Semi-Supervised Regression and System Identification

Henrik Ohlsson, Lennart Ljung

Division of Automatic Control

E-mail: ohlsson@isy.liu.se, ljung@isy.liu.se

25th April 2010

Report no.: LiTH-isy-R-2940

Accepted for publication in Three Decades of Progress in Systems and Control, Springer Verlag, 2010.

Address:

Department of Electrical Engineering

Linköpings universitet

SE-581 83 Linköping, Sweden

WWW: <http://www.control.isy.liu.se>

AUTOMATIC CONTROL
REGLERTEKNIK
LINKÖPINGS UNIVERSITET

Technical reports from the Automatic Control group in Linköping are available from

<http://www.control.isy.liu.se/publications>.

Abstract

System Identification and Machine Learning are developing mostly as independent subjects, although the underlying problem is the same: To be able to associate “outputs” with “inputs”. Particular areas in machine learning of substantial current interest are *manifold learning* and *unsupervised* and *semi-supervised regression*. We outline a general approach to semi-supervised regression, describe its links to Local Linear Embedding, and illustrate its use for various problems. In particular, we discuss how these techniques have a potential interest for the system identification world.

Keywords: Semi-supervised regression, system identification

Semi-Supervised Regression and System Identification

Henrik Ohlsson and Lennart Ljung

Dedicated to Chris and Anders at the peak of their careers

Abstract System Identification and Machine Learning are developing mostly as independent subjects, although the underlying problem is the same: To be able to associate “outputs” with “inputs”. Particular areas in machine learning of substantial current interest are *manifold learning* and *unsupervised* and *semi-supervised regression*. We outline a general approach to semi-supervised regression, describe its links to Local Linear Embedding, and illustrate its use for various problems. In particular, we discuss how these techniques have a potential interest for the system identification world.

1 Introduction

A central problem in many scientific areas is to link certain observations to each other and build *models* for how they relate. In loose terms, the problem could be described as relating y to φ in

$$y = f(\varphi) \quad (1)$$

where φ is a vector of observed variables and y is a characteristic of interest.

In system identification φ could be observed past behavior of a dynamical system, and y the predicted next output. In classification problems φ would be the vector of features and y the class label. Following statistical nomenclature, we shall

Henrik Ohlsson

Division of Automatic Control, Department of Electrical Engineering, Linköpings universitet, SE-583 37 Linköping, Sweden, e-mail: ohlsson@isy.liu.se

Lennart Ljung

Division of Automatic Control, Department of Electrical Engineering, Linköpings universitet, SE-583 37 Linköping, Sweden, e-mail: ljung@isy.liu.se

generally call φ the *regression vector* containing the *regressors* and following classification nomenclature we call y the corresponding *label*.

The information available could be a collection of labeled pairs

$$y(t) = f(\varphi(t)) + e(t), \quad t = 1, \dots, N_l \quad (2)$$

where e accounts for possible errors in the measured labels. Constructing an estimate of the function f from labeled data $\{(y(t), \varphi(t)), t = 1, \dots, N_l\}$ is a standard *regression problem* in statistics, see e.g. [9].

We shall in this contribution generally not seek explicit constructions of the estimate f , but be content by having a scheme that provides an estimate of $f(\varphi^*)$ for any given regressor φ^* . This approach has been termed *Model-on-Demand*, [23] or *Just-In-Time* modeling, [5].

The term *supervised learning* is also used for such algorithms, since the construction of f is “supervised” by the measured information in y . In contrast to this, *unsupervised learning* only has the information of the regressors $\{\varphi(t), t = 1, \dots, N_u\}$. In unsupervised classification, e.g. [11], the classes are constructed by various clustering techniques. *Manifold learning*, e.g. [24, 21] deals with unsupervised techniques to construct a manifold in the regressor space that houses the observed regressors.

Semi-supervised algorithms are less common. In semi-supervised algorithms, both labeled and unlabeled regressors,

$$\{(y(t), \varphi(t)), t = 1, \dots, N_l, \varphi(t), t = N_l + 1, \dots, N_l + N_u\} \quad (3)$$

are used to construct f . This is particularly interesting if extra effort is required to measure the labels. Thus costly labeled regressors are supported by less costly unlabeled regressors to improve the result.

It is clear that unsupervised and semi-supervised algorithms are of interest only if the regressors have a pattern that is unknown *a priori*.

Semi-supervised learning is an active area within classification and machine learning (see [4, 28] and references therein). In classification, it is common to make the assumption that class labels do not change in areas with a high density of regressors. Figure 1 gives an illustration of this situation. To estimate the high density areas, unlabeled data are useful.

The main reason that semi-supervised algorithms are not often seen in regression and system identification may be that it is less clear when unlabeled regressors can be of use. We will try to bring some clarity to this through this chapter. Let us start directly by a pictorial example. Consider the 5 regressors shown in the left of Fig. 2. Four of the regressors are labeled and their labels are written out next to them. One of the regressors is unlabeled. To estimate that label, we could compute the average of the two closest regressors’ labels, which would give an estimate of 2.5. Let us now add the information that the regressors and the labels were sampled from an in time continuous process and that the value of the regressor was evolving along the curve shown in the right part of Fig. 2. Knowing this, a better estimate of the label would probably be 1. The knowledge of that the regressors are restricted to a certain region in the regressor space can hence make us reconsider our estimation strategy.

