Automatica 80 (2017) 75–87

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Leading impulse response identification via the Elastic Net criterion ☆

Giuseppe C. Calafiore, Carlo Novara [1], Michele Taragna

Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi 24, I–10129, Torino, Italy

ARTICLE INFO

Article history:
Received 22 September 2015
Received in revised form
14 July 2016
Accepted 18 January 2017

Keywords:
FIR identification
ℓ_1 regularization
Elastic Net
Lasso
Sparsity

ABSTRACT

This paper deals with the problem of finding a low-complexity estimate of the impulse response of a linear time-invariant discrete-time dynamic system from noise-corrupted input–output data. To this purpose, we introduce an identification criterion formed by the average (over the input perturbations) of a standard prediction error cost, plus an ℓ_1 regularization term which promotes sparse solutions. While it is well known that such criteria do provide solutions with many zeros, a critical issue in our identification context is *where* these zeros are located, since sensible low-order models should be zero in the tail of the impulse response. The flavor of the key results in this paper is that, under quite standard assumptions (such as i.i.d. input and noise sequences and system stability), the estimate of the impulse response resulting from the proposed criterion is indeed identically zero from a certain time index n_l (named the *leading order*) onwards, with arbitrarily high probability, for a sufficiently large data cardinality N. Numerical experiments are reported that support the theoretical results, and comparisons are made with some other state-of-the-art methodologies.

1. Introduction

A large part of the literature on identification of linear time-invariant (LTI) dynamic systems follows a statistical approach (Ljung, 1999a; Söderström & Stoica, 1989), where probabilistic assumptions are made, at least on the noise corrupting the measurements. The techniques available in this context may be classified into two main categories: parametric and nonparametric. Parametric techniques are mainly based on the prediction error method (PEM) or on the maximum likelihood approach, if Gaussian noise is assumed. The identified models belong to finite-dimensional spaces of given order, like FIR, ARX, ARMAX, OE, Laguerre, Kautz or orthonormal basis function models. In order to limit the model complexity and to avoid possible overfitting, a tradeoff between bias and variance is usually considered, and the model order selection is performed by optimizing some suitable cost function — such as the Akaike's information criterion AIC (Akaike, 1974), the Rissanen's Minimum Description Length MDL, or the Bayesian information criterion BIC (Rissanen, 1978; Schwarz, 1978) – and by applying some form of cross validation (CV), like hold-out or leave-one-out. Possible limits of these parametric methods have been pointed out in Chen, Ohlsson, and Ljung (2012), Pillonetto and De Nicolao (2010) and Pillonetto, Chiuso, and De Nicolao (2011), where it is shown that the sample properties of PEM approaches equipped with, e.g., AIC and CV, may be rather unsatisfactory and quite far from those predicted by standard (i.e., without model selection) statistical theory.

Nonparametric techniques aim to obtain the overall system's impulse response as a suitable deconvolution of observed input–output data. In particular, very promising approaches have been recently developed, based on results coming from the machine learning field, see, e.g., Pillonetto, Dinuzzo, Chen, De Nicolao, and Ljung (2014) and the references therein. Rather than postulating finite-dimensional hypothesis spaces, the estimation problem is tackled in an infinite-dimensional space, and the intrinsical ill-posedness of the problem is circumvented by using suitable regularization methods. In particular, the system's impulse response is modeled as a zero-mean Gaussian process, and the prior information is introduced by simply assigning a specific covariance, named *kernel* in the machine learning literature. This procedure can be interpreted as the counterpart of model order selection in the parametric PEM approach and, in some cases, it is shown to be more robust.

In the present paper, a novel nonparametric method is presented, whereby an estimate of the system's impulse response is

☆ The material in this paper was partially presented at the 55th IEEE Conference on Decision and Control, December 12–14, 2016, Las Vegas, NV, USA. This paper was recommended for publication in revised form by Associate Editor Martin Enqvist under the direction of Editor Torsten Söderström.

E-mail addresses: giuseppe.calafiore@polito.it (G.C. Calafiore), carlo.novara@polito.it (C. Novara), michele.taragna@polito.it (M. Taragna).

