Elektrotehniški vestnik 62(3-4): 191–198, 1995
Electrotechnical Review, Ljubljana, Slovenija

Discovering dynamics from measured data

Viljem Križman[1], Sašo Džeroski[1], Boris Kompare[2]

[1]*Laboratorij za umetno inteligenco, Institut Jožef Stefan, Jamova 39, 61111 Ljubljana, Slovenia*
E-mail: viljem.krizman@ijs.si, saso.dzeroski@ijs.si
[2]*Oddelek za hidrotehniko, Fakulteta za arhitekturo, gradbeništvo in geodezijo*
Hajdrihova 28, 61101 Ljubljana, Slovenia
E-mail: bkompare@fagg.uni-lj.si

Abstract. LAGRANGE is a machine discovery system intended to discover empirical laws that govern the behavior of dynamical systems. The main drawback of LAGRANGE is its sensitivity to noisy data, which is mostly due to the use of numerical derivation. This paper describes GOLDHORN, a system that upgrades LAGRANGE with the ability to handle noise. The main technique used in GOLDHORN is numerical integration instead of derivation. Preprocessing of the input data with digital filters is also possible, as well as the discovery of difference equations. We used GOLDHORN to discover differential/difference equation models from measured data in several domains, including fluid dynamics and algal growth.

Key words: automated modelling, dynamical system identification, machine discovery, machine learning

Odkrivanje dinamike iz meritev

Povzetek. LAGRANGE je sistem za avtomatsko odkrivanje zakonitosti, predvsem zakonitosti obnašanja dinamičnih sistemov. Poglavitna pomankljivost tega sistema je občutljivost na šum in napake v meritvah, ki izvira iz uporabe numeričnega odvajanja. V prispevku je predstavljen sistem GOLHORN, ki nadgrajuje sistem LAGRANGE z zmožnostjo obravnave šumnih podatkov. Namesto numeričnega odvajanja uporablja GOLDHORN numerično integriranje. Omogoča tudi predprocesiranje vhodnih podatkov z digitalnimi filtri ter odkrivanje diferenčnih enačb. Sistem GOLDHORN je bil uspešno uporabljen za odkrivanje zakonitosti obnašanja dinamike tekočin in rasti alg.

Ključne besede: avtomatizirano modeliranje, identifikacija dinamičnih sistemov, avtomatsko odkrivanje zakonitosti, strojno učenje

1 Introduction

Natural and artificial dynamical systems, i.e., systems that change their state over time, are abundant in our environment. System identification deals with the problem of building mathematical models of dynamical systems based on observed data from the system [Ljung 1987]. Usually, dynamical systems are described by a set of observable system variables. The most common mathematical models of dynamical systems are ordinary differential equations, which completely specify the rate of change of each of the system variables.

The task of identification of dynamical systems as addressed in this paper can be sumarized as follows: Given an example behavior of a dynamical system, find a set of laws that describe the dynamics of the system. More precisely, a set of real-valued system variables is measured at regular intervals over a period of time, as illustrated in Table 1. The laws to be discovered (also called model of the dynamical system) typically take the form of a set of differential equations of the form $dX_i/dt = f_i(X_1, \ldots, X_n), i = 1, \ldots, n$.

In mainstream system identification, as summarized by Ljung [1993], the assumption is that the model structure (i.e. the functional form of the functions f_i) is known. The task is then to determine suitable values for the parameters appearing in the model. This task is accordingly called *parameter identification*. Linear model structures are most often used, where the optimal parameter values can be determined by using the well known least squares method. This method can also be used for nonlinear models that are linear in the parameters.

The most important step in the identification process as a whole is the decision upon the model structure. In practice, typically a whole lot of them are tried out and the process of identification really becomes the process of evaluating and choosing between the resulting models in these different structures [Ljung 1993]. We refer to this task as the task of *structure identification*. Obviously, it includes the task of parameter identification.

One possible approach to search for the best structure is to perform parameter estimation for all possible

Received 15 June 1995
Accepted 23 June 1995

Time	System variables			
	X_1	X_2	...	X_n
t_0	x_{10}	x_{20}	...	x_{n0}
$t_1 = t_0 + h$	x_{11}	x_{21}	...	x_{n1}
⋮	⋮	⋮	⋱	⋮
$t_N = t_0 + Nh$	x_{1N}	x_{2N}	...	x_{nN}

Table 1. A behavior trace of a dynamical system.

structures and to choose one on the basis of comparing certain performance indices, such as the multiple correlation coeficient. Existing approaches of this kind are mostly input/output and assume that the possible models are difference equations linear in the parameters. A survey of such approaches is given by Haber and Unbehauen [1990].

The paper deals with a state-space approach to dynamical system identification. This approach searches through an implicitly specified space of model structures, which can be nonlinear in the parameters. The models have the form of differential equations, although algebraic equations may also be found. The approach explores all possible model structures in the specified space.

