Automatica 50 (2014) 657–682

Contents lists available at ScienceDirect

Automatica

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

Survey Paper

# Kernel methods in system identification, machine learning and function estimation: A survey[☆]

Gianluigi Pillonetto[a,1], Francesco Dinuzzo[b], Tianshi Chen[c], Giuseppe De Nicolao[d], Lennart Ljung[c]

[a] Department of Information Engineering, University of Padova, Padova, Italy
[b] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[c] Division of Automatic Control, Linköping University, Linköping, Sweden
[d] Department of Computer Engineering and Systems Science, University of Pavia, Pavia, Italy

ARTICLE INFO

*Article history:*
Received 16 November 2011
Received in revised form
13 January 2014
Accepted 15 January 2014
Available online 25 February 2014

*Keywords:*
Linear system identification
Prediction error methods
Model complexity selection
Bias-variance trade-off
Kernel-based regularization
Inverse problems
Reproducing kernel Hilbert spaces
Gaussian processes

ABSTRACT

Most of the currently used techniques for linear system identification are based on classical estimation paradigms coming from mathematical statistics. In particular, maximum likelihood and prediction error methods represent the mainstream approaches to identification of linear dynamic systems, with a long history of theoretical and algorithmic contributions. Parallel to this, in the machine learning community alternative techniques have been developed. Until recently, there has been little contact between these two worlds. The first aim of this survey is to make accessible to the control community the key mathematical tools and concepts as well as the computational aspects underpinning these learning techniques. In particular, we focus on kernel-based regularization and its connections with reproducing kernel Hilbert spaces and Bayesian estimation of Gaussian processes. The second aim is to demonstrate that learning techniques tailored to the specific features of dynamic systems may outperform conventional parametric approaches for identification of stable linear systems.



## 1. Preamble

System identification is about building mathematical models of dynamic systems from observed input–output data. It is a well established subfield of Automatic Control, with more than 50 years history of theoretical and algorithmic development as well as software packages and industrial applications.

*General aspects.* For time-invariant linear dynamical systems the output is obtained as a convolution between the input and the system's impulse response. This means that system identification is an example of an *inverse problem*: indeed, finding the impulse response from observed data is a *deconvolution problem*. Such problems are quite ubiquitous and appear in biology, physics, and engineering with applications e.g. in medicine, geophysics, and image restoration (Bertero, 1989; De Nicolao, Sparacino, & Cobelli, 1997; Hunt, 1970; Tarantola, 2005). The problem is non trivial as convolution is a continuous operator, e.g. on the space of square integrable functions, but its inverse may not exist or may be unbounded (Phillips, 1962).

The reconstruction of the continuous-time impulse response is always an ill-posed problem since such a function cannot be uniquely inferred from a finite set of observations. Also finite discretizations lead to an ill-conditioned problem, meaning that small errors in the data can lead to large estimation errors. Starting from the seminal works of Tikhonov and Phillips (Phillips, 1962; Tikhonov & Arsenin, 1977), a number of regularization methods have been proposed in the literature to solve the deconvolution problem, e.g. truncated singular value decompositions (Hansen, 1987) and gradient-based techniques (Hanke, 1995; Nemirovskii, 1986; Yao, Rosasco, & Caponnetto, 2007). This means that

☆ This research has been partially supported by the European Community under agreement no. FP7-ICT-223866-FeedNetBack, no. 257462-HYCON2 Network of excellence, by the MIUR FIRB project RBFR12M3AC-Learning meets time: a new computational approach to learning in dynamic systems as well as by the Linnaeus Center CADICS, funded by the Swedish Research Council, and the ERC advanced grant LEARN, no. 287381, funded by the European Research Council. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Editor John Baillieul.

*E-mail addresses:* giapi@dei.unipd.it (G. Pillonetto), fdinuzzo@tuebingen.mpg.de (F. Dinuzzo), tschen@isy.liu.se (T. Chen), giuseppe.denicolao@unipv.it (G. De Nicolao), ljung@isy.liu.se (L. Ljung).

[1] Tel.: +39 3481212375; fax: +39 049 827 7699.

http://dx.doi.org/10.1016/j.automatica.2014.01.001
0005-1098/

regularization should be an important topic and area for system identification.

*Identification techniques*. The most widespread approach to identification of dynamic systems relies on parametric prediction error methods (PEMs), for which a large corpus of theoretical results is available (Ljung, 1999; Söderström & Stoica, 1989). The statistical properties of prediction error (and maximum likelihood) methods are well understood under the assumption that the model class is fixed. They show that such procedures are in some sense optimal, at least for large samples. However, within this parametric paradigm, a key point is the selection of the most adequate model structure. In the "classical, frequentist" framework, this is a question of trade-off between bias and variance, and can be handled by various model validation techniques. This is often carried out by resorting to complexity measures, such as the Akaike's criterion (AIC) (Akaike, 1974) or cross validation (CV), but some inefficiencies related to these classical approaches have been recently pointed out (Chen, Ohlsson, & Ljung, 2012; Pillonetto, Chiuso, & De Nicolao, 2011; Pillonetto & De Nicolao, 2010). In particular, it has been shown that sample properties of PEM approaches, equipped e.g. with AIC or CV, may be unsatisfactory when tested on experimental data, departing sharply from the properties predicted by standard (i.e. without model selection) statistical theory, which suggests that PEM should be asymptotically efficient for Gaussian innovations.

