the *hubs* in Figure 7 representation of the direct relations actually correspond to the equations that govern the system. Specifically, the hub linking V_4, V_5, V_8 represents eq. 10, the hub linking V_6, V_7, V_9 represents eq. 11, the hub linking V_1, V_8, V_9 represents eq. 12, while two hubs linking V_2, V_3, V_8, V_9 represent eq. 13.

Thus, in conclusion, we find that Virtual Scientist procedures were fully successful in reconstructing from the observed data the prominently nonlinear mathematical model that generated those data.

Discussion

The set of Virtual Scientist procedures described in this paper automates the method of inductive inference to produce a theory of a studied subject that can explain and relate the observed phenomena. The theory emerges in a form of a quantitative model of the subject capable of performing elaborate deductive inferences.

Developing an understanding of a new research subject commonly involves creation of new, subject-specific *concepts* that reveal deeper relations, interdependencies among the subject's observed phenomena. One central function of concepts is to express hidden causal factors, which – once recognized – serve to reduce complexity of the observed relations. As an example, in the kitchen sink study described above, flows of water in the two pipes, F_L and F_R, were two centrally important, but hidden factors. Despite lack of knowledge of F_L and F_R, relations among the observed variables could still be defined, because F_L and F_R were implicit in the states of the observed variables. However, such relations would be more complex. For example, compare the following relation, which uses only the observed variables to express $t°$, with eq. 13, which uses hidden factors F_L and F_R:

$$t^\circ = \frac{(1-HC)(1-DIR_L-KP_L+2\cdot DIR_L\cdot KP_L)+HC\cdot(1-DIR_R-KP_R+2\cdot DIR_R\cdot KP_R)}{2-DIR_L-KP_L+2\cdot DIR_L\cdot KP_L-DIR_R-KP_R+2\cdot DIR_R\cdot KP_R} \quad . \qquad (14)$$

The more complex the inferential relations, the more difficult it will be to learn them; at some degree of complexity, learning will become impossible (Clark and Thornton 1997; Favorov and Ryder 2004). Thus, when dealing with complex subjects, generation of hidden factor concepts is absolutely necessary, if one were to learn interdependencies that otherwise would be too complex to be picked up directly from observations.

This insight is not a new one. Many different methods have been developed aimed at discovery of hidden factors, such as linear methods of correlation analysis, principal component analysis (PCA),

independent component analysis (ICA), or nonlinear methods such as IMAX, nonlinear PCA, and nonlinear ICA (see, for example, Hyvarinen et al. 2001). Linear methods are well advanced, but they are only of limited use, as most of the real-world problems tend to be nonlinear. Among nonlinear methods, our IMAX-based SINBAD method is fundamentally different from most other approaches, which are based on the idea of *data compression*. Data compression based approaches transform raw input information into compact representations that utilize high-level data descriptors; however, the compactness of a representation does not guarantee that the descriptors will turn out to correspond to real, *inferentially useful* hidden factors. Rather, the search for descriptors by compression of the observed data is most likely to produce artificial descriptors that do not reflect the true causal factors operating in the system. In contrast, SINBAD belongs to the class of unsupervised learning algorithms that are based on the IMAX principle of identifying the sources of mutual information among disjoint sources of information, which yields functionally important features in the data reflecting the underlying causal structure of the system. The current version of the SINBAD method makes use of error-backpropagation neural network architecture to implement its learning modules. However, there are no restrictions on the type of learning modules that can be used (neural network or otherwise).

SINBAD discovery of influential hidden factors is one of the cornerstones of the Virtual Scientist approach. Another cornerstone is its focus on learning those relations among the observed and derived variables that predict the state of *one* variable from the states of other variables. The aim is to learn as many ways to infer each variable from the others as can be found. By pursuing this simpleminded strategy of expressing each variable in many different ways in terms of other variables, which in turn are expressed in terms of yet other variables, etc., Virtual Scientist automatically acquires a rich web of inferential relations. In the ideal case, this web (which is grounded in the orderly structure of the studied dynamical system) would link each variable either directly or via intermediaries to every other variable. In this web, all the discovered inferential relations are tied together into a single functional entity – *an inferential model* – revealing the causal organization of the studied subject (Ryder 2004).

A theory constructed by Virtual Scientist can be developed further. Knowledge of the presence of a particular hidden factor, its connections to the observed variables, knowledge of its behavior under various conditions can all be used as clues for discovering its physical identity (as Mendel's inference of the existence of "hereditary factors," for example, led eventually to discovery of genes). Once the physical identities of factors are established, the ways might be devised to measure them directly (thus changing them into observed variables). Direct monitoring of such factors will result in an improved, more efficient and informative data collection, which might lead to discoveries of more deeply hidden factors and development of more insightful updates of the theory. On another tack, the background knowledge of the studied subject, the nature of sensors and their particular locations in the studied system (i.e., information about the subject that is not carried by the observed variables) can also be used to link the Virtual Scientist theory with the larger body of knowledge, fitting it in the context of the overall science.

