Technical report from Automatic Control at Linköpings universitet

Sliced Inverse Regression for the Identification of Dynamical Systems

Christian Lyzell, Martin Enqvist

Division of Automatic Control

E-mail: lyzell@isy.liu.se, maren@isy.liu.se

14th November 2011

Report no.: LiTH-ISY-R-3031
Submitted to 16th IFAC Symposium on System Identification

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

The estimation of nonlinear functions can be challenging when the number of independent variables is high. This difficulty may, in certain cases, be reduced by first projecting the independent variables on a lower dimensional subspace before estimating the nonlinearity. In this paper, a statistical nonparametric dimension reduction method called *sliced inverse regression* is presented and a consistency analysis for dynamically dependent variables is given. The straightforward system identification application is the estimation of the number of linear subsystems in a Wiener class system and their corresponding impulse response.

Keywords: System identification; Dimension reduction; Inverse regression.

Sliced Inverse Regression for the Identification of Dynamical Systems

Christian Lyzell, Martin Enqvist

2011-11-14

Abstract

The estimation of nonlinear functions can be challenging when the number of independent variables is high. This difficulty may, in certain cases, be reduced by first projecting the independent variables on a lower dimensional subspace before estimating the nonlinearity. In this paper, a statistical nonparametric dimension reduction method called *sliced inverse regression* is presented and a consistency analysis for dynamically dependent variables is given. The straightforward system identification application is the estimation of the number of linear subsystems in a Wiener class system and their corresponding impulse response.

1 Introduction

In this paper, we will consider the identification of dynamical systems in the form

$$y(t) = f(B^{\mathrm{T}}\varphi(t), e(t)), \quad B \in \mathbb{R}^{n_b \times d}, \tag{1}$$

where f is a time-independent nonlinear function and

$$\varphi(t) \triangleq \begin{pmatrix} u(t) & u(t-1) & \cdots & u(t-n_b+1) \end{pmatrix}^{\mathrm{T}} \tag{2}$$

is a vector consisting of time shifted inputs. In particular, the goal is to estimate a basis for the column space of B and its dimension d from data, without knowing or estimating the nonlinearity f. An important benefit from such an approach is that, if the dimension of the projected space is significantly lower than the number of regressors, then the ensuing estimation of the nonlinearity will be computationally less demanding and yield more accurate estimates. This is illustrated in Figure 1, where the correct projection reveals the nature of the nonlinearity.

Historically, the inference problem of estimating the matrix B in (1), given data, has been done using forward regression, that is, finding an suitable projection via the minimization of some cost function. Typically, this approach involves parameterizing the nonlinearity in some appropriate basis [Friedman and Stuetzle, 1981] and the corresponding optimization problems often becomes nonconvex with several local minima. Another difficulty with this approach is to determine the dimension of the projection and it is therefore often assumed to be known in advance [Friedman, 1987].

These difficulties can be circumvented if one restricts the distribution of the regressors. In the one-dimensional case, it is well known that if $\varphi(t)$ is elliptically

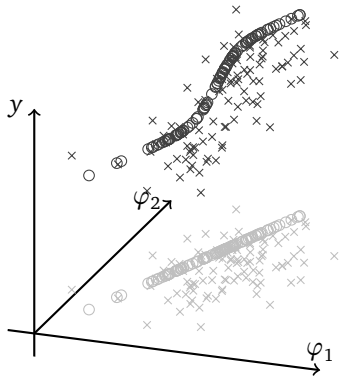

Figure 1: Data has been collected from a nonlinear dynamical system in the form (2) illustrated by crosses. With an appropriate projection, projected data depicted with circles, the nature of the nonlinearity becomes apparent.

distributed, the linear least-squares estimate of B is consistent [Bussgang, 1952]. In this paper, we are going to consider a related method called *sliced inverse regression* (SIR) [Li, 1991], which has had a considerable influence in the field of dimension reduction in the statistical community [see, for example, Bura and Cook, 2001, Li and Wang, 2007, Bura and Yang, 2011].

From a system identification perspective, a model structure related to (1) is the Wiener class of systems [Schetzen, 1980, Boyd and Chua, 1985]. This system structure consists of a finite number of parallel linear subsystems followed by a static multiple-input single-output nonlinearity. In particular, the system (1) corresponds to the Wiener class of systems where the linear subsystems have a *finite impulse response* (FIR).

The purpose of this paper is to introduce SIR for the estimation of systems in the form (1) and to analyze the consistency of the corresponding estimator when the regressors (2) are dynamically dependent. To this end, some assumptions on the involved signals are needed.

Definition 1. A stochastic processes $(x(t))_{t=-\infty}^{\infty}$ is strictly stationary if the joint probability density function of any set of variables $\{x(t+\tau), \tau \in D \subset \mathbb{Z}\}$ is independent of t.

Assumption 1. The elements appearing in (1) have the following properties:

(i) The signal $u(t)$ is a strictly stationary stochastic processes with finite variance.

(ii) The signal $e(t)$ is a strictly stationary stochastic process, independent of $u(t)$, with zero mean and finite variance.

(iii) The nonlinearity f is time independent and such that the joint probability density function of $y(t)$ and $\varphi(t)$ defined in (2) is locally integrable, that is, it is integrable on every bounded measurable set [see, for instance, Folland, 1999].

It is worth noting that item (iii) in Assumption 1 is fulfilled, for example, when the random variables $u(t)$ and $e(t)$ are continuous and the nonlinearity f is bounded on every compact set with only a finite number of jumping discontinuities.