Fig. 1 Left side shows three regressors, two labeled, with the class label next to them, and one unlabeled regressor. Desiring an estimate of the label of the unlabeled regressor, having no further information, we would probably guess that it belongs to class *B*. Now, let us assume that we are provided the information that regressors are constrained to lay on the black areas (on the elliptic curve on which the labeled regressor of class *A* lay or on the elliptic filled area in which the labeled regressor belonging to *B* belongs) shown in the right part of the figure. What would the guess be now?

Fig. 2 The left side shows 5 regressors, four labeled and one unlabeled. Desiring an estimate of the label of the unlabeled regressor, we could simply weight together the two closest regressors' labels and get 2.5. Say now that the process that generated our regressors, traced out the path shown in the right part of the figure. Would we still guess 2.5?

Notice also that to estimate the region to which the regressors are restricted, both labeled and unlabeled regressors are useful.

Generally, regression problems having regressors constrained to rather limited regions in the regressor space may be suitable for a semi-supervised regression algorithm. It is also important that unlabeled regressors are available and comparably “cheap” to get as opposed to the labeled regressors.

The chapter is organized as follows: We start off by giving a background to semi-supervised learning and an overview of previous work, Sect. 2. We thereafter formalize the assumptions under which unlabeled data has potential to be useful, Sect. 3. A semi-supervised regression algorithm is described in Sect. 4 and exemplified in Sect. 5. In Sect. 6 we discuss the application to dynamical systems and we end by a conclusion in Sect. 7.

2 Background

Semi-supervised learning has been around since the 1970s (some earlier attempts exist). *Fisher's linear discriminant rule* was then discussed under the assumption that each of the class conditional densities was Gaussian. *Expectation maximization* was applied using both labeled and unlabeled regressors to find the parameters of the Gaussian densities [10]. During the 1990s the interest for semi-supervised learning increased, mainly due to its application to text classification, see e.g. [15]. The first usage of the word *semi-supervised* learning, as it is used today, was not until 1992 [12].

The boost in the area of manifold learning in the 1990s brought with it a number of semi-supervised methods. *Semi-supervised manifold learning* is a type of semi-supervised learning in which the map found by an unsupervised manifold learning algorithm is restricted by giving a number of labeled regressors as examples for how that map should be. Most of the algorithms are extensions of unsupervised manifold learning algorithms, see among others [2, 26, 14, 20, 19, 16, 27]. Another interesting contribution is the developments by Rahimi in [18]. A time series of regressors, some labeled and some unlabeled, are considered there. The series of labels best fitting the given labels and at the same time satisfying some temporal smoothness assumption is then computed.

Most of the references above are to semi-supervised classification algorithms. They are however relevant since most semi-supervised classification methods can, with minor modifications, be applied to regression problems. The modification or the application to regression problems are however almost never discussed or exemplified. For more historical notes on semi-supervised learning, see [4].

3 The Semi-Supervised Smoothness Assumption

We are in regression interested in finding estimates for the conditional distribution $p(y|\varphi)$. For the unlabeled regressors to be useful, it is required that the regressor distribution $p(\varphi)$ brings information concerning the conditional $p(y|\varphi)$.

We saw from the pictorial example in Sect. 1 that one situation for which this is the case is when we make the assumption that the label changes continuously along high-density areas in the regressor space. This assumption is referred to as the *semi-supervised smoothness assumption* [4]:

Assumption 1 (Semi-Supervised Smoothness). *If two regressors $\varphi(1)$, $\varphi(2)$ in a high-density region are close, then so should their labels.*

“High density region” is a somewhat loose term: In many cases it corresponds to a manifold in the regressor space, such that the regressors for the application in question are confined to this manifold. That two regressors are “close” then means that the distance between them along the manifold (the geodesic distance) is small.

In classification, this smoothness assumption is interpreted as that the class labels should be the same in the high-density regions. In regression, we interpret this as a slowly varying label along high-density regions. Note that in regression, it is common to assume that the label varies smoothly in the regressor space; the semi-supervised smoothness assumption is less conservative since it only assumes smoothness in the high-density regions in the regressor space. Two regressors could be close in the regressor space metric, but far apart along the high density region (the manifold): think of the region being a spiral in the regressor space.