In conclusion, we believe that the set of Virtual Scientist procedures offers a powerful analytical tool for use in research of complex scientific subjects rich in multivariate and nonlinear relations. The current version of the Virtual Scientist approach described here does not take temporal information into account, as it was intended for problems in which temporal information is not available. The Virtual Scientist strategy, however, is flexible and can readily be adopted for exploration of temporal order as well.

Acknowledgements

The authors wish to acknowledge helpful discussions with Douglas Kelly, Dan Ryder, Sameer Joshi and Mark Tommerdahl. This work was supported, in part, by US Army Research Office grant P43077-LS.

References

Becker S (1999) Implicit learning in 3d object recognition: the importance of temporal context. *Neural Computation* 11: 347-374

Becker S (1996) Mutual information maximization: models of cortical self-organization. *Network* 7(1): 7-31

Becker S (1995) JPMAX: learning to recognize moving objects as a model-fitting problem. In: *Advances in neural information processing systems* 7, Morgan Kaufmann Publishers, San Mateo, CA, pp 933-940

Becker S, Hinton GE (1992) A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature* 355: 161-163

Bishop CM (1995) Neural Networks for Pattern Recognition. Oxford University Press, Oxford

Clark A, Thornton C (1997) Trading places: computation, representation, and the limits of uninformed learning. *Behavioral and Brain Sciences* 20: 57-90

Favorov OV, Ryder D (2004) SINBAD: A neocortical mechanism for discovering environmental variables and regularities hidden in sensory input. *Biological Cybernetics* (accepted)

Favorov OV, Ryder D, Hester JT, Kelly DG, Tommerdahl M (2003) The cortical pyramidal cell as a set of interacting error backpropagating networks: a mechanism for discovering nature's order. In: Hecht-Nielsen R and McKenna T (eds) *Computational Models for Neuroscience*, Springer Verlag, London, pp.25-64

Hanson SJ, Pratt LY (1989) Comparing biases for minimal network construction with backpropagation. In: Touretzky DS (ed), *Advances in Neural Information Processing Systems*, 1: 177-185. San Mateo, CA: Morgan Kaufmann

Hassibi B, Stork DG (1993) Second order derivatives for network pruning: optimal brain surgeon. In: Hanson SJ, Cowan JD, and Giles CL (eds) *Advances in Neural Information Processing Systems* 5:164-171. San Mateo, CA: Morgan Kaufmann

Hyvarinen A, Karhunen J, Oja E (2001) Independent Component Analysis. John Wiley&Sons, Toronto

Joshi S, Kursun O, Favorov OV (2003) Exploiting the structure of order: an application to natural images. The 7th World Multiconference on Systemics, Cybernetics and Informatics, Orlando, FL

Kramer MA (1991) Nonlinear principal component analysis using autoassociative neural networks. *Aiche Journal* 37(2): 233-243

Kursun O, Favorov OV (2003) Single-frame super-resolution by inference from learned features. *Istanbul University Journal of Electrical & Electronics Engineering* 3(1): 673-681

Kursun O, Favorov OV (2002) Single-frame super-resolution by a cortex based mechanism using high level visual features in natural images. IEEE Workshop on Applications of Computer Vision, Orlando, FL

Lang KJ, Hinton GE (1989) Dimensionality reduction and prior knowledge in e-set recognition. NIPS 1989: 178-185

Lappalainen H, Honkela A (2000) Bayesian nonlinear independent component analysis by multi-layer perceptrons. In: Girolami M (ed) *Advances in Independent Component Analysis*, pp 93-121. Springer-Verlag. MATLAB toolbox is available at http://www.cis.hut.fi/projects/bayes/software/

LeCun Y, Denker JS, Solla SA (1990) Optimal brain damage. In: Touretzky DS (ed) *Advances in Neural Information Processing Systems* 2: 598-605. Morgan Kaufmann, San Mateo, CA

Mjolsness E, DeCoste D (2001) Machine Learning for Science: State of the Art and Future Prospects. *Science* 293: 2051-2055

Phillips WA, Singer W (1997) In search of common foundations for cortical computation. *Behavioral and Brain Sciences* 20: 657-722

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: Rumelhart DE, Mcclelland JL, PDP Research Group (eds) *Parallel Distributed Processing: Explorations in the Microstructure Of Cognition*, MIT Press, Cambridge, Mass, 1: 318-362

Ryder D (2004) SINBAD neurosemantics: a theory of mental representation. *Mind & Language* (in press)

Ryder D, Favorov OV (2001) The new associationism: a neural explanation for the predictive powers of cerebral cortex. *Brain and Mind* 2: 161-194

Stone J (1996) Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. *Neural Computation* 8: 1463-1492

Valpola H, Raiko T, Karhunen J (2001) Building blocks for hierarchical latent variable models. *Proc. Int. Conf. on Independent Component Analysis and Signal Separation*, San Diego