The paper is outlined as follows: In Section 2, a summary of the basic ideas and limitations of SIR is given. In Section 3, an algorithm for calculating the SIR estimate of the projection matrix is given. Furthermore, the consistency of the corresponding estimator is analyzed under a simple *mixing assumption* on the input. The performance of the estimator is then evaluated in Section 4 on some artificial data and the paper is concluded in Section 5.

2 Inverse regression

The field of inverse regression commenced with the seminal paper by Li [1991] in which a simple method for estimating the projection in (1) is presented. The procedure is named *sliced inverse regression*, SIR for short, and operates without the need of knowing or estimating the nonlinearity. This is achieved, similarly to Bussgang [1952], by limiting the distribution of the regressors. The assumption used in Li [1991] is referred to as the *linear design condition*.

Assumption 2. For any $a \in \mathbb{R}^{n_b}$, there exist $c_0 \in \mathbb{R}$ and $c \in \mathbb{R}^d$ such that

$$\mathrm{E}(a^{\mathrm{T}}\varphi(t) \,|\, B^{\mathrm{T}}\varphi(t)) = c_0 + c^{\mathrm{T}}B^{\mathrm{T}}\varphi(t). \tag{3}$$

Before discussing the implications of the above condition on the distribution of $\varphi(t)$, let us turn our attention to the fundamental result on which SIR is based upon.

Theorem 1. *Let the system be given by* (1)*. If Assumption 1 and 2 are fulfilled, it holds that*

$$\mathrm{E}(\varphi(t) \,|\, y(t)) - \mathrm{E}(\varphi(t)) \in \mathcal{R}\left(\mathrm{Cov}(\varphi(t))B\right), \tag{4}$$

where the operator $\mathcal{R}$ returns the linear subspace spanned by the columns of its argument.

Proof. The proof given in Li [1991] is valid with a change in the notation or see Appendix A for a different approach. □

According to Theorem 1, the only information needed to estimate the projection in (1) is the conditional expectation appearing in (4). This entity is easily estimated from data and a simple nonparametric method based on Li [1991] is presented below. The result of Theorem 1 is the foundation of more advanced methods [see, for instance, Bura and Cook, 2001, Bura, 2003] but has one major drawback. If the nonlinearity in (1) is even about the origin, then the conditional expectation (4) will be zero and no information regarding the projection will be present in the data. To remedy this limitation, several methods involving second order moments have been proposed in the literature, see, for example, Cook and Weisberg [1991], Li and Wang [2007], Bura and Yang [2011]. The evaluation of some of the above methods on systems from the Wiener class can be found in Lyzell and Enqvist [2011].

The linear design condition, Assumption 2, restricts the distributions of the regressors that are applicable. It can be shown [Cook and Weisberg, 1991] that (3)

is fulfilled, for instance, when $\varphi(t)$ is elliptically distributed, that is, when the level curves of the probability density function are ellipsoids [see, for example, Eaton, 1986]. Furthermore, Hall and Li [1993] proved that if B is a random matrix with columns from the uniform distribution on the n_b-dimensional unit sphere, the linear design condition holds as $n_b \to \infty$, see also Li and Wang [2007]. In addition, Cook and Nachtsheim [1994] presents a method for reweighting the regressors to follow a elliptical distribution [see also, Enqvist, 2007].

3 Statistical Inference

In this section, an algorithm for estimating a basis for $\mathcal{R}(B)$ is formulated and a consistency analysis for dynamically dependent regressors is provided. In the following, consider that a dataset $(\varphi(t), y(t))_{t=0}^{N-1}$ is given in accordance to (1) and (2).

3.1 Sliced Inverse Regression

A common method for improving the numerical accuracy is to *standardize* the regressors by removing the mean and apply scaling to achieve an identity covariance matrix,

$$\zeta(t) \triangleq \operatorname{Cov}(\varphi(t))^{-1/2}\big(\varphi(t) - \mathrm{E}(\varphi(t))\big), \tag{5}$$

and (4) can now be written as

$$\mathrm{E}(\zeta(t) \,|\, y(t)) \in \mathcal{R}\left(\operatorname{Cov}(\varphi(t))^{1/2} B\right). \tag{6}$$

Thus, to estimate a basis for $\mathcal{R}(B)$ one only need to estimate the conditional expectation in (6). This can be done in a number of different ways [see, for instance, Bura and Cook, 2001, Bura, 2003]. The original work of Li [1991] uses a rather simple estimator based on a discretized version of the output. Let

$$\{\mathcal{Y}_j,\ j = 1, 2, \ldots, J\}, \tag{7}$$

be a partition of the output space, where the number of *slices* J is chosen by the user. Then the SIR estimator is given by

$$\mathrm{E}(\zeta(t) \,|\, y(t) \in \mathcal{Y}_j) \approx \sum_{t=0}^{N-1} \frac{I_{\mathcal{Y}_j}(y(t))}{\sum_{t=1}^{N} I_{\mathcal{Y}_j}(y(t))} \zeta(t), \tag{8}$$

where $I_A(x)$ is the indicator function which returns one if x is a member of A and zero otherwise.

Algorithm 1 (SIR)**.** Given a dataset, an estimate $\hat{T}_B$ of a basis for $\mathcal{R}(B)$ is returned.

1) Standardize $\zeta(t) \triangleq \hat{\Sigma}^{-1/2}(\varphi(t) - \hat{\mu})$, where $\hat{\mu}$ and $\hat{\Sigma}$ is the sample mean and covariance, respectively.

2) Determine $M \in \mathbb{R}^{n_b \times J}$ with columns according to (8) for some choice of partition (7).

3) Find the d leading singular values of M and the corresponding left singular vectors U_d.