The approach is an extension of techniques developed in the area of machine learning, a subfield of artificial intelligence. Induction is the process whereby general laws are derived from specific observations. Machine learning [Bratko 1989] is concerned with automating this process. The approach from machine learning considered in this paper is machine discovery of empirical laws [Langley, *et al.* 1987].

While several sytems exist that discover empirical numerical laws from data, e.g., BACON [Langley, *et al.* 1987], ABACUS [Falkenheiner & Michalski, 1990] and FAHRENHEIT [Żytkow & Zhu, 1991], few have so far addressed the problem of discovering laws that govern dynamical systems. LAGRANGE [Džeroski & Todorovski 1993, 1995, Todorovski 1993] and GPDD [Džeroski & Petrovski, 1994] are unique in this respect. They take as input a behavior of a dynamical system (as specified in Table 1) and produce a set of laws that describe the dynamics of the underlying system.

In addition to the input behavior, LAGRANGE has to be provided with the values of several parameters: the order o of the dynamical system (the order of the highest derivative appearing in the dynamics equations), the maximum depth d of new terms introduced by combining old terms (variables), and the maximum number r of independent regression variables used for generating equations. The LAGRANGE algorithm consists of three main stages. Taking the set of system variables, LAGRANGE first introduces their time derivatives (up to order o). It then introduces new variables (terms) by repeatedly applying multiplication to variables from S and their time derivatives. Finally, given the set V of all variables (terms), LAGRANGE generates and tests equations by using linear regression.

Roughly speaking, each subset of V sized at most $r+1$ is used to generate a linear equation. The term with greatest depth (complexity) is chosen as the dependent variable and is expressed as a linear combination of the remaining ones. The constant coefficients in the linear equation are calculated by applying linear regression. If the equation appears to be significant, it is added to the model. The significance of an equation is judged by the multiple correlation coefficient R and the normalized deviation S, calculated as $R^2 = 1 - E/\sum_{j=0}^{N}(y_j - \overline{y})^2$ and $S^2 = E/[(N+1)(y^2 + e^{-\overline{y}^2})]$, where E is the sum of squared errors of the dependent variable y, y_j is the value of y at time t_0+jh, and $\overline{y}$ is the average value of y. Smaller values of S and larger values of R correspond to more significant equations.

LAGRANGE has been applied to reconstruct the models of several dynamical systems [Džeroski & Todorovski 1993, 1995], the most complex being the inverted pendulum, a standard benchmark problem for dynamic system control. However, the experiments have been performed on simulated data. Applications to modelling real dynamical systems have been hindered by the sensitivity of LAGRANGE to noise and other minor problems.

This paper presents an extension of LAGRANGE that effectively deals with noisy data and is applicable to modelling real dynamical systems from measured data. We first describe in detail some of the problems encountered in LAGRANGE, in particular the sensitivity to noise and the choice of dependent variables. We then describe GOLDHORN and the techniques it uses, including numerical integration and digital filtering. We proceed with a description of experiments on noisy data with synthetically introduced noise. The application of GOLDHORN to modelling three real dynamical systems from measured data is described next. Two systems from the area of fluid dynamics and algal growth in the Lagoon of Venice are successfully modelled. We conclude with a brief discussion of related work and the contributions of this paper.

2 Problems with LAGRANGE

We have identified two problems with LAGRANGE that are of statistical nature. These are the choice of the dependent variable for linear regression and the sensi-

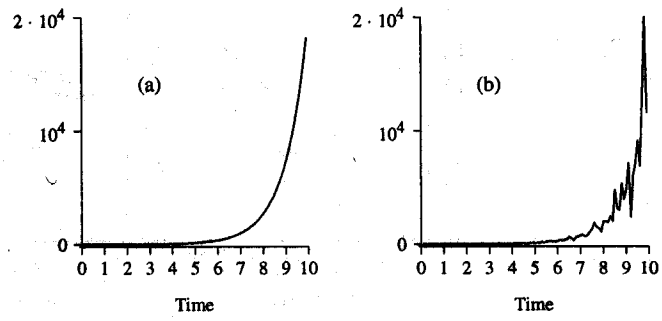

Figure 1. The effect of noise on numerical derivation.

tivity to noise in the data. The latter is especially important, as LAGRANGE performs numerical derivation. The problem of choosing the dependent variable is more serious in the presence of noise.

To understand the effects of noise in LAGRANGE, consider the error made when calculating the derivatives by using aproximation polynomials. The error due to measurement errors in the derivated variable is proportional to $1/h$, and the error of the derivation method is proportional to h^{i+1}, where i is the order of the approximation polynomial used. Therefore, high errors in numerically calculated derivatives are unavoidable.

