Using Manifold Learning for Nonlinear System Identification

Henrik Ohlsson, Jacob Roll, Torkel Glad, Lennart Ljung

Division of Automatic Control

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

13th June 2007

Report no.: [LiTH-ISY-R-2795](#)

Accepted for publication in 7th IFAC Symposium on Nonlinear Control Systems (NOLCOS 2007)

Address:

Department of Electrical Engineering

Linköpings universitet

SE-581 83 Linköping, Sweden

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

AUTOMATIC CONTROL
REGLERTEKNIK
LINKÖPINGS UNIVERSITET

Abstract:

A high-dimensional regression space usually causes problems in nonlinear system identification. However, if the regression data are contained in (or spread tightly around) some manifold, the dimensionality can be reduced. This paper presents a use of dimension reduction techniques to compose a two-step identification scheme suitable for high-dimensional identification problems with manifold-valued regression data. Illustrating examples are also given.

Keywords: Nonlinear system identification; Dimension reduction techniques; Manifold learning

USING MANIFOLD LEARNING FOR NONLINEAR SYSTEM IDENTIFICATION

Henrik Ohlsson * Jacob Roll * Torkel Glad *
Lennart Ljung *

** Division of Automatic Control
Department of Electrical Engineering,
Linköpings universitet, SE-581 83 Linköping, Sweden
{ohlsson, roll, torkel, ljung}@isy.liu.se*

Abstract: A high-dimensional regression space usually causes problems in nonlinear system identification. However, if the regression data are contained in (or spread tightly around) some manifold, the dimensionality can be reduced. This paper presents a use of dimension reduction techniques to compose a two-step identification scheme suitable for high-dimensional identification problems with manifold-valued regression data. Illustrating examples are also given.

Keywords: Nonlinear system identification; Dimension reduction techniques; Manifold learning

1. INTRODUCTION

While the theory and methods of linear system identification are generally well developed, large parts of nonlinear system identification still remain an open field. Classical nonlinear identification methods commonly fail to handle identification of high-dimensional problems well. One reason for this is what is generally called the curse of dimensionality, which is exemplified in most standard books of system identification (Ljung, 1999) and statistical learning (Hastie *et al.*, 2001) and a known problem.

In many practical problems, however, the real underlying structure has an inherent lower-dimensionality, which could be exploited to reduce the complexity of the problem. We can distinguish between two cases:

- (1) The outputs are functions only of, e.g., a small number of linear combinations of the regressors.
- (2) The regression vectors are manifold-valued, i.e., constrained to a lower-dimensional manifold.

A way to treat the former problem is to project high-dimensional data down to a hyperplane prior to iden-

tification. Different approaches for finding an appropriate projection for this purpose are also presented in (Li, 1991; Li, 1992; Fukumizu *et al.*, 2004; Lindgren, 2005). The projection is linear, which works well for many problems, but not all kinds of nonlinear problems can be handled satisfactorily. Nonlinear approaches, such as Nonlinear Component Analysis (Schölkopf *et al.*, 1998), can work well with a good choice of kernels, but without insight, the choice of kernels can be considerably hard.

In this paper, we will mainly concentrate on problem (2), when data are contained in an unknown, lower-dimensional manifold. In many examples there are connections between the inputs/regressors, which constrain data to some manifold. One example is a sensor fusion problem, where by combining the many measurements from a number of sensors, a better measure can be obtained compared to using a single sensor. The number of sensors can be large, which gives a large number of regressors or inputs from which a function has to produce an as accurate estimate of the position, velocity etc. as possible. In this case, if noise is neglected, the regression data are all on a

manifold embedded in the high-dimensional space of inputs/regressors. For instance, six inaccurate sensors, used to measure the acceleration for an object in a three-dimensional space, would yield an input data space of dimension 18, while the measurements would all tend to some three-dimensional manifold.