4 Semi-Supervised Regression: WDMR

Given a particular regressor φ^* , consider the problem of finding an estimate for $f(\varphi^*)$ given the measurements $\{(y(t), \varphi(t))\}_{t=1}^{N_l}$ generated by

$$y = f(\varphi) + e, \quad e \sim \mathcal{N}(0, \sigma). \quad (4)$$

This is a supervised regression problem. If unlabeled regressors $\{\varphi(t)\}_{t=N_l+1}^{N_l+N_u}$ are used as well, the regression becomes semi-supervised. Since we in the following will make no difference between the unlabeled regressor φ^* and $\{\varphi(t)\}_{t=N_l+1}^{N_l+N_u}$, we simply include φ^* in the set of unlabeled regressors to make the notation a bit less cluttered. We let $\hat{f}_t$ denote the estimates of $f(\varphi(t))$ and assume that $f : \mathcal{R}^{n_\varphi} \rightarrow \mathcal{R}$ for simplicity. In the following we will also need to introduce kernels as distance measure in the regressor space. To simplify the notation, we will use K_{ij} to denote a kernel $k(\cdot, \cdot)$ evaluated at the regressor pair $(\varphi(i), \varphi(j))$ i.e., $K_{ij} \triangleq k(\varphi(i), \varphi(j))$. A popular choice of kernel is the Gaussian kernel

$$K_{ij} = e^{-\|\varphi(i) - \varphi(j)\|^2 / 2\sigma^2}. \quad (5)$$

Since we will consider regressors constrained to certain regions of the regressor space (often manifolds), kernels constructed from manifold learning techniques, see Sect. 4.1, will be of particular interest. Notice however that we will allow us to use a kernel like

$$K_{ij} = \begin{cases} \frac{1}{K}, & \text{if } \varphi(j) \text{ is one of the } K \text{ closest} \\ & \text{neighbors of } \varphi(i), \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

and K_{ij} will therefore not necessarily be equal to K_{ji} . We will also always use the convention that $K_{ij} = 0$ if $i = j$.

Under the semi-supervised smoothness assumption, we would like the estimates belonging to two regressors which are close in a high-density region to have similar values. Using a kernel, we can express this as

$$\hat{f}_t = \sum_{i=1}^{N_l+N_u} K_{ti} \hat{f}_i, \quad t = 1 \dots N_l + N_u \quad (7)$$

where K_{ti} is a kernel giving a measure of distance between $\varphi(t)$ and $\varphi(i)$, relevant to the assumed region.

So the sought estimates $\hat{f}_i$ should be such that they are smooth over the region. At the same time, for regressors with measured labels, the estimates should be close to those, meaning that

$$\sum_{t=1}^{N_l} (y(t) - \hat{f}_t)^2 \quad (8)$$

should be small. The two requirements (7) and (8) can be combined into a criterion

$$\lambda \sum_{i=1}^{N_l+N_u} (\hat{f}_i - \sum_{j=1}^{N_l+N_u} K_{ij} \hat{f}_j)^2 + (1-\lambda) \sum_{t=1}^{N_l} (y(t) - \hat{f}_t)^2 \quad (9)$$

to be minimized with respect to $\hat{f}_t$, $t = 1, \dots, N_l + N_u$. The scalar λ decides how trustworthy our labels are and is seen as a design parameter.

The criterion (9) can be given a Bayesian interpretation as a way to estimate $\hat{f}$ in (8) with a “smoothness prior” (7), with λ reflecting the confidence in the prior.

Introducing the notation

$$\begin{aligned} J &\triangleq [I_{N_l \times N_l} \ 0_{N_l \times N_u}], \\ \mathbf{y} &\triangleq [y(1) y(2) \dots y(N_l)]^T, \\ \hat{\mathbf{f}} &\triangleq [\hat{f}_1 \hat{f}_2 \dots \hat{f}_{N_l} \hat{f}_{N_l+1} \dots \hat{f}_{N_l+N_u}]^T, \\ K &\triangleq \begin{bmatrix} K_{11} & K_{12} & \dots & K_{1,N_l+N_u} \\ K_{21} & K_{22} & & K_{2,N_l+N_u} \\ \vdots & & \ddots & \vdots \\ K_{N_l+N_u,1} & K_{N_l+N_u,2} & \dots & K_{N_l+N_u,N_l+N_u} \end{bmatrix}, \end{aligned}$$

(9) can be written as

$$\lambda (\hat{\mathbf{f}} - K\hat{\mathbf{f}})^T (\hat{\mathbf{f}} - K\hat{\mathbf{f}}) - (1-\lambda) (\mathbf{y} - J\hat{\mathbf{f}})^T (\mathbf{y} - J\hat{\mathbf{f}}) \quad (10)$$

which expands into

$$\hat{\mathbf{f}}^T (\lambda (I - K - K^T + K^T K) - (1-\lambda) J^T J) \hat{\mathbf{f}} + 2(1-\lambda) \hat{\mathbf{f}}^T J^T \mathbf{y} + (1-\lambda) \mathbf{y}^T \mathbf{y}. \quad (11)$$

Setting the derivative with respect to $\hat{\mathbf{f}}$ to zero and solving gives the linear kernel smoother

$$\hat{\mathbf{f}} = (1-\lambda) (\lambda (I - K - K^T + K^T K) - (1-\lambda) J^T J)^{-1} J^T \mathbf{y}. \quad (12)$$

This regression procedure uses all regressors, both unlabeled and labeled, and is hence a semi-supervised regression algorithm. We call the kernel smoother *Weight Determination by Manifold Regularization* (WDMR, [16]). In this case the unlabeled

beled regressors are used to get a better knowledge for what parts of the regressor space that the function f varies smoothly in.