4) Then $\hat{T}_B = \hat{\Sigma}^{-1/2} U_d$ is an estimate of a basis for $\mathcal{R}(B)$, compare with (6).

There are several alternatives for constructing a partition (7) of the output data. In the simulation studies that follow, we are going to use a simple data dependent partitioning scheme. Assume that the output has been sorted in ascending order, that is, $y(s) \leq y(t)$ for all $s \leq t$, and that $\zeta(t)$ has been rearranged accordingly. Let k be the largest integer such that $kJ \leq N$. Then

$$\mathcal{Y}_j \triangleq \{y(t),\ (j-1)k \leq t < jk\}, \quad j = 1, 2, \ldots, J-1, \tag{9}$$

and $\mathcal{Y}_J \triangleq \{y(t),\ (J-1)k \leq t \leq N-1\}$ is a partition of the output data. With this partitioning scheme, the SIR estimator (8) can be written as

$$\mathrm{E}(\zeta(t) \,|\, y(t) \in \mathcal{Y}_j) \approx \frac{1}{|\mathcal{I}_j|} \sum_{i \in \mathcal{I}_j} \zeta(i), \tag{10}$$

where $\mathcal{I}_j$ is the set of indices belonging to slice j and $|\mathcal{I}_j|$ the number of such elements, and be used in Step 2 of Algorithm 1.

3.2 Consistency

In this section, the consistency of the SIR estimator (8) for a certain class of stationary stochastic processes will be analyzed. The first assumption concerns the mixing properties of the signals involved. First, a definition is given [see, for example, Chung, 1974].

Definition 2. A stochastic process $(x(t))_{t=-\infty}^{\infty}$ is called *m-dependent* if and only if there exists an integer m such that for every n, $\{x(k),\ 1 \leq k \leq n\}$ and $\{x(n+j),\ j \geq m+1\}$ are independent.

This is a slightly stricter assumption than the one given in Chung [1974, p. 214]. From a dynamical systems point of view, an m-dependent process is generally generated by filtering a white noise sequence through an FIR system.

Assumption 3. The processes $(u(t))_{t=-\infty}^{\infty}$ and $(e(t))_{t=-\infty}^{\infty}$ are m_u-dependent and m_e-dependent, respectively.

An immediate consequence of Assumption 1 and 3 is that the processes $(\varphi(t))_{t=-\infty}^{\infty}$ and $(y(t))_{t=-\infty}^{\infty}$ are m_φ-dependent and m_y-dependent where $m_\varphi \triangleq m_u + n_b - 1$ and $m_y \triangleq \max(m_\varphi, m_e)$, respectively. The first result treats the convergence of the estimator when the partition (7) is kept fixed.

Theorem 2. *Let the system be given by* (1) *and let* $\mathcal{Y}_j$ *be an open interval in the partition* (7)*. Under Assumption 1 and 3, it holds that*

$$\lim_{N\to\infty} \sum_{t=1}^{N} \frac{I_{\mathcal{Y}_j}(y(t))}{\sum_{t=0}^{N} I_{\mathcal{Y}_j}(y(t))} \varphi(t) = \frac{\mathrm{E}(I_{\mathcal{Y}_j}(y(t))\varphi(t))}{\mathrm{E}(I_{\mathcal{Y}_j}(y(t)))} \ \text{a.s.} \tag{11}$$

Proof. See Appendix A. □

Now, we need to investigate what happens when the partitioning (7) is refined.

Theorem 3. *Let the system be given by* (1) *and let* $\mathcal{Y}_j$ *be a symmetric open interval around a point* $\bar{y}$ *with length* $\delta > 0$ *in the range of* $y(t)$*. Under Assumption 1, it holds that*

$$\lim_{\delta \to 0} \frac{\mathrm{E}(I_{\mathcal{Y}_j}(y(t))\varphi(t))}{\mathrm{E}(I_{\mathcal{Y}_j}(y(t)))} = \mathrm{E}(\varphi(t) \,|\, y(t) = \bar{y}). \tag{12}$$

Proof. See Appendix A. □

The results of Theorem 2 and 11 implies that SIR is a consistent estimator of the conditional expectation $\mathrm{E}(\varphi(t) \,|\, y(t))$, and thus Algorithm 1 provides a procedure to estimate $\mathcal{R}(B)$ consistently.

4 Simulations

In this section, the performance of the SIR method will be evaluated on artificial data generated from some realizations of (1). The first simulation study concerns the case when $d = 1$, where the effects of different input signals and nonlinearities are analyzed. Furthermore, the method will be compared to the least-squares estimate, which is consistent when the input is, for instance, elliptically distributed [Nuttall, 1958]. In the second simulation study, the dimension d is considered to be unknown and the ability of SIR to indicate the number of linear subsystems correctly is evaluated.

As a measure of quality of an estimate the *angle between subspaces* will be used [see, for example, Li and Wang, 2007]. Let A and B be matrices. The angle between $\mathcal{R}(A)$ and $\mathcal{R}(B)$ is defined by

$$\operatorname{angle}(A, B) \triangleq \arcsin\left(\|\Pi_A - \Pi_B\|_2\right), \tag{13}$$

where Π_A denotes the orthogonal projection on $\mathcal{R}(A)$. This generalizes the angle between vectors and can be determined, for instance, in MATLAB with the command `subspace`.

The quality of the data will be measured in terms of the *signal to noise ratio* defined here by

$$\mathrm{SNR} \triangleq \frac{\operatorname{Var}(y_{\mathrm{nf}}(t))}{\operatorname{Var}(y(t) - y_{\mathrm{nf}}(t))}, \tag{14}$$

where $y_{\mathrm{nf}}(t)$ denotes the noise-free part of $y(t)$.