Before proceeding further, let us describe the process of introducing noise to a given variable X (used in our experiments). Let the noise level be p_N, corresponding to at most $100p_N\%$ of relative error caused by the noise process. For each value x_i, we generate a new value $\hat{x}_i$ by adding a random amount of error, according to the equation $\hat{x}_i = x_i + x_i p_N Z$. Z is a random variable, obtained from a random variable with a normal distribution (mean 0 and standard deviation 1) by eliminating outcomes with absolute value greater than 1.

Let us now illustrate the above problem by example. Consider the function $y(t) = e^t$ and its derivative $y' = y$. Figure 1 depicts the function $y(t)$ and its derivative, calculated numerically (with a fourth degree approximation polynomial). Noise was artificially introduced in the function y, so that the error for y at each time point is at most 1 %. While the errors are not visible on the graph (a) depicting y, they are very large for the numerically calculated derivative, depicted on graph (b).

Given a set of terms that should appear in a linear equation, LAGRANGE choses the term of highest degree (depth) as the dependent variable for linear regression. Ties are resolved by taking the first term according to lexicographic order. However, the quality criteria for the equation, including the correlation coefficient, the normalized deviation, and the relative error of the equation coefficients, strongly depend on the choice of dependent variable.

Consider a simple dynamical system representing a chemical reaction with linear kinetics. The behavior of this system is depicted in Figure 2. The law of mass conservation holds for the three substances in the reaction, i.e., $a(t) + b(t) + c(t) = 100$. This equation is discovered by LAGRANGE, in addition to the equations describing dynamics ($\dot{a} = -2a$, $\dot{b} = 2a - 3b$, and $\dot{c} = 3b$; $\dot{X} = dX/dt$ denotes the time derivative of X). Consider now the effects of introducing different amounts of noise in the values of a, b, and c on the correlation coefficient, normalized deviation and the relative error of the constant 100 when each of a, b and c is chosen as the dependent variable for the above equation.

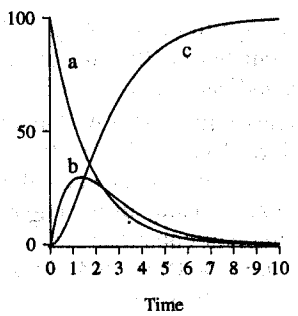

Figure 2. An example from chemical kinetics.

We added noise at six different levels (1, 4, 8, 12, 16, and 20%) to the data obtained by numerical integration of the above three differential equations. For each noise level, we performed linear regression with each of the variables a, b, and c as a dependent variable and the remaining two as independent variables. We recorded the correlation coefficient R, the normalized deviation S, and the relative error of the constant term (with correct value of 100) in each equation. Each experiment was repeated 10 times with different randomizations in the noise addition process. The values depicted in Figure 3 are averages over the 10 experiments.

From Figure 3, we can conclude that choosing c as the dependent variable gives the best results, while choosing b gives the worst results. The choice of the dependent variable can greatly influence the quality of the derived equation. While the quality of the generated equations uniformly decreases as the noise level increases, it does much less so when c is chosen as a dependent variable. Note that the averages/variances for a, b and c are 5.05/229.21, 3.32/55.57 and 91.62/450.71. Seemingly, greater average/variance of the dependent variable implies higher R and lower S. Given same values of the sum of squared errors E, greater average/variance of the dependent variable does imply higher R and lower S, as E is divided by the variance and the

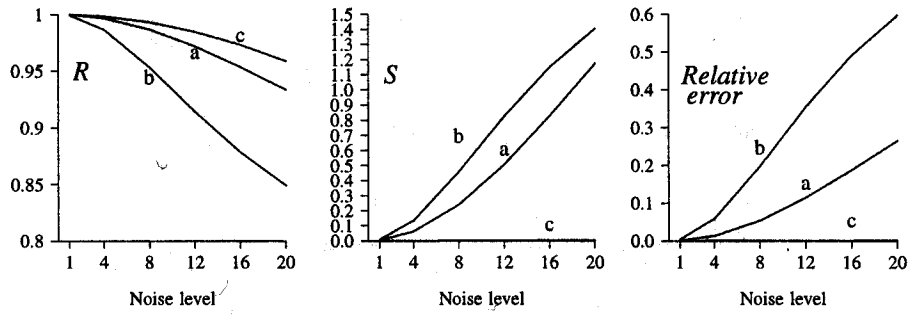

Figure 3. The effect of chosing the dependent variable for linear regression.

square of the average to obtain R^2 and S^2. The relative errors of the calculated coefficients are also smaller for higher averages and variances of the dependent variable.

Other problems with LAGRANGE have been identified [Džeroski & Todorovski, 1995]. These include the limited ability to use mathematical functions (only sin and cos), the redundancy of the equations generated, and generating implicit equations (where the highest order derivatives are not expressed explicitly). In fact, there is no guarantee that LAGRANGE will find a set of differential equations that can be used to simulate the dynamical system that we intend to model. Finally, LAGRANGE can use no background knowledge: even expressions on existing variables have to be calculated externally and then provided to LAGRANGE.