Similar methods to the one presented here has also been discussed in [8, 26, 3, 2, 25]. [26] discusses manifold learning and construct a semi-supervised version of the manifold learning technique *Locally Linear Embedding* (LLE, [21]) which coincides with a particular choice of kernel in (9). More details about this kernel choice will be given in the next section. [8] studies graph based semi-supervised methods for classification and derives a similar objective function as (9). [3, 25] discuss a classification method called *label propagation* which is an iterative approach converging to (12). In [2], support vector machines is extended to work under the semi-supervised smoothness assumption.

4.1 LLE: A Way of Selecting the Kernel in WDMR

Local Linear Embedding, LLE, [21] is a technique to find lower dimensional manifolds to which an observed collection of regressors belong. A brief description of it is as follows:

Let $\{\varphi(i), i = 1, \dots, N\}$ belong to $U \subset R^{n_\varphi}$ where U is an unknown manifold of dimension n_z . A coordinatization $z(i)$, ($z(i) \in R^{n_z}$) of U is then obtained by first minimizing the cost function

$$\varepsilon(l) = \sum_{i=1}^N \left\| \varphi(i) - \sum_{j=1}^N l_{ij} \varphi(j) \right\|^2 \quad (13a)$$

under the constraints

$$\begin{cases} \sum_{j=1}^N l_{ij} = 1, \\ l_{ij} = 0 \text{ if } \|\varphi(i) - \varphi(j)\| > C_i(K) \text{ or if } i = j. \end{cases} \quad (13b)$$

Here, $C_i(K)$ is chosen so that only K weights l_{ij} become nonzero for every i . K is a design variable. It is also common to add a regularization to (13a) not to get degenerate solutions.

Then for the determined l_{ij} find $z(i)$ by minimizing

$$\sum_{i=1}^N \left\| z(i) - \sum_{j=1}^N l_{ij} z(j) \right\|^2 \quad (14)$$

wrt $z(i) \in R^{n_z}$ under the constraint

$$\frac{1}{N} \sum_{i=1}^N z(i) z(i)^T = I_{n_z \times n_z}$$

$z(i)$ will then be the coordinate for $\varphi(i)$ in the lower dimensional manifold.

The link between WDMR and LLE is now clear: If we pick the kernel K_{ij} in (9) as l_{ij} from (13) and have no labeled regressors ($N_l = 0$) and add the constraint $\frac{1}{N_u} \hat{\mathbf{f}}^T \hat{\mathbf{f}} = I_{n_z \times n_z}$, minimization of the WDMR criterion (9) will yield $\hat{f}_i$ as the LLE coordinates $z(i)$.

In WDMR with labeled regressors, the addition of the criterion (8) will replace the constraint $\frac{1}{N_u} \hat{\mathbf{f}}^T \hat{\mathbf{f}} = I_{n_z \times n_z}$ as an anchor to prevent a trivial zero solution. Thus WDMR is a natural semi-supervised version of LLE, [16].

4.2 A Comparison with K Nearest Neighbor Averages: K -NN

It is interesting to notice the difference between using the kernel given in (6) and

$$K_{ij} = \begin{cases} \frac{1}{K}, & \text{if } \varphi(j) \text{ is one of the } K \text{ closest} \\ & \text{labeled neighbors of } \varphi(i), \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

To illustrate the difference, let us return to the pictorial example discussed in Fig. 2. We now add 5 unlabeled regressors to the 5 previously considered. Hence we have 10 regressors, 4 labeled and 6 unlabeled, and we desire an estimate of the label marked with a question mark in Fig. 3. The left part of Fig. 3 shows how WDMR solves the estimation problem if the kernel in (15) is used. Since the kernel will cause the searched label to be similar to the label of the K closest labeled regressors, the result will be similar to using the algorithm *K-nearest neighbor average* (K -NN, see e.g. [9]). In the right part of Fig. 3, WDMR with the kernel given in (6) is used. This kernel grants estimates of the K closest regressors (labeled or not) to be similar. Since the closest regressors, to the regressor for which we search the label, are unlabeled, information is propagated from the labeled regressors towards the one for which we search a label along the chain of unlabeled regressors. The shaded regions in both the left and right part of the figure symbolize the way information is propagated using the different choices of kernels. In the left part of the figure we will therefore obtain an estimate equal to 2.5 while in the right we get an estimate equal to 1.

5 Examples

We give in the following two examples of regression problem for which the semi-supervised smoothness assumption is motivated. Estimates are computed using WDMR and comparisons to conventional supervised regression methods are given.

Fig. 3 An illustration of the difference of using the kernel given in (15) (left part of the figure) and (6) (right part of the figure).