In the following, let $\mathcal{U}(a, b)$ denote the uniform distribution on the interval (a, b) and let $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution with mean μ and variance σ^2.

In the two following examples, we will consider systems in the form

$$y(t) = f(B^{\mathrm{T}}\varphi(t)) + v(t), \quad v(t) \sim \mathcal{N}(0, \sigma_v^2), \tag{15a}$$

where $B \in \mathbb{R}^{13}$ has elements given by

$$B_i = \begin{cases} 3/2^i, & i = 4, 5, \ldots, 10, \\ 0, & \text{otherwise.} \end{cases} \tag{15b}$$

The regressors $\varphi(t)$ will be constructed from, for each example, different input signals. The additive noise $v(t)$ is generated independently of $u(t)$ with variance σ_v^2 chosen to yield a specific SNR (14).

The setup of the Monte Carlo simulations is as follows: for each level of SNR, 100 datasets consisting of 1,000 data points for identification are collected, where a new input and noise sequence are sampled at each instance. The number of regressors and slices used in the experiments are fixed to $n_b = 13$ and $J = \lfloor \sqrt{N} \rfloor = 31$, respectively.

Example 1: Elliptically distributed input signal

Let the input to the system (15) be given by

$$u(t) = G_u(q)\eta_u(t), \quad \eta_u(t) \sim \mathcal{N}(0, 0.15), \tag{16}$$

where the linear filter has the impulse response

$$g_{u,i} = \begin{cases} 2/3^i, & i = 0, 1, \ldots, 5, \\ 0, & \text{otherwise.} \end{cases}$$

Then $\varphi(t)$, defined in (2), is elliptically distributed and both LS and SIR are expected to work well. To show the effects for different nonlinearity, two different noninvertible functions are considered, namely $f(x) = \text{sign}(x)$ and $f(x) = x\cos(x)$, respectively.

Figure 2 shows the resulting angle (13) between the true impulse response and the corresponding estimates for different levels of SNR. For $f(x) = \text{sign}(x)$ (to the left), we see that both methods perform equally well and it is difficult to tell one estimate from the other. For the more oscillating nonlinearity $f(x) = x\cos(x)$ (to the right), we see that SIR significantly outperforms LS. The LS estimator is still consistent, but the convergence is slow due to difficulty of averaging out the nonlinearity. The estimate given by SIR is not perfect, especially for low SNR, and a larger data set is needed, but the convergence is faster than that of LS, at least for this example.

Figure 3 shows the true impulse response and the LS and SIR estimates with 3 standard deviation errors for SNR $= 20$ dB respectively. The results here reflect the ones given in Figure 2 with $f(x) = \cos(x)$ being the more difficult nonlinearity.

Example 2: Nonelliptically distributed input

Let the input to the system (15) now be given by

$$u(t) = 2\sin(2\pi\eta_u(t)), \quad \eta_u(t) \sim \mathcal{U}(0, 1). \tag{17}$$

Then $\varphi(t)$, defined in (2), is nonelliptically distributed and there is no guarantee that either LS or SIR will work well. The Monte Carlo results for the nonlinearity $f(x) = \text{sign}(x)$ are given in Figure 4 where the left plot shows the angle between the true impulse response and the estimates (13) and the one on the right shows the estimated impulse responses for 20 dB SNR, respectively. The LS and SIR estimators seems have similar accuracy and are clearly biased in this case.

Now, let us consider a more interesting example.

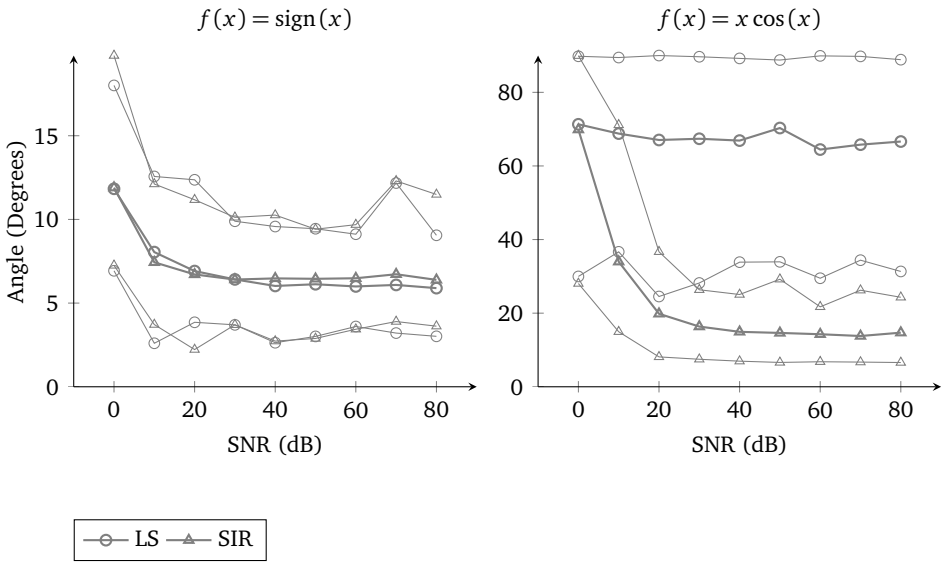

Figure 2: (Example 1). The angle (13) between the true impulse response and the estimates for different levels of the SNR (14). The thick lines shows the mean value, while the thin lines shows the maximum and the minimum values attained, respectively.