Parallel to this development in system identification, other techniques have been developed in the machine learning community. Until very recently, there has been little contact between these concepts and system identification.

Recent research has shown that the model selection problems can be successfully faced by a different approach to system identification that leads to an interesting cross fertilization with the machine learning field (Pillonetto & De Nicolao, 2010). Rather than postulating finite-dimensional hypothesis spaces, e.g. using ARX, ARMAX or Laguerre models, the new paradigm formulates the problem as function estimation possibly in an infinite-dimensional space. In the context of linear system identification, the elements of such space are all possible impulse responses. The intrinsical ill-posedness of the problem is circumvented using regularization methods that also admit a Bayesian interpretation (Rasmussen & Williams, 2006). In particular, the impulse response is modeled as a zero-mean Gaussian process. In this way, prior information is introduced in the identification process just assigning a covariance, named also kernel in the machine learning literature (Schölkopf & Smola, 2001). In view of the increasing importance of these kernel methods also in the general system identification scenario, the first aim of this survey is to make accessible to the control community some of the key mathematical tools and concepts underlying these learning techniques, e.g. reproducing kernel Hilbert spaces (Aronszajn, 1950; Cucker & Smale, 2001; Saitoh, 1988), kernel methods and regularization networks (Evgeniou, Pontil, & Poggio, 2000; Suykens, Gestel, Brabanter, De Moor, & Vandewalle, 2002; Vapnik, 1998), the representer theorem (Schölkopf, Herbrich, & Smola, 2001; Wahba, 1990) and the connection with the theory of Gaussian processes (Hengland, 2007; Rasmussen & Williams, 2006). It is also pointed out that a straight application of these techniques in the control field is doomed to fail unless some key features of the system identification problem are taken into account. First, as already recalled, the relationship between the unknown function and the measurements is not direct, as typically assumed in the machine learning setting, but instead indirect, through the convolution with the system input. This raises significant analogies with the literature on inverse problems (Bertero, 1989; Tikhonov & Arsenin, 1977). Furthermore, in system identification it is essential that the estimation process be informed on the stability of the impulse response. In this regard, a recent major advance has been the introduction of new kernels which include information on impulse response exponential stability (Chen et al., 2012; Pillonetto & De Nicolao, 2010). These kernels depend on some hyperparameters which can be estimated from data e.g. using marginal likelihood maximization. This procedure is interpretable as the counterpart of model order selection in the classical PEM paradigm but, as it will be shown in the survey, it turns out to be much more robust, appearing to be the real reason of success of these new procedures. Other research directions recently developed have been the justification of the new kernels in terms of Maximum Entropy arguments (Pillonetto & De Nicolao, 2011), the analysis of these new approaches in a classical deterministic framework leading to the derivation of the *optimal kernel* (Chen et al., 2012), as well as the extension of these new techniques to the estimation of optimal predictors (Pillonetto et al., 2011).

*Outline of the survey*. The present survey will deal with this meeting between conventional system identification of linear models and learning techniques. It is divided into three Parts with sections which are relevant, but can be skipped without interrupting the flow of the discussion, marked with a star ⋆.

Part I will describe the status in traditional parametric system identification in discrete-time with an account of how the bias-variance trade-off can be handled also by *regularization techniques*, including their Bayesian interpretation.

Part II is an account of general function estimation – or function learning – theory in a general and abstract setting. This includes the role of RKHS theory for this problem.

Part III treats linear system identification, mainly in continuous-time, as an application of learning the impulse response function from observed data, leaning on general function estimation and its adaptation to the specific properties of impulse responses of dynamic systems. This will link back to the regularizations techniques from the simplistic perspective in Part I. Considerations on computational issues are also included while some mathematical details are gathered in the Appendix.

In conclusion, the scope of this work is twofold. Firstly, our aim is to survey essential results in kernel methods for estimation, that are mostly published outside the control audience, and hence not so well known in this community. Secondly, we want to show that these results have much to offer for estimation problems in the control community, in particular for system identification.

# Part I. Estimating system impulse responses in discrete time

In this part, we study the problem of estimating system impulse responses in discrete time.

## 2. "Classical" system identification

### 2.1. System identification

There is a very extensive literature on system identification, with many text books, like Ljung (1999) and Pintelon and Schoukens (2012a). Most of the techniques for system identification have their origins in estimation paradigms from mathematical statistics, and classical methods like Maximum Likelihood (ML) have been important elements in the area. In Part I the main ingredients of this "classical" view of system identification will be reviewed. For convenience, we will only focus on the single input–single output (SISO) linear time-invariant, stable and causal systems. We will also set the stage for the "kernel methods" for estimating the main characteristics of a system.