Other examples that lead to regressors being connected occur in the field of differential-algebraic equations (DAE:s, (Kunkel and Mehrmann, 2006)). DAE:s have states that must satisfy some algebraic conditions which makes the states manifold-valued. See Section 2 for a more detailed description. Likewise, data collected under (unknown) feedback will also lead to manifold-valued regression vectors, although in this case, no substantial dimensionality reduction can be expected in general.

Yet other examples can be found in the fields of face and speech recognition, where the dimensionality of input data is an obvious problem. From a picture of a face consisting of a huge number of pixels, the dimensionality has to be substantially reduced, to allow sorting into subgroups of people of similar face characteristics. This is a classification problem where data also tend to lie on manifolds. All faces have a mouth, a nose, etc., and these kinds of similarities among faces restrict the data to certain areas/volumes and make the data manifold-valued. Interest in this field has developed many theories for how to reduce the dimension taking into account that data are manifold-valued. These techniques are commonly called manifold learning techniques, see, e.g., (Roweis and Saul, 2000; Tenenbaum *et al.*, 2000). A recent overview is given in (Brun, 2006).

Manifold learning algorithms have previously been used for classification. See for example (de Ridder *et al.*, 2003; Kouropteva *et al.*, 2002; Zhao *et al.*, 2005), which propose supervised local linear embedding (SLLE).

Our goal in this paper is to examine the possibilities of using a manifold learning techniques for nonlinear system identification, by making a prior dimension reduction followed by an identification, similarly to the approach in (Li, 1991; Li, 1992; Fukumizu *et al.*, 2004; Lindgren, 2005), but using nonlinear techniques for the dimension reduction. The focus will be on the case of manifold-valued regression data. We will use manifold learning algorithms in combination with an identification step, in order to create an attractive two-step identification procedure for high-dimensional problems. The results presented here should be regarded as a first step in this direction. As an example, a nonlinear system with manifold-valued regressors is identified using a reduction method named *Local Linear Embedding* (LLE, (Roweis and Saul, 2000)).

The paper is structured as follows: Section 2 investigates connections between discrete-time DAE:s and

manifold-valued data. Section 3 gives an overview of the suggested method, which is then illustrated by two examples in Section 4.

2. DIFFERENTIAL-ALGEBRAIC EQUATIONS AND MANIFOLD-VALUED DATA

As mentioned in the introduction, working in the context of differential-algebraic equations (DAE:s) may give rise to regressors being connected. DAE:s naturally occur in many physical modeling contexts, in particular when using object-oriented modeling tools, such as Modelica (Fritzson, 2004; Tiller, 2001). A general DAE can be written as (Kunkel and Mehrmann, 2006)

$$F(\dot{x}, x, t) = 0 \quad (1)$$

of which state-space models form a special case. For “true” DAE:s (that cannot be directly written in state-space form), (1) may implicitly define algebraic relations between the states. More explicitly, for a large class of DAE:s (for more details, see (Kunkel and Mehrmann, 2006)) x may be partitioned into three parts, and (1) may be reformulated as

$$\begin{aligned} \dot{x}_1 &= \mathcal{L}(x_1, x_2, \dot{x}_2, t) \\ x_3 &= \mathcal{R}(x_1, x_2, t) \end{aligned}$$

Here, the second relation is an algebraic relation, which constrains x to a lower-dimensional manifold.

In the field of system identification, it is usually assumed that the experimental data is sampled at discrete timepoints, and hence it is natural to consider discrete-time models. A discrete-time DAE corresponding to (1) can be written as

$$F(x_{t+1}, x_t, t) = 0$$

For a time-invariant DAE with measured inputs and outputs, it is generally possible to reformulate (2) into

$$\mathcal{F}(x_{1,t}, x_{2,t}, x_{1,t+1}, x_{2,t+1}, u_t) = 0 \quad (2)$$

$$g(x_{1,t}, x_{2,t}, u_t) = 0 \quad (3)$$

$$h(x_{1,t}, x_{2,t}, u_t) = y_t \quad (4)$$

where we have partitioned x into two parts, and where y_t denotes our measured outputs. Assuming that (4) can be solved for $x_{1,t}$, i.e.,