[1] Fax: +39 011 0907099.

http://dx.doi.org/10.1016/j.automatica.2017.01.011
0005-1098/

obtained by minimizing a suitable cost function that directly takes into account the resulting model complexity. The aim is indeed to obtain a low-complexity model of the system, in the form of a reduced-order FIR (in this sense, the approach is not so far from parametric techniques). A key feature of the proposed approach, representing a relevant improvement over the state of the art, is that it allows for an effective model order selection, without using strong a-priori information on the true system. More specifically, we propose the use of an identification criterion which is a weighted combination of (a) a standard prediction error term, (b) an ℓ_2 regularization term, and (c) an ℓ_1 penalty term which promotes sparse solutions; a full justification for such criterion is given in Section 3.2. This type of criterion corresponds to the so-called Elastic Net cost, which recently has become popular in the machine learning community, see, e.g., De Mol, De Vito, and Rosasco (2009) and Zou and Hastie (2005). Notice that, while it is well known that the use of ℓ_1 regularization leads to sparse solutions, sparsity alone is not a very interesting feature in our identification context. Indeed, reduced-order FIR models are obtained only if the sparsity of the solution follows a specific pattern, whereby the zeros are all concentrated in the tail of the impulse response. Obtaining such a pattern is not obvious, nor a-priori granted by the ℓ_1 regularization. One of the key contributions of this paper is to prove that, under standard assumptions, the impulse response estimated via our Elastic-Net type of criterion has the property of being indeed nonzero only on the initial part of the impulse response (which we shall name the *leading response*), with arbitrarily high probability, if the number of data N is sufficiently large.

The present paper is organized as follows. In Section 2 the notation is set, and some preliminary results on a Chebyshev's type of convergence for random variables are stated. Section 3 describes the linear identification problem of interest, and contains the derivations of the Elastic Net cost. The main results on the recovery of the leading part of the impulse response are contained in Section 4. Section 5 illustrates a practical procedure for implementing the proposed identification scheme. Numerical experiments, including a comparative discussion with other identification methods, are given in Section 6. All proofs are contained in the Appendix.

2. Notation and preliminaries

2.1. Notation

For a vector $x \in \mathbb{R}^N$, we denote by $[x]_i$ the ith entry of x, and we define its *support* as

$$\text{supp}(x) \doteq \{i \in \{1, \dots, N\} : [x]_i \neq 0\}.$$

The notation $\|x\|_p$ represents the standard ℓ_p norm of x, and $\|x\|_0$ denotes the cardinality of supp(x), that is the number of nonzero entries of x.

For a matrix $X \in \mathbb{R}^{N,M}$ (with M possibly equal to ∞), we denote by $[X]_{i,j}$ the entry of X in row i and column j. For $n \leq M$, we denote by $X_{\uparrow n} \in \mathbb{R}^{N,n}$ the sub-matrix formed by the first n columns of X, with $X_{\downarrow n} \in \mathbb{R}^{N,M-n}$ the sub-matrix formed by the columns of X of indices $n+1, \dots, M$, and with $X_{\sharp n}$ the $n \times n$ principal sub-matrix of X. The identity matrix is denoted by I, or by I_n, if we wish to specify its dimension. We denote by $X^\dagger$ the Moore–Penrose pseudo-inverse of X; if X has full column rank, then $X^\dagger = (X^\top X)^{-1} X^\top$.

If x is a random variable, then $\mathbb{E}\{x\}$ denotes the expected value of x, and $\text{var}\{x\}$ denotes its variance: $\text{var}\{x\} = \mathbb{E}\{(x - \mathbb{E}\{x\})^2\}$. $\mathbb{P}$ denotes a probability measure on x. The symbol $\rightsquigarrow$ implies almost sure convergence, and it is formally defined in Section 2.2.1.