5.1 fMRI

Functional Magnetic Resonance Imaging, fMRI is a technique to measure brain activity. The fMRI measurements give a measure of the degree of oxygenation in the blood, it measures the *Blood Oxygenation Level Dependent* (BOLD) response. The degree of oxygenation reflects the neural activity in the brain and fMRI is therefore an indirect measure of brain activity.

Measurements of brain activity can with fMRI be acquired as often as once a second and are given as an array, each element giving a scalar measure of the average activity in a small volume element of the brain. These volume elements are commonly called *voxels* (short for *volume pixel*) and they can be as small as one cubic millimeter. The fMRI measurements are heavily affected by noise.

In this example, we consider measurements from an $8 \times 8 \times 2$ array covering parts of the visual cortex gathered with a sampling period of 2 seconds. To remove noise, data was prefiltered by applying a spatial and temporal filter with a Gaussian kernel. The filtered fMRI measurements at each time t were vectorized into the regression vector $\phi(t)$. fMRI data was acquired during 240 seconds (giving 120 samples, since the sampling period was 2 seconds) from a subject that was instructed to look away from a flashing checkerboard covering 30% of the field of view. The flashing checkerboard moved around and caused the subject to look to the left, right, up and down. The direction in which the person was looking was seen as the label. The label was chosen to be 0 when the subject was looking to the right, $\pi/2$ when looking up, π when looking to the left and $-\pi/2$ when looking down.

The direction in which the person was looking is described by its angle, a scalar. The fMRI data should hence be constrained to a one-dimensional closed manifold residing in the 128 dimensional regressor space (since the regressors can be parameterized by the angle). If we assume that the semi-supervised smoothness assumption holds, WDMR therefore seems like a good choice.

The 120 labeled regressors were separated into two sets, a training set consisting of 80 labeled regressors and a test set consisting of 40 labeled regressors. The training set was further divided into an estimation set and a validation set, both of the same size. The estimation set and the regressors of the validation set were used

in WDMR. The estimated labels of the validation regressors were compared to the measured labels and used to determine the design parameters. λ in (9) was chosen as 0.8 and K (using the kernel determined by LLE, see (13)) as 6. The tuned WDMR regression algorithm was then used to predict the direction in which the person was looking. The result from applying WDMR to the 40 regressors of the test set are shown in Fig. 4.

The result is satisfactory but it is not clear to what extent the one-dimensional manifold has been found. The number of unlabeled regressors used are rather low and it is therefore not surprising that K-NN can be shown to do almost as good as WDMR in this example. One would expect that adding more unlabeled regressors would improve the result obtained by WDMR. The estimates of K-NN would however stay unchanged since K-NN is a supervised method and therefore not affected by unlabeled data.

Fig. 4 WDMR applied to brain activity measurements (fMRI) of the visual cortex in order to tell in what direction the subject in the MR scanner was looking. Thin gray line shows the direction in which the subject was looking and thick black line, the estimated direction by WDMR.

5.2 Climate Reconstruction

There exist a number of climate recorders in nature from which the past temperature can be extracted. However, only a few natural archives are able to record climate fluctuations with high enough resolution so that the seasonal variations can be reconstructed. One such archive is a bivalve shell. The chemical composition of a shell of a bivalve depends on a number of chemical and physical parameters of the water in which the shell was composed. Of these parameters, the water temperature is probably the most important one. It should therefore be possible to estimate the water temperature for the time the shell was built, from measurements of the shell's chemical composition. This would e.g. give climatologists the ability to estimate past water temperatures by analyzing ancient shells.

In this example, we used 10 shells grown in Belgium. Since the temperature in the water had been monitored for these shells, this data set provides excellent means

to test the ability to predict water temperature from chemical composition measurements. For these shells, the chemical composition measurements had been taken along the growth axis of the shells and paired up with temperature measurements. Between 30 and 52 measurements were provided from each shell, corresponding to a time period of a couple of months. The 10 shells were divided into an estimation set and a validation set. The estimation set consisted of 6 shells (a total of 238 labeled regressors) grown in Terneuzen in Belgium. Measurements from five of these shells are shown in Fig. 6. The figure shows measurements of the relative concentrations of Sr/Ca, Mg/Ca and Ba/Ca (Pb/Ca is also measured but not shown in the figure). The line shown between measurements connects the measurements coming from a shell and gives the chronological order of the measurements (two in time following measurements are connected by a line).

Fig. 5 A plot of the Sr/Ca, Mg/Ca and Ba/Ca concentration ratio measurements from five shells. Lines connects measurements (ordered chronologically) coming from the same shell. The temperatures associated with the measurements were color coded and are shown as different gray scales on the measurement points.

As seen in the figure, measurements are highly restricted to a small region in the measurement space. Also, the water temperature (gray level coded in Fig. 6) varies smoothly in the high-density regions. This together with that it is a biological process generating data, motivates the semi-supervised smoothness assumption when trying to estimate water temperature (labels) from chemical composition measurements (4-dimensional regressors).