$$x_{1,t} = \tilde{h}(x_{2,t}, u_t, y_t),$$

x_1 can be eliminated from (2) to give an expression in x_2 , u and y . If we further assume that x_2 can be extracted from the latter, it can be inserted into (3) to give an equation in delayed u and y ,

$$\tilde{g}(\{u_i, y_i\}_{i=t-k}^{t+1}) = 0 \quad (5)$$

for some integer $k \geq 0$. This equation gives additional relations between inputs u and outputs y , which constrain them to a manifold. These additional relations can be interpreted as inherent constraints on u , either through an implicit feedback mechanism or because of dependencies between different elements of u .

As a side remark, one could argue that a constrained u is no “real” input, since an input should be possible to choose arbitrarily. However, in many applications (for instance in many biological applications) it is difficult to conclude in advance which of the measured signals actually will act as inputs, and which will be outputs. An alternative view is not to distinguish between inputs and outputs, and just view them as different measured signals (see, e.g., (Willems, 1986)).

3. GENERAL SCHEME

If the regression data are contained in (or spread tightly around) some manifold embedded in the regression space, the system identification problem can be split into two steps. The first step amounts to finding a new coordinate representation on the manifold: the *intrinsic* or *embedded* coordinates. From the new, low-dimensional, coordinate space a nonlinear function can then be identified to predict the output. See Figure 1 for an overview.

Fig. 1. Overview of the identification steps for a system having manifold-valued regression data.

For the first step, a manifold learning technique should be used. Several techniques have been developed in the fields of face and speech recognition to reduce dimensionality and find the intrinsic coordinate frame for manifold-valued data. Each of them aims to reduce dimensionality, but with special features. The LLE (Roweis and Saul, 2000), for example, has as its goal to preserve neighbors, while Isomap (Tenenbaum *et al.*, 2000) tries to preserve the geodesic distance between points. Since local properties are often important in nonlinear system identification, we have chosen to use LLE in this paper. Another advantage with LLE is its simplicity and computational efficiency. The algorithm is briefly described in Section 3.1.

A drawback for many manifold learning algorithms is that they seldom produce an explicit mapping from the high-dimensional space down to the intrinsic coordinates, but the algorithm has to be re-run if new data is introduced. For the identification application, such a mapping is important, since the model should be able to predict the output also for new regression vectors (e.g., for validation data). One solution is to make a linear interpolation of the implicit mapping. This is the approach taken in this paper. Similar approaches have been used previously (see for example (Bengio *et al.*, 2003)).

3.1 Local Linear Embedding

The LLE algorithm classifies as a nonlinear unsupervised learning algorithm. Aiming to preserve neighborhoods, with a low complexity and with no risk of getting stuck in local minima, the LLE is an attractive choice for dimension reduction for manifold-valued data.

Algorithm (Roweis and Saul, 2000) Given data consisting of N real-valued vectors $\bar{X}_i$ of dimension D , minimize the cost function

$$\varepsilon(W) = \sum_{i=1}^N \left| \bar{X}_i - \sum_{j=1}^N W_{ij} \bar{X}_j \right|^2$$

under the constraints

$$\begin{cases} \sum_{j=1}^N W_{ij} = 1 \\ W_{ij} = 0 \text{ if } |\bar{X}_i - \bar{X}_j| > C(i) \text{ or if } i = j \end{cases}$$

Here, $C(i)$ is chosen so that only a number of K weights W_{ij} are nonzero. The number K is the only design parameter of the LLE method, along with the choice of lower dimension $d \leq D$. Now, let $\bar{Z}_i$ be of dimension d and minimize

$$\Phi(Z) = \sum_{i=1}^N \left| \bar{Z}_i - \sum_{j=1}^N W_{ij} \bar{Z}_j \right|^2$$

with respect to $Z = (\bar{Z}_1 \dots \bar{Z}_N)$, and subject to

$$\begin{cases} \sum_{i=1}^N \bar{Z}_i = 0 \\ \frac{1}{N} \sum_{i=1}^N \bar{Z}_i \bar{Z}_i^T = I \end{cases}$$

using the previously computed W_{ij} 's. Z is the new low-dimensional data set. The above algorithm can be realized as a least square problem followed by an eigenvalue problem.