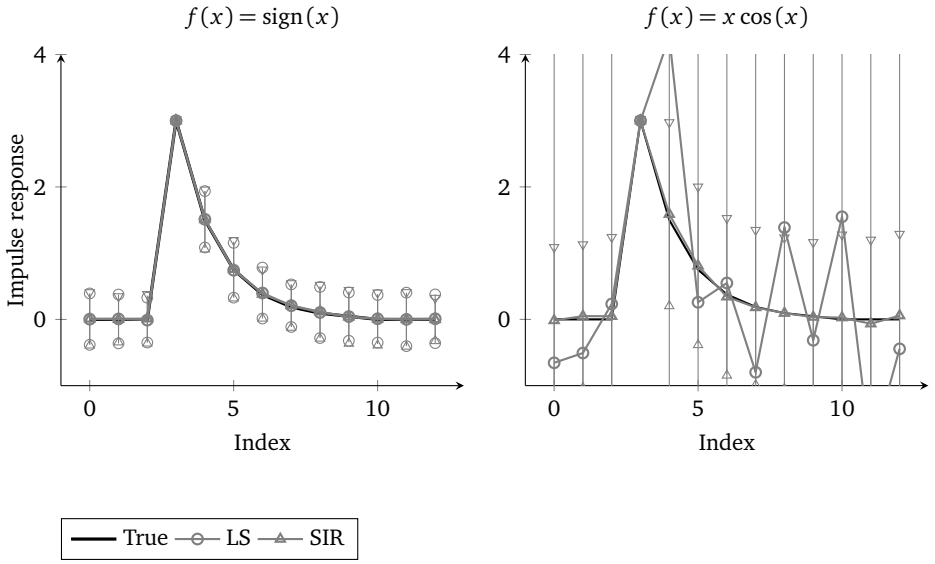

Figure 3: (Example 1). The mean value of the impulse response estimates for 20 dBSNR with 3 standard deviation errors when all estimates have been scaled to coincide for the first nonzero coefficient.

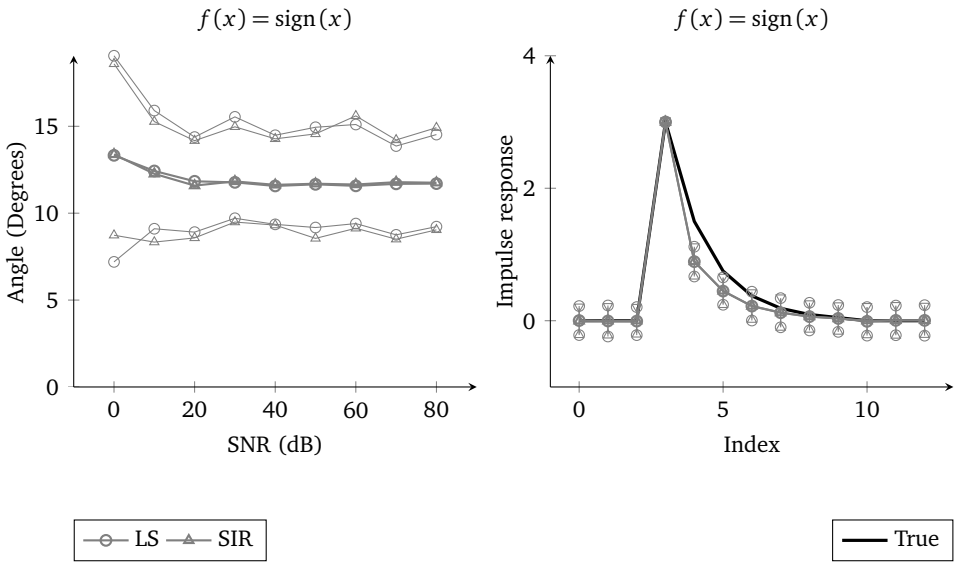

Figure 4: (Example 2). The plot to the left shows the angle (13) between the true impulse response and the estimates for different SNR, respectively. The plot to the right shows the true impulse response together with the estimates for $\text{SNR} = 20\,\text{dB}$ with 3 standard deviation errors, respectively.

Example 3: Output Error Linear System

Let the system be given by

$$y(t) = f\left(G(q)u(t) + w(t)\right) + v(t), \quad v(t) \sim \mathcal{N}(0, \sigma^2) \tag{18}$$

where the measurement noise $v(t)$ and the process noise

$$w(t) = \frac{1}{1 + 0.3q^{-1}}\eta_w(t), \quad \eta_w(t) \sim \mathcal{N}(0, \sigma^2), \tag{19}$$

are drawn independently. The input is generated as

$$u(t) = \frac{1}{1 + 0.6q^{-1}}\eta_u(t), \quad \eta_u(t) \sim \mathcal{N}(0, 1), \tag{20}$$

and $\varphi(t)$ is therefore elliptically distributed. One difference from the previous examples is that the input $u(t)$ is no longer m-dependent which is one of the requirements of Theorem 2. Furthermore, the linear system $G(q)$ does not have a finite impulse response and we will compare the estimates of LS and SIR with the truncated impulse response. The Monte Carlo results are shown in Figure 5. The left plot shows that SIR is quite accurate for small values of σ but that the performance degrades as σ increases. The LS estimator on the other hand performs quite badly for small σ but eventually is more accurate than SIR. This is a bit confusing but somehow the noise helps the LS estimator in this particular case. If the pole in the input filter (20) is mirrored to be 0.6 instead of -0.6, the LS estimator performance is more similar to that of SIR but still not as accurate. The plot to the right in Figure 5 shows the estimates of the truncated impulse response for the