The four shells in the validation set came from four different sites (Terneuzen, Breskens, Ossensisse, Knokke) and from different time periods. The estimated temperatures for the validation data obtained by using WDMR with the kernel determined by LLE (see (13)) are shown in Fig. 6. For comparison purpose, it could be mentioned that K-NN had a *Mean Absolute Error* (MAE) nearly twice as high as WDMR.

A more detailed discussion of this example is presented in [1]. The data sets used were provided by Vander Putten and colleagues [17] and Gillikin and colleagues [6, 7].

Fig. 6 Water temperature estimations using WDMR for validation data (thick line) and measured temperature (thin line). From top to bottom figure: Terneuzen, Breskens, Ossenisse, Knokke.

6 Dynamical Systems

6.1 Analysis of a Circadian Clock

The circadian rhythmic living among people and animals are kept by robustly coupled chemical processes in cells of the suprachiasmatic nucleus (SCN) in the brain. The whole system is affected by light and goes under the name *biological clock*. The biological clock synchronizes the periodic behavior of many chemical processes in the body and is crucial for the survival of most species.

The chemical processes cause protein and messenger RNA (mRNA) concentrations in the cells of the SCN to fluctuate. The “freerunning” rhythm (no external input) of the fluctuations is however not the same as the light/dark cycle and environmental cues, such as light, cause it to synchronize with the environmental rhythm. We use the nonlinear biological clock model by [22]

$$\frac{dM(t)}{dt} = \frac{r_M(t)}{1 + P(t)^2} - 0.21M(t), \quad (16)$$

$$\frac{dP(t)}{dt} = M(t-4)^3 - 0.21P(t), \quad (17)$$

to generate simulated data and simulate the affect of the light cue by letting the mRNA production rate r_M vary periodically. M and P are the relative concentrations of mRNA and the protein. Figure 7 shows the (periodic) response of P to the (periodic) stimuli r_M . We see r_M as input and P as output and we want to predict P from measured $r_M(t)$. The measurements of P are rather costly in real applications, while $r_M(t)$ can be inferred from simple measurements of the light.

We seek to describe the output P as a nonlinear FIR (NFIR) model from two previous inputs $[r_M(t) \ r_M(t-4)]^T$ (the regression vector) and collect 230 measurements of this regression vector. Only 6 out of these are labeled by the corresponding $P(t)$. We thus have a situation (3) with $N_l = 6$ and $N_u = 224$. Applying the WDMR algorithm (12) with $\lambda = 0.5$ and the kernel defined in (6) (with $K = 4$) gives an estimate of P corresponding to all the 230 time points. This estimate is shown in Fig. 8 together with the true values.

We note that the estimate is quite good, despite the very small number of labeled measurements. In this case the two-dimensional regression vector is confined to a 1-dimensional manifold (this follows since r_M is periodic: one full period will create a track in the regressor space that can be parameterized by the scalar time variable over one period). This means that this application can make full use of the dimension reduction that is inherent in WDMR. On the other hand, the model is tailored to the specific choice of input. (This, by the way, is true for any non-linear identification method)

Fig. 7 The circadian clock is affected by light. This is in Example 6.1 modeled by letting r_M vary in a periodic manner. One period of r_M (thin gray line) and a period of P (thick black line) are shown in the figure. The synchronization between r_M and P is characteristic for a circadian clock and crucial for surviving.

Let us compare with the estimates obtained by K-NN, using the K-NN kernel given in (15). The dashed line in Fig. 8 shows the estimated protein levels using the

K-NN (using only the labeled regressors). Since using only one neighbor ($K = 1$) gave the best result in K-NN, only this result is shown.

Fig. 8 Estimated relative protein concentration by K-NN ($K = 1$ gave the best result and therefore shown) and WDMR using the K-nearest neighbor kernel ($K = 4$ gave the best result and therefore shown). K-NN: dashed gray line; true P : solid black line; WDMR: solid gray line; estimation data: filled circles.

The result shown in Fig. 8 confirms the previous discussion around the pictorial example, see Fig. 3. K-NN average together the Euclidean closest regressors' labels while WDMR search for labeled regressors along the manifold and then assumes a slowly varying function along the manifold.

6.2 The Narendra–Li System

Let us now consider a standard test example from [13], “the Narendra–Li example”:

$$x_1(t+1) = \left(\frac{x_1(t)}{1+x_1^2(t)} + 1 \right) \sin(x_2(t)) \quad (18a)$$

$$x_2(t+1) = x_2(t) \cos(x_2(t)) + x_1(t) \exp\left(-\frac{x_1^2(t) + x_2^2(t)}{8}\right) + \frac{u^3(t)}{1 + u^2(t) + 0.5 \cos(x_1(t) + x_2(t))} \quad (18b)$$

$$y(t) = \frac{x_1(t)}{1 + 0.5 \sin(x_2(t))} + \frac{x_2(t)}{1 + 0.5 \sin(x_1(t))} + e(t) \quad (18c)$$

This dynamical system was simulated with 2000 samples using a random binary input, giving input output data $\{y(t), u(t), t = 1, \dots, 2000\}$. A separate set of 50 validation data were also generated with a sinusoidal input. The chosen regression vector was

$$\boldsymbol{\varphi}(t) = [y(t-1) \ y(t-2) \ y(t-3) \ u(t-1) \ u(t-2) \ u(t-3)]^T \quad (19)$$

A standard Sigmoidal Neural Network model with one hidden layer with 18 (gave the best result) units was applied to this data set, and a corresponding NLARX model $y(t) = f(\boldsymbol{\varphi}(t))$ was constructed. The prediction performance for the validation data is illustrated in Fig. 9. As a numerical measure of how good the prediction is, the “fit” is shown in the figure. The fit is the relative norm of the difference between the curves, expressed in %. 100% is thus a perfect fit.