A drawback with LLE is that it may run into problems while processing high-dimensional data that are not equally distributed on the manifold. More specifically, if there exist $d + 1$ clusters of points, such that for each point, all its K closest neighbors can be found within the same cluster, then the algorithm will have a solution where all points map to only $d + 1$ lower-dimensional points. Similar problems are seen also for a lower or higher degree of clustering. This kind of mapping result is of course a problem in a second identification step. Fortunately, it is fairly easy to get around it by increasing the number of neighbors K used and re-run the algorithm. However, it is not trivial to design an automated algorithm for selecting K , and in the following examples K was increased manually when needed. Note also that it is at the same

time desirable to keep K low, in order to keep the neighborhoods small.

4. MANIFOLD-VALUED REGRESSION DATA — EXAMPLES

To illustrate the two-step procedure of combining a manifold learning algorithm with a classical identification procedure for the case of manifold-valued regression data, two simple examples are given. A comparison with a direct identification approach without any dimension reduction is also given.

Example 1. Consider the system

$$\begin{aligned} x_1(t) &= 6w(t) \cos 6w(t) \\ x_2(t) &= 6w(t) \sin 6w(t) \\ y(t) &= \sqrt{x_1^2(t) + x_2^2(t)}. \end{aligned}$$

Assume that the output y is measured along with x_1 and x_2 , with some measurement error, i.e., that the measured signals are

$$\begin{aligned} x_1^m(t) &= x_1(t) + e_1(t), & e_1(t) &\sim N(0, \sigma_{e_1}^2) \\ x_2^m(t) &= x_2(t) + e_2(t), & e_2(t) &\sim N(0, \sigma_{e_2}^2) \\ y^m(t) &= y(t) + e_y(t), & e_y(t) &\sim N(0, \sigma_{e_y}^2) \end{aligned}$$

and that a set of regression data, or learning data, is generated by the system with 400 w -values equally distributed over an interval $[0.8, 3.2]$.

Fig. 2. The estimation data or training set, $\sigma_{e_1} = \sigma_{e_2} = 0.25$ and the number of data points $N = 400$.

As long as the measurement noise is not too large, the regressors, x_1^m and x_2^m , are situated around a one-dimensional manifold, a spiral. A plot of the regression vectors of the estimation data (or training set) is given in Figure 2 for the case $\sigma_{e_1} = 0.25$, $\sigma_{e_2} = 0.25$ and $\sigma_{e_y} = 0.25$, and the output is plotted against the regressors in Figure 3. By just a look at the plot of the estimation data in Figure 2, it is easy to see that any linear projection method would not perform well on this example, the reason being that no matter which hyperplane is chosen for projection,

Fig. 3. Estimation data together with the output, $\sigma_{e_1} = \sigma_{e_2} = \sigma_{e_y} = 0.25$ and the number of data points $N = 400$.

there will be estimation data with different outputs that are projected to the same point. Instead, the two-step procedure seems to be an attractive choice. Even though the dimensionality of the problem is not an issue in this particular example, the manifold-valued regression data makes it a suitable example.

Using the LLE algorithm, a set of (one-dimensional) intrinsic coordinates can be computed for the estimation data. The number of closest neighbors was set to $K = 10$. A plot showing both the intrinsic coordinates and the output is given in Figure 4.

Fig. 4. The intrinsic coordinates, z , plotted against the output y .