3 GOLDHORN

This section describes the system GOLDHORN that upgrades LAGRANGE in several directions. First, it allows for introducing new variables through the use of the functions e^x, $\log(x)$, $|x|$, and $\text{sign}(x)$, in addition to $\sin(x)$ and $\cos(x)$ that are used in LAGRANGE. Second, it expresses the highest order derivatives explicitly as rational functions of the system variables and their lower order derivatives. By doing this, GOLDHORN avoids the need to use the highest order numerical derivatives. To estimate equation coefficients, as well as the quality of the explicit equations, GOLDHORN uses numerical integration, rather than derivation. In addition, GOLDHORN allows the measured data to be pre-processed with digital filters that alleviate the effects of noise to a certain degree. Finally, GOLDHORN can be used to discover difference equations, thus eliminating the problems with numerical derivation.

LAGRANGE looks for a set of equations of the following form, where $X_1, \ldots, X_n$ are the system variables, o the order of the dynamic system (i.e., of the highest order derivatives in the dynamics equations) and F is a polynomial of degree d or less.

$$F(X_1, \ldots, X_n, \dot{X}_1, \ldots, \dot{X}_n, \ldots, X_1^{(o)}, \ldots, X_n^{(o)}) = 0 \qquad (1)$$

In GOLDHORN, we can restrict the search to equations that are linear in the highest order derivatives $X_i^{(o)}$ and can be rewritten as

$$X_i^{(o)} = F_i(X_1, \ldots, X_n, \dot{X}_1, \ldots, \dot{X}_n, \ldots, X_1^{(o-1)}, \ldots, X_n^{(o-1)}), \qquad (2)$$

where F_i is a rational function. In this way, we obtain equations that are suitable for simulation of the dynamical system. As a bonus, the number of equations considered is greatly reduced. Namely, when introducing new variables with multiplication, only variables that contain at most one highest order derivative are introduced. Furthermore, only equations where at least one term contains a highest order derivative are considered.

GOLDHORN first introduces all derivatives by numerical derivation. It then considers all implict equations, i.e., equations of form (1), that can be rewritten in explict form, i.e., as equations of form (2). The term with largest variance is chosen as the dependent variable and the initial coefficients are determined by linear regression. GOLDHORN then expresses $X_i^{(o)}$ explicitly, i.e., $X_i^{(o)} = F_i$, and uses nonlinear optimization and numerical integration to fit the coefficents. More precisely, instead of fitting the coefficients in F_i to minimize

$$E_1 = \sum_{j=0}^{N} (X_i^{(o)}(t_0 + jh) - F_i(t_0 + jh))^2,$$

it fits them to minimize

$$E_2 = \sum_{j=0}^{N} (X_i^{(o-1)}(t_0 + jh) - \int_{t=0}^{t_0+jh} F_i(t_0 + jh)dt)^2.$$

The downhill simplex method of nonlinear optimization [Press, *et al.* 1986] is used. The quality of an equation is then judged by the quantity A, defined as

$$A = E_2 / \sum_{j=0}^{N} (X_i^{(o)}(t_0 + jh) - \overline{X_i^{(o)}})^2.$$

The lower A, the more significant the equation.

Digital filtering can be applied to measured signals (dynamical system behaviors) to selectively remove noise, e.g., a low-pass filter can be applied to remove high frequency noise. GOLDHORN includes several linear filters with finite impulse response, which have the form $y_n = \sum_{i=0}^{M} c_i x_{n-i}$. Only even filter lengths M

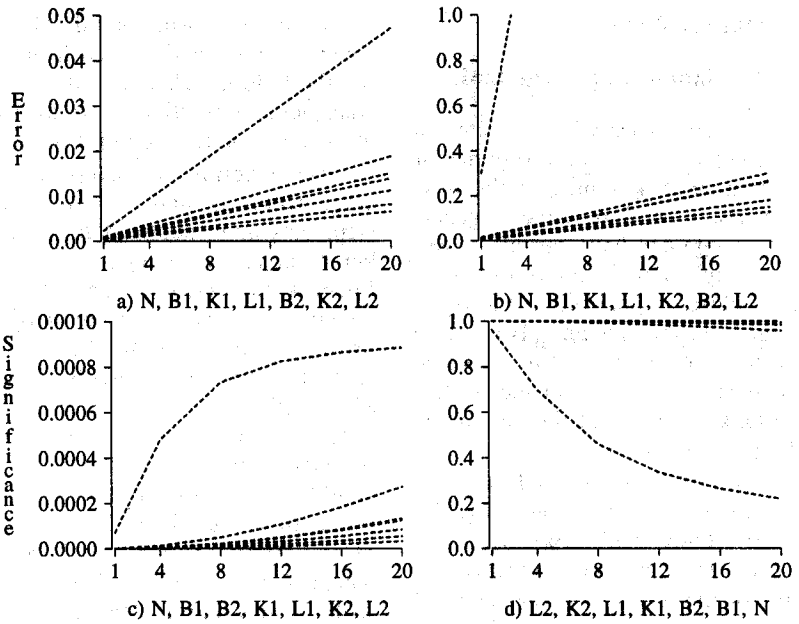

Figure 4. The effect of noise on equation discovery in GOLDHORN and LAGRANGE.