Fig. 9 One step ahead prediction for the models of the Narendra-Li example. Top: Neural Network (18 units); middle: WDMR; bottom: K-Nearest Neighbor (K=15). Thin line: true validation outputs; thick line: model output.

The semi-supervised algorithm, WDMR (12) with a kernel determined from LLE (as described in (13)) was also applied to these data. Then unlabeled regression vectors from the validation data was appended to the estimation data. The resulting prediction performance is also shown in Fig. 9.

We see that WDMR gives a significantly better model than the standard neural network technique. In this case it is not clear that the regressors are constrained to a manifold. Therefore *semi*-supervised aspect is not so pronounced, and anyway the (validation) set of unlabeled regressors is quite small in comparison to the labeled

(estimation) ones. WDMR in this case can be seen as a kernel method, and the message is perhaps that the neural network machinery is too heavy an artillery for this application. For comparison we also computed a K-nearest neighbor model for the same data. Experiments showed that $K = 15$ neighbors gave the best prediction fit to validation data, and the result is also depicted in Fig. 9. It is better than a Neural Network, but worse than WDMR.

In System Identification it is common that the regression vector contains old outputs as in (19). Then it is not so natural to think of “unlabeled” regressor sets, since they would contain outputs = “labels”, for other regressors. But WDMR still provides a good algorithms as we saw in the example.

Also, one may discuss how common it is in system identification that the regressors are constrained to a manifold. The input signal part of the regression vector should according to identification theory be “persistently exciting” which is precisely the opposite of being constrained. However, in many biological applications and in DAE (differential algebraic equation) modeling such structural constraints are frequently occurring.

Anyway, even in the absence of manifold constraints it may be a good idea to require smoothness in dense regressor regions as in (18). The Narendra-Li example showed the benefits of WDMR also in this more general context.

7 Conclusion

The purpose of this contribution was to explore what current techniques typical in machine learning has to offer for system identification problems. We outlined the ideas behind semi-supervised learning, that even regressors without corresponding outputs can improve the model fit, due to inherent constraints in the regressor space. We described a particular method, WDMR, which we believe to be novel, how to use both labeled and unlabeled regressors in regression problems. The usefulness of this method was illustrated on a number of examples, including some problems of a traditional non-linear system identification character. Even though WDMR compared favorably to more conventional methods for these problems, further analysis and comparisons must be made before a full evaluation of this approach can be made.

Acknowledgements This work was supported by the Strategic Research Center MOVIII, funded by the Swedish Foundation for Strategic Research, SSF, and CADICS, a Linnaeus center funded by the Swedish Research Council.