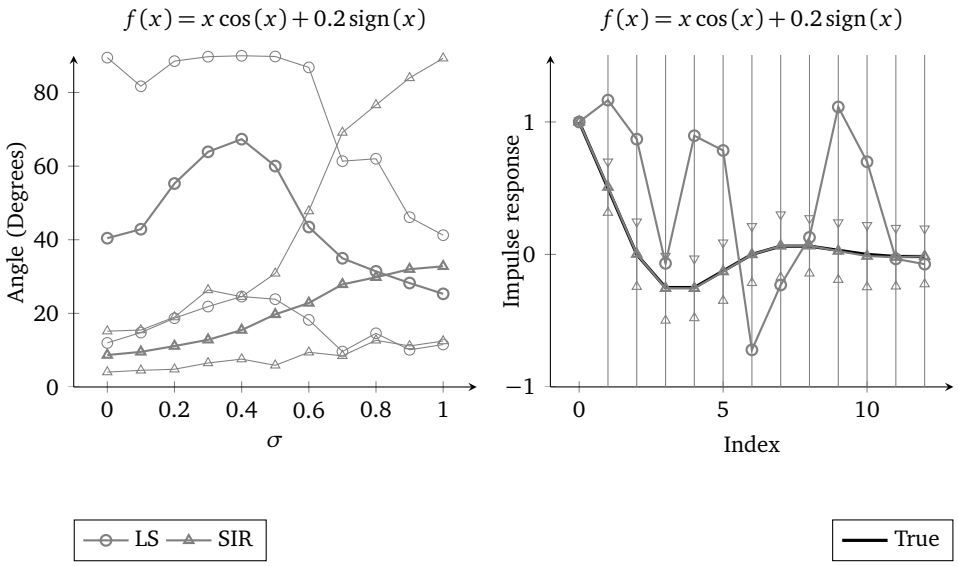

Figure 5: (Example 3). The plot to the left shows the angle (13) between the true impulse response and the estimates for different σ. The plot to the right shows the true impulse response together with the estimates for $\sigma = 0.2$, SNR ≈ 12.5 dB, with 3 standard deviation errors, respectively.

case $\sigma = 0.2$. The SIR estimate is close to the true response which shows that the usefulness of the SIR method is not restricted to the m-dependence case.

Example 4

Consider a system in the form

$$y(t) = f\left(G_1(q)u(t) + w_1(t), G_2(q)u(t) + w_2(t)\right) + v(t), \tag{21}$$

with the linear subsystems

$$G_1(z) = \frac{1}{z - 0.75}, \quad G_2(z) \triangleq \frac{z^2 - 0.5z}{z^2 - z + 0.5}, \tag{22}$$

and the nonlinearity

$$f(x_1, x_2) = \tanh(x_1)\,\mathrm{sat}(x_2) + \mathrm{sign}(x_2). \tag{23}$$

The input signal is generated according to

$$u(t) = \frac{1}{1 + 0.6q^{-1}}\eta_u(t), \quad \eta_u(t) \sim \mathcal{N}(0, 1), \tag{24}$$

and the measurement noise $v(t) \sim \mathcal{N}(0, 0.1)$. On each channel, process noise is added according to (21) which is given by

$$w_1(t) = \frac{1}{z + 0.25}\eta_{w,1}(t), \qquad \eta_{w,1}(t) \sim \mathcal{N}(0, 0.1), \tag{25}$$

$$w_2(t) = \frac{1}{z - 0.3}\eta_{w,2}(t), \qquad \eta_{w,2}(t) \sim \mathcal{N}(0, 0.1). \tag{26}$$

In the Monte Carlo simulation, 100 datasets consisting of 10,000 data points for identification are collected with new realizations of the input and noise sequence at each instance. The singular values of the matrix M in Algorithm 1 are presented in Figure 6. Here one sees that there is a large gap in the singular values with

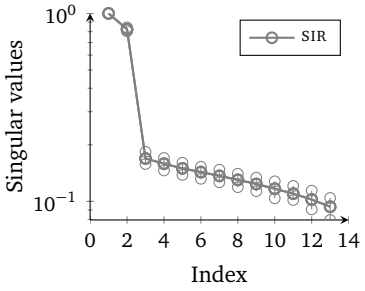

Figure 6: (Example 4). The singular values with the bars indicating the minimum and maximum values attained for the Monte Carlo simulation.

two values that significantly larger then the rest, which hints that $d = 2$. With $d = 2$ the mean value of the angle (13) between the truncated impulse response corresponding to (22) and the SIR estimate is 2.27 degrees with a maximum value of 3.25.

The example above implies that the SIR method may be of help as an initial estimator of the Wiener class of systems when there are more than one linear subsystem [see also Lyzell and Enqvist, 2011].

5 Conclusions

In this paper, the well known statistical method SIR is introduced for the purpose of identifying dynamical systems. Furthermore, the corresponding estimator has been shown to be consistent for m-dependent inputs. The identification results on artificial data looks promising and the estimates seems to be at least as good as the estimates based upon Bussgang [1952].

Future research includes improving the consistency analysis for a more general class of inputs and investigating if the ideas of inverse regression may be conferred with subspace identification [see, for example, Westwick and Verhaegen, 1996].

References

S. Boyd and L. O. Chua. Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Curcuits and Systems*, 32 (11), 1985.

L. Breiman. *Probability*. Addison-Wesley, 1968.

E. Bura. Using linear smoothers to assess the structural dimension of regressions. *Statistica Sinica*, 13, 2003.

E. Bura and R. D. Cook. Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society: Series B*, 63(2):393–410, 2001.

E. Bura and J. Yang. Dimension estimation in sufficient dimension reduction: A unifying approach. *Journal of Multivariate Analysis*, 102(1):130–142, 2011.