are used (for symmetry reasons). Setting $L = \frac{1}{2}M$ and $S = L + 1$, the three types of low-pass filters used in GOLDHORN, named respectively linear (L), quadratic (K), and binomial (B), are defined by the following equations. All coefficients for a single filter sum up to one.

$$y_{n+L} = \sum_{i=-L}^{L} \frac{S - |i|}{S^2} x_{n+i},$$

$$y_{n+L} = \sum_{i=-L}^{L} \frac{3 * (S - |i|)^2}{S(2S^2 + 1)} x_{n+i},$$

$$y_{n+L} = \sum_{i=-L}^{L} \frac{\binom{M}{L+i}}{2^M} x_{n+i}. \qquad (3)$$

Note that due care should be exercised when applying filters to the measured signals: applying an inappropriate filter may not only remove noise but also distort the original signal.

Finally, let us note that difference equations, instead of differential equations can be produced by GOLDHORN, thus avoiding the need for numerical derivation altogether. Instead of introducing $X'(t) = dX(t)/d(t)$, we introduce $X'(t) = X(t + h)$. In this case, numerical integration and derivation are not used. In addition to the correlation coefficient R and the normalized deviation S, the sum of squared error E can be used to estimate the significance of equations: the lower E, the more significant the equation.

4 Experiments with synthetic data

To investigate the performance of GOLDHORN on noisy data, we performed several experiments where noise was synthetically introduced to data generated by numerical integration of a target set of differential equations. Noise was introduced at several different levels, as described in Section 2. The noisy data were filtered by each of the three kinds of digital filters in GOLDHORN, each with length 12, 24, 36, 48, and 60. Both GOLDHORN and LAGRANGE were run on the nonfiltered and filtered data.

Experiments were performed in several synthetic domains [Križman, 1994], most of them taken from [Todorovski, 1993]. For illustration, we present the results in the domain of chemical kinetics, described in Section 2. Figure 4 depicts the performance of GOLDHORN (graphs a) and c)) and LAGRANGE (graphs b) and d)) at discovering the equation $\dot{a} = -2a$.

The x-axes represent the noise level, the y-axes represent the relative error for the coefficient -2 (graphs a) and b)) and the significance of the equation (A in graph c) and R in graph d)). Note that lower values of A denote higher significance, while lower values of R denote lower significance! The different lines in each graph represent results obtained by six different filters: L1, K1, and B1 of length 24, L2, K2, and B2 of length 48. The letters under each graph tell us what filter each line corresponds to, starting from the top (N means no filter was used). For example, the top line in graph a) depicts the relative error of the coefficent on nonfiltered data, the next line on data filtered with the B1 filter, and the bottom line on data filtered with the L2 filter.

It can be immediately noticed that the relative error of the coefficient is larger in LAGRANGE (graph b)) by an order of magnitude. The error skyrockets on nonfiltered data. The relative error decreases gradually with the increase in the noise level, and the significance decreases slightly, except on nonfiltered data.

5 Modelling real dynamical systems

Modelling water level oscillations in a surge tank

A surge tank is a container that is usually connected to the pressure pipes that conduct water to the turbines of a hydro-power plant. If a sudden change of the flow through the turbines occurs, a pressure surge is generated in the pipeline. The surge can have serious consequences on the pipeline (explosion or implosion), so a surge tank is situated as close as possible to the place of formation of the surge, i.e., the turbine valve. Surge pressure is transformed to water movement in the surge tank, resulting in an increase or decrease of the steady-state water level in the tank. The rest of the pipeline is thus not exposed to the pressure shocks.

In this experiment, GOLDHORN was used to model the water level oscillations in a laboratory model (replica) of a surge tank. Only one variable is measured, i.e., the water level H. The measurements and the numerical derivative of the water level are depicted in Figure 5. The system is a second order one. In addition to the water level rate of change $\dot{H}$, its magnitude $|\dot{H}|$ was provided.