2.2. Chebyshev's inequality for certain empirical means

Let x_i, $i = 1, \dots,$ be a sequence of (not necessarily independent) random variables such that $\mathbb{E}\{x_i\} = \mu < \infty$ for all i, $\text{var}\{x_i\} = \sigma_i^2 \leq \overline{\sigma}^2 < \infty$ for all i, and $\mathbb{E}\{(x_i - \mu)(x_j - \mu)\} = 0$ for all $i \neq j$. For given $N \geq 1$, define the empirical mean

$$\hat{x}_N \doteq \frac{1}{N} \sum_{i=1}^{N} x_i.$$

Obviously, from linearity of the expectation, it holds that $\mathbb{E}\{\hat{x}_N\} = \mu$. Further, we have that

$$\begin{aligned}
\sigma^2 \doteq \text{var}\{\hat{x}_N\} &= \mathbb{E}\left\{(\hat{x}_N - \mu)^2\right\} = \frac{1}{N^2} \mathbb{E}\left\{\left[\sum_{i=1}^{N} (x_i - \mu)\right]^2\right\} \\
&= \frac{1}{N^2}\left[\sum_{i=1}^{N} \mathbb{E}\left\{(x_i - \mu)^2\right\} \right. \\
&\quad \left. + \sum_{i=1}^{N} \sum_{j=1,\, j\neq i}^{N} \mathbb{E}\left\{(x_i - \mu)(x_j - \mu)\right\}\right] \\
&= \sum_{i=1}^{N} \sigma_i^2 / N^2 \leq \overline{\sigma}^2 / N,
\end{aligned}$$

where the last passages follow from the fact that the x_is are uncorrelated, and have first moment μ and variance $\sigma_i^2 \leq \overline{\sigma}^2$. Chebyshev's inequality applied to the random variable $\hat{x}_N$ thus states that, for any $\eta > 0$,

$$\mathbb{P}\{|\hat{x}_N - \mu| \geq \eta\sigma\} \leq 1/\eta^2. \tag{1}$$

Since $\eta\sigma \leq \eta\overline{\sigma}/\sqrt{N}$, we have that $\mathbb{P}\{|\hat{x}_N - \mu| \geq \eta\overline{\sigma}/\sqrt{N}\} \leq \mathbb{P}\{|\hat{x}_N - \mu| \geq \eta\sigma\}$, whence, from (1), we obtain that $\mathbb{P}\{|\hat{x}_N - \mu| \geq \eta\overline{\sigma}/\sqrt{N}\} \leq 1/\eta^2$. Equivalently, we can state that, for any $\epsilon > 0$, it holds that

$$\mathbb{P}\{|\hat{x}_N - \mu| \geq \epsilon\} \leq \overline{\sigma}^2 \big/ \left(N\epsilon^2\right).$$

We thus conclude that, for any given accuracy $\epsilon > 0$ and probability $\beta \in (0, 1)$, it holds that

$$\mathbb{P}\{|\hat{x}_N - \mu| \geq \epsilon\} \leq \beta, \quad \forall N \geq \left\lceil \overline{\sigma}^2 \big/ \left(\beta\epsilon^2\right) \right\rceil.$$

Notice that (1) implies that $\mathbb{P}\{|\hat{x}_N - \mu| > \eta\sigma\} \leq 1/\eta^2$; hence, by considering the complementary event, it also holds that $\mathbb{P}\{|\hat{x}_N - \mu| \leq \eta\sigma\} \geq 1 - 1/\eta^2$, from which it follows that

$$\mathbb{P}\{|\hat{x}_N - \mu| \leq \epsilon\} \geq 1 - \overline{\sigma}^2 \big/ \left(N\epsilon^2\right).$$

2.2.1. Meaning of the convergence symbol $\rightsquigarrow$

For a random variable z_N that depends on N and for a given real value $\bar{z}$, the notation $z_N \rightsquigarrow \bar{z}$ means that for any given $\epsilon > 0$ and $\beta \in (0, 1)$ there exists a *finite* integer $N_{\epsilon,\beta}$ such that

$$\mathbb{P}\{|z_N - \bar{z}| \geq \epsilon\} \leq \beta, \quad \forall N \geq N_{\epsilon,\beta}. \tag{2}$$

Notice that $z_N \rightsquigarrow \bar{z}$ implies that z_N converges to $\bar{z}$ *almost surely* (that is, with probability one), as N tends to infinity. However, we are specifically interested in the property in (2), that holds for possibly large, but finite, N.

2.3. Lipschitz functions of random variables

If z_N is the empirical mean of N uncorrelated variables with common mean μ and variance bounded by $\overline{\sigma}^2$ then, from the