J. J. Bussgang. Crosscorrelation functions of amplitude-distorted gaussian signals. Technical Report 216, MIT Research Laboratory of Electronics, Cambridge, Massachusetts, 1952.

K. L. Chung. *A Course in Probability Theory*. Academic Press, 2nd edition, 1974.

R. D. Cook and C. J. Nachtsheim. Reweighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association*, 89(426): 592–599, 1994.

R. D. Cook and S. Weisberg. Comments on "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association*, 86(414):328–332, 1991.

M. L. Eaton. A characterization of spherical distributions. *Journal of Multivariate Analysis*, 20(2):272–276, 1986.

M. Enqvist. A weighting method for approximate nonlinear system identification. In *Proceedings of the 46th IEEE Conference on Decision and Control*, pages 5104–5109, New Orleans, USA, December 2007.

G. W. Folland. *Real Analysis*. John Wiley & Sons, Ltd, 2nd edition, 1999.

J. H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82(397), 1987.

J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981.

P. Hall and K. C. Li. On almost linearity of low-dimensional projections from high-dimensional data. *The Annals of Statistics*, 21(2):867–889, 1993.

B. Li and S. Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.

K. C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

C. Lyzell and M. Enqvist. Inverse regression for the Wiener class of systems. Technical Report LiTH-ISY-R-3032, Department of Electrical Engineering, Linköping University, 2011.

A. H. Nuttall. Theory and application of the separable class of random processes. Technical Report 343, MIT Research Laboratory of Electronics, Cambridge, Massachusetts, 1958.

M. Schetzen. *The Volterra and Wiener Theories of Nonlinear Systems*. Wiley-Interscience, 1980.

D. Westwick and M. Verhaegen. Identifying MIMO Wiener systems using subspace model identification methods. *Signal Processing*, 52(2):235 – 258, 1996.

D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

A Theory

In this section, the proofs of the theorems present in the paper are given. Certain knowledge of probability theory is needed, which may be found in, for instance, Breiman [1968] or Chung [1974]. An extensive treatment of conditional expectations is given in Williams [1991].

Lemma 1. *Let X and Y be random variables with $\mathrm{E}(X^2) < \infty$. Then $h(Y) \triangleq \mathrm{E}(X \mid Y)$ is the random variable that minimizes $\mathrm{E}(X - h(Y))^2$.*

Proof. See, for example, Williams [1991, p. 85]. □

Theorem 1. The proof given here is different from the original presented in Li [1991], where the statement is validated via a proof by contradiction. Without loss of generality, assume that $\mathrm{E}(\varphi(t)) = 0$. The tower property of conditional expectations [Williams, 1991, p. 88] yields

$$\mathrm{E}(\varphi(t) \mid y(t)) = \mathrm{E}(\mathrm{E}(\varphi(t) \mid B^T\varphi(t), y(t)) \mid y(t)) = \mathrm{E}(\mathrm{E}(\varphi(t) \mid B^T\varphi(t)) \mid y(t)), \tag{27}$$

where the last equality follows from (1) and the fact that $e(t)$ is independent of $\varphi(t)$ according to Assumption 1. Lemma 1 together with Assumption 2, keeping in mind that $\mathrm{E}(\varphi(t)) = 0$, implies that $\mathrm{E}(\varphi(t) \mid B^T\varphi(t)) = \widehat{C}B^{\mathrm{T}}\varphi(t)$, where

$$\begin{aligned}\widehat{C} &= \operatorname*{arg\,min}_{C \in \mathbb{R}^{n_b \times d}} \mathrm{E}\left((\varphi(t) - CB^{\mathrm{T}}\varphi(t))^{\mathrm{T}}(\varphi(t) - CB^{\mathrm{T}}\varphi(t))\right) \\ &= \mathrm{Cov}(\varphi(t))B(B^{\mathrm{T}}\,\mathrm{Cov}(\varphi(t))B)^{-1}\end{aligned}$$

Inserting the above into (27) yields

$$\mathrm{E}(\varphi(t) \mid y(t)) = \mathrm{Cov}(\varphi(t))B\Lambda,$$

where

$$\Lambda \triangleq (B^{\mathrm{T}}\,\mathrm{Cov}(\varphi(t))B)^{-1}B^{\mathrm{T}}\,\mathrm{E}(\varphi(t) \mid y(t)).$$

This shows that $\mathrm{E}(\varphi(t) \mid y(t))$ lies in the column space of $\mathrm{Cov}(\varphi(t))B$, which completes the proof of the theorem. □

Lemma 2. *Let $\{X_n,\ n \geq 0\}$ be a strictly stationary sequence with $\mathrm{Var}(X_n) = \sigma^2 < \infty$. Then it holds that $X_n/n \to 0$ a.s. as $n \to \infty$.*

Proof. Let $\varepsilon > 0$ be given. By the subaddativity of P and the Chebyshev's inequality, it holds that

$$\mathrm{P}\Big(\bigcup_{n=m}^{\infty}\{|X_n| > n\varepsilon\}\Big) \leq \frac{\sigma^2}{\varepsilon^2}\sum_{n=m}^{\infty}\frac{1}{n^2} \to 0,$$

as $m \to \infty$ and the statement of the lemma follows from Theorem 4.2.2 in Chung [1974]. □