The maximum depth of terms was set to two and the number of independent regression variables to three. Several filters were applied to the data, namely linear, quadratic, and binomial filters of length 2, 4, 6, 8, 10, 12, and 14. Best results were obtained with linear filters of length 6 to 10. The following equation was the best according to the A measure (data were filtered with a linear filter of length 6): $\ddot{H} = -0.057H - 3.629\dot{H}|\dot{H}|$. When simulated, this equation reproduces the observed behavior almost exactly (no visible difference). On the nonfiltered data, GOLDHORN produced the equation $\ddot{H} = -0.059H - 2.737\dot{H}|\dot{H}|$. For comparison, LAGRANGE produced the equation $\ddot{H} = 0.239 - 0.184H - 23.9777\dot{H}|\dot{H}|$ from the nonfiltered data, and $\ddot{H} = -0.058H - 3.829\dot{H}|\dot{H}|$ from data filtered with a linear filter of length twelve.

The surge tank model is a special case of the U-tube model, described below. In the surge tank case, friction losses can be regarded as constant.

Modelling water level oscillations in the U-tube

The U-tube is a pipe in the shape of the letter U, filled with fluid. The fluid levels at each side are equal in the steady state of the system. If the fluid column is disturbed, it begins to oscillate around its steady state. These oscillations are dumpened due to the friction (viscosity) of the fluid. The oscillating frequency is mostly affected by the length L of the fluid column. The general equation describing the oscillations, derived from first principles, has the form $\ddot{H} = c_1 H + c_2 \lambda \dot{H}|\dot{H}|$, where λ represents friction losses and is a complicated function of $\dot{H}$. The term $c_2\lambda\dot{H}|\dot{H}|$ can be of the order between $\dot{H}$ and $\dot{H}^2$, depending on the magnitude of $\dot{H}$.

We constructed a simple U-tube model from a pipe with a retively large diameter to avoid influence of the surface tension of the liquid. The U-tube was filled with water, which was disturbed from stationary state by tilting the U-tube. This caused the water to oscillate. The water level oscillations were measured by an ultra-sonic probe and are depicted on Figure 6.

The experimental setup was the same as for the surge tank: $\dot{H}$ and $|\dot{H}|$ were introduced in the first step, maximum depth of terms was two and the number of independent regression variables three. Several filters were applied to the data. Best results were obtained with linear filters of length 14 to 22. The following equation was the best according to the A measure (data were filtered with a linear filter of length 22): $\ddot{H} = -25.61H - 0.38\dot{H}$. When simulated, this equation reproduces the observed behavior almost exactly (no visible difference). On the nonfiltered data, GOLDHORN produced the equation $\ddot{H} = -21.74H - 1.96\dot{H}$. For comparison, LAGRANGE produced the equation $\ddot{H} = -0.08 - 23.41H - 12.28\dot{H}$ from the nonfiltered data, and $\ddot{H} = 0.01 - 27.91H - 0.42\dot{H}$ from data filtered with a linear filter of length fourteen.

As the diameter of the pipe is just large enough, the water flow is mostly in the laminary regime and not affected with surface tension, which involves linear dependence on $\dot{H}$. Thus, the discovered equation is consistent with the general equation given above.

Modelling algal growth in the Lagoon of Venice

The Lagoon of Venice measures 550 km^2, but is very shallow, with an average depth of less than 1m. It is heavily influenced by anthropogenic inflow of nutrients - 7 mio kg/year of nitrogen and 1.4 mio kg/year of phosphorus [Bendoricchio, *et al.* 1994]. These loads (mainly nitrogen) are above the Lagoon's admissible trophic limit and generate its dystrophic behavior, which is characterized by excessive growth of algae, mainly *Ulva rigida.*

Four sets of measured data were available [Coffaro, *et al.* 1993]. The data were sampled weekly for slightly more than one year at four different locations in the Lagoon. Location 0 was sampled in 1985/86, locations 1, 2, and 3 in 1990/91. The sampled quantities are nitrogen in ammonia NH_3, nitrogen in nitrate NO_3, phosphorus in orthophosphate PO_4 (all in μg/l), dissolved oxygen DO (in % of saturation), temperature T (degrees C), and algal biomass B (dry weight in g/m^2). In addition to the measured variables, GOLDHORN was provided with the growth μ and mortality ω rates, which are relatively well defined relations and can be calculated from the measured variables.

We applied GOLDHORN to model algal growth at Station 0. Difference equations were sought that express $B(t+1)$, i.e., the algal biomass at week $t+1$, in terms of the measured variables and the growth/mortality rates at week t, i.e., $NH_3(t)$, $NO_3(t)$, $PO_4(t)$, $DO(t)$, $T(t)$, $B(t)$, $\mu(t)$, and $\omega(t)$. The depth of variables was set to two and the number of independent regression variables to eight. The best equation, according to the sum of squared errors E, was $B(t+1) = -\frac{0.611}{\omega(t)} - 2077\omega(t) + 0.653DO(t) + 0.662B(t) + 7.490T(t)$.