References

1. Bauwens, M., Ohlsson, H., Barbé, K., Beelaerts, V., Dehairs, F., Schoukens, J.: On climate reconstruction using bivalve shells: Three methods to interpret the chemical signature of a shell. In: 7th IFAC Symposium on Modelling and Control in Biomedical Systems (2009)
2. Belkin, M., Niyogi, P., Sindhiani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.* **7**, 2399–2434 (2006)
3. Bengio, Y., Delalleau, O., Le Roux, N.: Label propagation and quadratic criterion. In: O. Chapelle, B. Schölkopf, A. Zien (eds.) *Semi-Supervised Learning*, pp. 193–216. MIT Press (2006)
4. Chapelle, O., Schölkopf, B., Zien, A. (eds.): *Semi-Supervised Learning*. MIT Press, Cambridge, MA (2006)
5. Cybenko, G.: Just-in-time learning and estimation. In: S. Bittanti, G. Picci (eds.) *Identification, Adaptation, Learning. The Science of Learning Models from data*, NATO ASI Series, pp. 423–434. Springer (1996)
6. Gillikin, D.P., Dehairs, F., Lorrain, A., Steenmans, D., Baeyens, W., André, L.: Barium uptake into the shells of the common mussel (*Mytilus edulis*) and the potential for estuarine paleochemistry reconstruction. *Geochimica et Cosmochimica Acta* **70**(2), 395–407 (2006)
7. Gillikin, D.P., Lorrain, A., Bouillon, S., Willenz, P., Dehairs, F.: Stable carbon isotopic composition of *Mytilus edulis* shells: relation to metabolism, salinity, $\delta^{13}C_{DIC}$ and phytoplankton. *Organic Geochemistry* **37**(10), 1371–1382 (2006)
8. Goldberg, A.B., Zhu, X.: Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In: *HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing* (2006)
9. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA (2001)
10. Hosmer David W., J.: A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics* **29**(4), 761–770 (1973)
11. Kohonen, T.: *Self-Organizing Maps*, *Springer Series in Information Sciences*, vol. 30. Springer, Berlin (1995)
12. Merz, C., St. Clair, D., Bond, W.: Semi-supervised adaptive resonance theory (smart2). In: *Neural Networks, 1992. IJCNN., International Joint Conference on*, vol. 3, pp. 851–856 (1992)
13. Narendra, K.S., Li, S.M.: Neural networks in control systems. In: P. Smolensky, M.C. Mozer, D.E. Rumelhard (eds.) *Mathematical Perspectives on Neural Networks*, chap. 11, pp. 347–394. Lawrence Erlbaum Associates (1996)
14. Navaratnam, R., Fitzgibbon, A., Cipolla, R.: The joint manifold model for semi-supervised multi-valued regression. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* pp. 1–8 (2007)
15. Nigam, K., McCallum, A., Thrun, S., Mitchell, T.: Learning to classify text from labeled and unlabeled documents. In: *AAAI '98/IAAI '98: Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, pp. 792–799. American Association for Artificial Intelligence, Menlo Park, CA, USA (1998)
16. Ohlsson, H., Roll, J., Ljung, L.: Manifold-constrained regressors in system identification. In: *Proc. 47th IEEE Conference on Decision and Control*, pp. 1364–1369 (2008)
17. Putten, E.V., Dehairs, F., André, L., Baeyens, W.: Quantitative in situ microanalysis of minor and trace elements in biogenic calcite using infrared laser ablation - inductively coupled plasma mass spectrometry: a critical evaluation. *Analytica Chimica Acta* **378**(1-3), 261–272 (1999)
18. Rahimi, A., Recht, B., Darrell, T.: Learning to transform time series with a few examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**(10), 1759–1775 (2007)
19. de Ridder, D., Duin, R.: Locally linear embedding for classification (2002). Tech Report, PH-2002-01, Pattern Recognition Group, Dept. of Imaging Science & Technology, Delft University of Technology, Delft, The Netherlands.

20. de Ridder, D., Kouropteva, O., Okun, O., Pietikäinen, M., Duin, R.: Artificial Neural Networks and Neural Information Processing – ICANN/ICONIP 2003, *Lecture Notes in Computer Science*, vol. 2714, chap. Supervised locally linear embedding, pp. 333–341. Springer Berlin / Heidelberg (2003)
21. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **290**(5500), 2323–2326 (2000)
22. Scheper, T.o., Klinkenberg, D., Pennartz, C., van Pelt, J.: A Mathematical Model for the Intracellular Circadian Rhythm Generator. *J. Neurosci.* **19**(1), 40–47 (1999)
23. Stenman, A.: Model on demand: Algorithms, analysis and applications. Linköping Studies in science and Technology. Thesis No 571, Linköping University, SE-581 83 Linköping, Sweden (1999)
24. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. *Science* **290**(5500), 2319–2323 (2000)
25. Wang, F., Zhang, C.: Label propagation through linear neighborhoods. *Knowledge and Data Engineering, IEEE Transactions on* **20**(1), 55–67 (2008)
26. Yang, X., Fu, H., Zha, H., Barlow, J.: Semi-supervised nonlinear dimensionality reduction. In: ICML '06: Proceedings of the 23rd international conference on Machine learning, pp. 1065–1072. ACM, New York, NY, USA (2006)
27. Zhao, L., Zhang, Z.: Supervised locally linear embedding with probability-based distance for classification. *Comput. Math. Appl.* **57**(6), 919–926 (2009)
28. Zhu, X.: Semi-supervised learning literature survey. Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison (2005)

Avdelning, Institution
Division, Department

Division of Automatic Control
Department of Electrical Engineering

Datum
Date

2010-04-25

Språk

Language

☐ Svenska/Swedish

☒ Engelska/English

☐ _____

Rapporttyp

Report category

☐ Licentiatavhandling

☐ Examensarbete

☐ C-uppsats

☐ D-uppsats

☒ Övrig rapport

☐ _____

ISBN

—

ISRN

—

Serietitel och serienummer

Title of series, numbering

ISSN

1400-3902

URL för elektronisk version

<http://www.control.isy.liu.se>

LiTH-ISY-R-2940

Titel

Title

Semi-Supervised Regression and System Identification

Författare

Author

Henrik Ohlsson, Lennart Ljung

Sammanfattning

Abstract

System Identification and Machine Learning are developing mostly as independent subjects, although the underlying problem is the same: To be able to associate “outputs” with “inputs”. Particular areas in machine learning of substantial current interest are *manifold learning* and *unsupervised* and *semi-supervised regression*. We outline a general approach to semi-supervised regression, describe its links to Local Linear Embedding, and illustrate its use for various problems. In particular, we discuss how these techniques have a potential interest for the system identification world.

Nyckelord

Keywords

Semi-supervised regression, system identification