Theorem 2. Assumption 1 and 3 imply that the process $(y(t))_{t=-\infty}^{\infty}$ is m_y-dependent as discussed in Section 3.2 above. Let $q, r \in \mathbb{Z}$ be such that $N = qm_y + r$ with $q > 1$ and $0 \leq r < m_y$. Then it holds that

$$\frac{1}{N}\sum_{t=1}^{N} I_{\mathcal{Y}_j}(y(t)) = \frac{q}{N}\sum_{i=1}^{m_y}\frac{1}{q}\sum_{j=0}^{q-1} I_{\mathcal{Y}_j}(y(i+jm_y)) + S_r, \tag{28}$$

where $S_r \triangleq \frac{1}{N}\sum_{t=qm_y+1}^{N} I_{\mathcal{Y}_j}(y(t))$ is a sum over at most $m_y - 1$ random variables. From the definition of q and r, it is clear that $q/N \to 1/m_y$ as $N \to \infty$ and that

$$S_r = \frac{1}{N}\sum_{t=qm_y+1}^{N} I_{\mathcal{Y}_j}(y(t)) \leq \frac{r}{N} \to 0, \text{ as } N \to \infty.$$

Furthermore, according to Assumption 3, the random variables in the second sum of (28) are independent identically distributed and the strong law of large numbers [see, for instance, Chung, 1974, p. 126] yields

$$\lim_{N\to\infty}\frac{1}{q}\sum_{j=0}^{q-1} I_{\mathcal{Y}_j}(y(i+jm_y)) = \mathrm{E}(I_{\mathcal{Y}_j}(y(t))),$$

almost surely. Thus, it holds that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N} I_{\mathcal{Y}_j}(y(t)) = \mathrm{E}(I_{\mathcal{Y}_j}(y(t))), \tag{29}$$

almost surely. In a similar manner, it holds that

$$\frac{1}{N}\sum_{t=1}^{N} I_{\mathcal{Y}_j}(y(t))\varphi(t) = \frac{q}{N}\sum_{i=1}^{m_y}\frac{1}{q}\sum_{j=0}^{q-1} I_{\mathcal{Y}_j}(y(i+jm_y))\varphi(i+jm_y) + \tilde{S}_r,$$

where $\tilde{S}_r \triangleq \frac{1}{N}\sum_{t=qm_y+1}^{N} I_{\mathcal{Y}_j}(y(t))\varphi(t)$ is a sum over a finite number of stochastic vectors. By applying Lemma 2 componentwise, it follows that $\tilde{S}_r \to 0$ a.s. as $N \to \infty$. Arguing as above, it holds that

$$\frac{1}{N}\sum_{t=1}^{N} I_{\mathcal{Y}_j}(y(t))\varphi(t) \to \mathrm{E}(I_{\mathcal{Y}_j}(y(t))\varphi(t)), \tag{30}$$

as $N \to \infty$. Finally, combining (29) and (30), using the fact that the quotient of two almost surely convergent sequences converges to the quotient of their respective limits almost surely, yields the desired result (11). □

Theorem 3. It follows from Assumption 1 and the Lebesgue differentiation theorem [Folland, 1999, p. 98] that

$$\lim_{\delta\to 0}\frac{1}{|A|}\mathrm{E}(I_{\mathcal{Y}_j}(y(t))) = \lim_{\delta\to 0}\frac{1}{|A|}\int I_{\mathcal{Y}_j}(y)p_{y(t)}(y)\,dy = p_{y(t)}(\bar{y}), \tag{31}$$

where $|A| \triangleq \int I_{\mathcal{Y}_j}(x)\,dx = \delta$ is the volume of A. Similarly,

$$\begin{aligned}\lim_{\delta\to 0}\frac{1}{|A|}\,\mathrm{E}(I_{\mathcal{Y}_j}(y(t))\varphi(t)) &= \int \varphi\Big[\lim_{\delta\to 0}\frac{1}{|A|}\int I_{\mathcal{Y}_j}(y)p_{\varphi(t),y(t)}(\varphi,y)\,dy\Big]\,d\varphi \\ &= \int \varphi p_{\varphi(t),y(t)}(\varphi,\bar{y})\,d\varphi.\end{aligned} \tag{32}$$

Finally, combining (31) and (32) yields

$$\begin{aligned}\lim_{\delta\to 0}\frac{\mathrm{E}(I_{\mathcal{Y}_j}(y(t))\varphi(t))}{\mathrm{E}(I_{\mathcal{Y}_j}(y(t)))} &= \lim_{\delta\to 0}\frac{|A|}{\mathrm{E}(I_{\mathcal{Y}_j}(y(t)))}\frac{\mathrm{E}(I_{\mathcal{Y}_j}(y(t))\varphi(t))}{|A|} \\ &= \frac{1}{p_{y(t)}(\bar{y})}\int \varphi p_{\varphi(t),y(t)}(\varphi,\bar{y})\,d\varphi = \mathrm{E}(\varphi(t)\,|\,y(t)=\bar{y}),\end{aligned}$$

which completes the proof of the theorem. □

Avdelning, Institution
Division, Department

Division of Automatic Control
Department of Electrical Engineering

Datum
Date

2011-11-14

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

LiTH-ISY-R-3031

URL för elektronisk version

`http://www.control.isy.liu.se`

Titel
Title

Sliced Inverse Regression for the Identification of Dynamical Systems

Författare
Author

Christian Lyzell, Martin Enqvist

Sammanfattning
Abstract

The estimation of nonlinear functions can be challenging when the number of independent variables is high. This difficulty may, in certain cases, be reduced by first projecting the independent variables on a lower dimensional subspace before estimating the nonlinearity. In this paper, a statistical nonparametric dimension reduction method called *sliced inverse regression* is presented and a consistency analysis for dynamically dependent variables is given. The straightforward system identification application is the estimation of the number of linear subsystems in a Wiener class system and their corresponding impulse response.

Nyckelord
Keywords

System identification; Dimension reduction; Inverse regression.