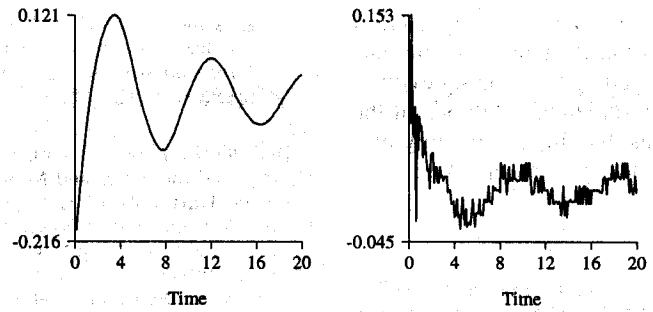

Figure 5. Water level (left) oscillations and the corresponding flow (right) in a surge tank.

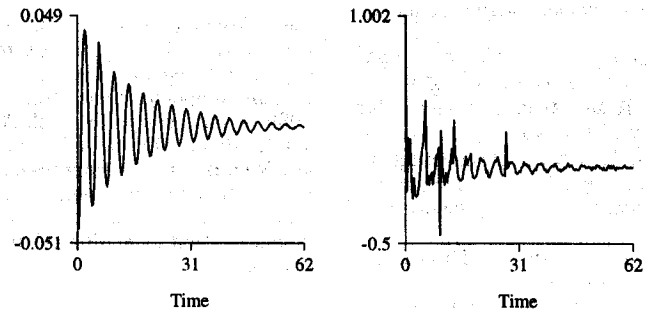

Figure 6. Water level (left) oscillations and the corresponding flow (right) in the U-tube.

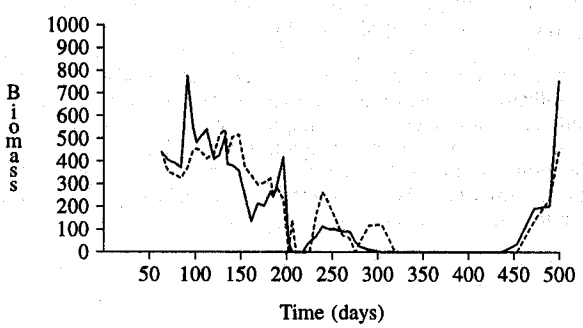

Figure 7. Algal biomass in the Venice lagoon as measured (solid line) and predicted by the equation discovered by GOLDHORN (dashed line).

Figure 7 depicts the measured biomass and the biomass predicted by the above equation. While the fit is not perfect, one should take into account that measurement errors for the biomass are of the order 20-50 %. The equation predicts correctly most of the peaks and crashes, both in time and to a certain degree in magnitude. These quantites are more important to ecologists than a very close degree of fit.

6 Discussion

We have presented the GOLDHORN system for machine discovery of empirical laws that govern dynamical systems. As compared to its predecessor LAGRANGE, GOLDHORN has the important ability to handle noisy data. The particular techniques used are digital filtering, numerical integration, and the discovery of difference equations. While the discovery of difference equations has already been proposed [Todorovski & Džeroski, 1994], it has not been applied to measured data.

We applied GOLDHORN to model three real dynamical systems from measured data. Good models, with almost perfect fit in two cases and a good qualitative fit in the third, were obtained. Two of the three systems were laboratory replicas of real systems and were measured under controlled conditions. The third process, namely algal growth in the Lagoon of Venice, takes place in areal dynamical system. It has all the characteristics that make automated modelling difficult: measurement errors for biomass are between 20 and 50 %, important factors that influence the measured biomass are not taken into account, e.g., winds and tidal currents, etc. Nevertheless, the model constructed by GOLDHORN captures important properties of algal growth as it is able to predict correctly most of the algal blooms (peaks) and crashes: it is the blooms and crashes that ecologists are most worried about.

Just as LAGRANGE, GOLDHORN is related to the work on discovery of empirical numerical laws [Langley, *et al.* 1987; Falkenheiner & Michalski, 1990; Żytkow & Zhu, 1991; Moulet, 1994]. Unfortunately, there have been few applications of machine discovery methods to practical domains. A notable exception is E^* [Schaffer, 1990], which was applied to a wide range of modelling problems from the area of physics.

GOLDHORN is also related to GPDD [Džeroski & Petrovski, 1994]. GPDD uses genetic programming search techniques to explore a space of models that can be larger than that of LAGRANGE/GOLDHORN.

GPDD uses numerical integration just as GOLDHORN does. Both the structure and the parameters of the target model are determined: genetic programming evolves the structure and numerical optimization methods fit the parameters. However, it seems that the genetic programming search can easily get stuck in local optima.

7 References

[1] Bendoricchio, G., Coffaro, G., and De Marchi, C. (1994). A trophic model for Ulva Rigida in the Lagoon of Venice. *Ecological Modelling*, 75/76: 485–496.