Having the intrinsic coordinates, a linear regression can be performed to find a map from the intrinsic coordinates to the output. For this step, linear, quadratic, and cubic polynomials were tried. The performance of the resulting mappings (from (x_1^m, x_2^m) to the predicted output $\hat{y}$) was then evaluated by computing the sum of squared prediction errors for a validation data set consisting of 400 data points, generated in the same manner as the estimation data set. For each new regression vector from the validation data set, an interpolation using the three closest estimation data points was done, in order to extend the manifold learning mapping to out-of-sample points. For comparison, a

Table 1. Results for Example 1 (low-dimensional case). The mean of the summed squared errors (based on 400×50 estimations and validations) $\frac{1}{50} \sum_{j=1}^{50} \sum_{i=1}^{400} (y^j(i) - \hat{y}^j(i))^2$, for the two step method and direct regression of different orders.

Method \ Order	1	2	3
LLE and regression	0.885	0.491	0.364
Direct regression	31.7	1.05	1.00

Table 2. Results for Example 2 (high-dimensional case). The mean of the summed squared errors (based on 400×50 estimations and validations) $\frac{1}{50} \sum_{j=1}^{50} \sum_{i=1}^{400} (y^j(i) - \hat{y}^j(i))^2$, for the two step method and direct regression of different orders.

Method \ Order	1	2
LLE and regression	36	33.8
Direct regression	280	700

linear, a quadratic, and a cubic polynomial were fitted directly to the estimation data. For both methods, the whole estimation-validation procedure was performed 50 times, and the average of the sums of squared errors was computed. Because of the clustering problem described in Section 3.1, the numbers of neighbors, K , used in the LLE algorithm had to be increased in 10 of the 50 estimation-validation procedure. The number of neighbors used never exceeded 15. The results are summarized in Table 1. To better evaluate the prediction power of the identified models, no measurement noise was added to the validation output.

As we can see, the performance of the two-step method is very good even in this low-dimensional case. However, to obtain the full benefit of the dimension reduction from the LLE algorithm, the dimension of the regression space needs to be increased. This is done in the next example.

Example 2. To exemplify the behavior for a high-dimensional case, the previous example was extended as follows. x_1^m and x_2^m from Example 1 were used to compute

$$\begin{aligned}
&(\tilde{x}_1^m, \tilde{x}_2^m, \tilde{x}_3^m, \tilde{x}_4^m, \tilde{x}_5^m, \tilde{x}_6^m, \tilde{x}_7^m, \tilde{x}_8^m) \\
&= (e^{x_1^m}, e^{x_2^m}, x_2^m e^{-x_1^m}, x_1^m e^{-x_2^m}, \\
&\quad \log |x_1^m|, \log |x_2^m|, 1/x_1^m, 1/x_2^m)
\end{aligned}$$

which were now seen as the new regressors. Using the same estimation and validation procedure (with $K = 10$, $N = 400$) as in Example 1 gave the result shown in Table 2. The number of neighbors used in the LLE algorithm had to be increased 11 times due to clustering, but it never exceeded 16.

Note that in this example the LLE algorithm reduces the dimension from eight to one compared to from two to one in the previous example. The results indicates

the usefulness of using a dimension reduction before identifying in high-dimensional cases.

5. CONCLUSIONS

A two-step method has been presented as a way to identify nonlinear functions of high dimensionality or of special character when the regressors are manifold-valued. The approach was successfully applied and illuminated the possibility to use manifold learning techniques together with system identification. This paper should be seen as a first step of using manifold learning techniques in combination with nonlinear system identification.

For the more general case when data is scattered over the regression space, a projection step would be needed to be able to benefit from a dimension reduction, as is done in (Lindgren, 2005; Li, 1991; Li, 1992; Fukumizu *et al.*, 2004). In all these papers, the projection performed is linear, and an interesting topic for further research would be to find an extension by making use of ideas from manifold learning techniques.