[2] Bratko, I. (1989). Machine learning. In Gilhooly, K., editor, *Human and Machine Problem Solving*. Academic Press, London.

[3] Coffaro, G., Carrer, G., and Bendoricchio, G. (1993). *Model for Ulva Rigida Growth in the Lagoon of Venice.* Report UNESCO MURST Research Project Venice Lagoon Ecosystem. University of Padova.

[4] Džeroski, S. and Petrovski, I. (1994). Discovering dynamics with genetic programming. In *Proc. Seventh European Conference on Machine Learning*, pages 347–350. Springer, Berlin.

[5] Džeroski, S. and Todorovski, L. (1993). Discovering dynamics. In *Proc. Tenth International Conference on Machine Learning*, pages 97–103. Morgan Kaufmann, San Mateo, CA, 1993.

[6] Džeroski, S. and Todorovski, L. (1995). Discovering dynamics: from inductive logic programming to machine discovery. *Journal of Intelligent Information Systems*, 4: 89–108.

[7] Falkenheiner, B. and Michalski, R. (1990). Integrating quantitative and qualitative discovery in the ABACUS system. In Kodratoff, Y. and Michalski, R., editors, *Machine Learning: An Artificial Intelligence Approach*, pages 153–190. Morgan Kaufmann, San Mateo, CA.

[8] Haber, R., and Unbehauen, H. (1990). Structure identification of nonlinear dynamic systems - a survey of input/output approaches. *Automatica*, 26(4):651–677.

[9] Križman, V. (1994) Handling noisy data in automated modeling of dynamical systems. MSc Thesis, Faculty of Electrical and Computer Engineering, University of Ljubljana, Slovenia.

[10] Langley, P., Simon, H., Bradshaw, G., and Żytkow, J. (1987). *Scientific discovery*. MIT Press, Cambridge, MA.

[11] Ljung, L. (1987). *System Identification: Theory for the User*. Prentice Hall, Englewood Cliffs, NJ.

[12] Ljung, L. (1993). Modelling of industrial systems. In *Proc. Seventh International Symposium on Methodologies for Intelligent Systems*, pages 338–349. Springer, Berlin.

[13] Moulet, M. (1994). Iterative model construction with regression. In *Proc. Eleventh European Conference on Artificial Intelligence*, pages 448–452. John Wiley & Sons, Chichester.

[14] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1986). *Numerical Recipes*. Cambridge University Press, Cambridge, MA.

[15] Schaffer, C. (1990) *Domain-Independent Scientific Function Finding*. PhD Thesis, Department of Computer Science, Rutgers University (Technical Report LCSR-TR-149).

[16] Todorovski, L. (1993). *Modeling Dynamic Systems with Machine Discovery*. (1993). BSc Thesis, Faculty of electrical engineering and computer science, University of Ljubljana, Slovenia.

[17] Todorovski, L. and Džeroski, S. (1994). Modeling dynamic systems with machine discovery. *Electrotechnical Review*, 61(1-2): 55–64. In Slovenian.

[18] Żytkow, J. and Zhu, J. (1991). Application of empirical discovery in knowledge acquisition. In *Proc. Fifth European Working Session on Learning*, pages 101–117. Springer, Berlin.

Viljem Križman was born on 3 April 1966 in Koper, Slovenija. He received the B.Sc. and M.Sc. degrees in computer science from the University of Ljubljana in 1990 and 1994, respectively. He is currently a Research Assistant at the Jožef Stefan Institute, and is working toward Ph.D. degree in computer science. His research interests include machine learning, machine discovery, modeling and control of dynamic systems. Lately he is investigating the use of multidimensional optimisation algorithms in the field of machine discovery.

Sašo Džeroski was born on 31 May 1968 in Ohrid, Makedonija. He holds a B.Sc. degree (1989) and a M.Sc. (1991) degree in Computer Science. In 1995 he received a Ph.D. degree in computer science from University of Ljubljana. Since 1989 he has been a Research Assistant at the Jožef Stefan Institute. Since 1995 he has been a visiting researcher (ERCIM) at GMD, Bonn. His main research interest is in machine learning, in particular in teoretical aspects of Inductive Logic Programming and in applications of machine learning in practical domains, such as ecological modelling and dynamic systems control.

Boris Kompare is an Assistant Professor at the Faculty of Civil Engineering and Geodesy, University of Ljubljana, Slovenia. His background is civil engineering and environmental chemistry. His main fields of interest are sanitary engineering (potable water supply, potable and waste water treatment, urban drainage, groundwater pollution, etc.) and environmental engineering (restoration of lakes, watercourses, and sea, revitalization of reclaimed land etc.). He constructs and uses mathematical models for solving the listed problems. His latest modelling work employs AI tools for automatic system identification and automatic or semi automatic model construction - i.e. he uses models constructed by AI tools as an inspiration how to improve his conceptual models.