The possible lack of robustness for manifold learning algorithms, as discussed in the last paragraph of Section 3.1, indicates that more work is needed in this area, for instance in order to get an automated selection of the number of neighbors K used in the LLE algorithm. Another interesting question would be to evaluate which manifold learning algorithm to use in which context. This should be a topic for further research.

6. ACKNOWLEDGMENT

This work was supported by the Strategic Research Center MOVIII, funded by the Swedish Foundation for Strategic Research, SSF.

REFERENCES

- Bengio, Y., J.-F. Paiement and P. Vincent (2003). Out-of-sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering. Technical Report 1238. Département d'Informatique et Recherche Opérationnelle.
- Brun, A. (2006). Manifold learning and representation for image analysis and visualization. Linköpings Universitet, LiU-TEK-LIC-2006:16, licentiate Thesis No. 1235.
- de Ridder, D., O. Kouropteva, O. Okun, M. Pietikäinen and R. Duin (2003). *Artificial Neural Networks and Neural Information Processing — ICANN/ICONIP 2003*. Chap. Supervised locally linear embedding, pp. 333–341. Vol. 2714/2003 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg.

- Fritzson, P. (2004). *Principles of Object-Oriented Modeling and Simulation with Modelica 2.1*. Wiley.
- Fukumizu, K., F. R. Bach and M. I. Jordan (2004). Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research* **5**, 73–99.
- Hastie, T., R. Tibshirani and J. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- Kouropyteva, O., O. Okun, A. Hadid, M. Soriano, S. Marcos and M. Pietikäinen (2002). Beyond locally linear embedding algorithm. Technical Report MVG-01-2002. University of Oulu, Machine Vision Group, Information Processing Laboratory. pp 49.
- Kunkel, P. and V. Mehrmann (2006). *Differential-Algebraic Equations — Analysis and Numerical Solution*. EMS Publishing House, Zürich.
- Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**(414), 316–327.
- Li, K.-C. (1992). On principal hessian directions for data visualizations and dimension reduction: Another application of Stein’s lemma. *Journal of the American Statistical Association* **87**(420), 1025–1039.
- Lindgren, D. (2005). Projection Techniques for Classification and Identification. PhD thesis. Linköpings universitet. Dissertation no 915.
- Ljung, L. (1999). *System Identification - Theory For the User*. 2 ed. PTR Prentice Hall. Upper Saddle River, N.J.
- Roweis, S. T. and L. K. Saul (2000). Nonlinear dimensionality reduction by local linear embedding. *Science* **290**(5500), 2323–2326.
- Schölkopf, B., A. Smola and K.-R. Müller (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comp.* **10**(5), 1299–1319.
- Tenenbaum, J. B., V. de Silva and J. C. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290**(5500), 2319–2323.
- Tiller, M. (2001). *Introduction to Physical Modeling with Modelica*. Kluwer Academic Publishers.
- Willems, J. C. (1986). From time series to linear system. part i: Finite dimensional linear time invariant systems. *Automatica* **22**(5), 561–580.
- Zhao, Q., D. Zhang and H. Lu (2005). Supervised LLE in ICA space for facial expression recognition. In: *Neural Networks and Brain, 2005. ICNN&B ’05. International Conference on*. Vol. 3. pp. 1970–1975. ISBN: 0-7803-9422-4.

Avdelning, Institution Division, Department Division of Automatic Control Department of Electrical Engineering		Datum Date 2007-06-13
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-2795
Titel Title Using Manifold Learning for Nonlinear System Identification		
Författare Author Henrik Ohlsson, Jacob Roll, Torkel Glad, Lennart Ljung		
Sammanfattning Abstract A high-dimensional regression space usually causes problems in nonlinear system identification. However, if the regression data are contained in (or spread tightly around) some manifold, the dimensionality can be reduced. This paper presents a use of dimension reduction techniques to compose a two-step identification scheme suitable for high-dimensional identification problems with manifold-valued regression data. Illustrating examples are also given.		
Nyckelord Keywords Nonlinear system identification; Dimension reduction techniques; Manifold learning		