Annual Reviews in Control 000 (2016) 1–15

Contents lists available at ScienceDirect

Annual Reviews in Control

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

Review

Regularization and Bayesian learning in dynamical systems: Past, present and future☆

A. Chiuso

Department of Information Engineering, University of Padova, Italy

ARTICLE INFO

Article history:
Received 4 November 2015
Revised 19 February 2016
Accepted 8 March 2016
Available online xxx

Keywords:
Identification
Learning
Numerical methods
Linear systems

ABSTRACT

Regularization and Bayesian methods for system identification have been repopularized in the recent years, and proved to be competitive w.r.t. classical parametric approaches. In this paper we shall make an attempt to illustrate how the use of regularization in system identification has evolved over the years, starting from the early contributions both in the Automatic Control as well as Econometrics and Statistics literature. In particular we shall discuss some fundamental issues such as compound estimation problems and exchangeability which play and important role in regularization and Bayesian approaches, as also illustrated in early publications in Statistics. The historical and foundational issues will be given more emphasis (and space), at the expense of the more recent developments which are only briefly discussed. The main reason for such a choice is that, while the recent literature is readily available, and surveys have already been published on the subject, in the author's opinion a clear link with past work had not been completely clarified.

1. Introduction

About sixty years have passed since the seminal paper by Zadeh (1956), which has coined the name "identification" and triggered research of the Automatic Control community in the broad area of data based dynamical modeling.

In the subsequent ten years the field had achieved such an importance that *IFAC* decided to start the series of *Symposiums on System Identification* (formerly *Symposium on Identification in Automatic Control Systems*) in 1967, two years after the work by Åström and Bohlin (1965) which has laid the foundations of Maximum Likelihood methods (and thus Prediction Error Methods in the Gaussian case) for ARMAX models. I have had the honor and privilege of being a plenary speaker at the 17th Symposium of the series in Beijing, and this paper has been written as a companion to the plenary presentation, which of course could not enter into many of the details that can be found here.

Despite this long history, the field is still lively and active. We believe there are two main reasons why this is so: the first is definitely the increasing importance that data centric methods are playing in many areas of Engineering and Applied Sciences with new challenges arising from the need to process high dimensional data, possibly in real time, and with little (if any) human supervision. A very nice overview of such challenges has been discussed by Lennart Ljung in his plenary at the joint IEEE Conference on Decision and Control and European Control Conference in 2011 (Ljung, Hjalmarsson, & Ohlsson, 2011) and by Mario Sznaier in his SYSID 2012 semi plenary lecture, Sznaier (2012).

The second reason, which is the topic of this paper, has to do with the revitalization of (old) techniques which are rooted in the theory of regularization and Bayesian statistics.

This paper shall be focused on the role these latter techniques play in the recent developments of (linear) system identification, with the main objective to guide the reader from the early developments to the present days, with a (brief) outlook into the future. For reasons of space we will only address linear system identification, even though we believe the methods and tools discussed here have high potential in the nonlinear scenario as well (see Johansen, 1997; Pillonetto, Quang & Chiuso, 2011b; Suykens, 2001 and references therein). We also warn the reader that only the discrete time case will be presented. It is worth stressing that most of the recent results, just briefly discussed in Section 4 and 7, can be framed in a continuous time scenario (see e.g. Pillonetto & De Nicolao, 2010) so that also non-uniform sampling can be handled (see e.g. Neve, De Nicolao, & Marchesi, 2008 for applications with pharmacokinetic data).

More specifically, after having introduced the problem and defined notation in Section 2, we shall provide in Section 3 an overview of parametric Maximum Likelihood/Prediction Error Methods (PEM) as formulated in Åström and Bohlin (1965). We

☆ This work has been partially supported by the FIRB project "Learning meets time" (RBFR12M3AC) funded by MIUR.

E-mail address: chiuso@dei.unipd.it

http://dx.doi.org/10.1016/j.arcontrol.2016.04.013
1367-5788/

warn the reader that *this is not* a paper about ML/PEM and thus no attempt is made to discuss its developments over the years. The main goal of Section 3 will be just to set the notation and to pinpoint the weaknesses of the parametric approach. Section 4 will introduce the regularization approach, with an attempt to provide a complete historical overview, including early work in Statistics and Econometrics where these type of approaches have been first advocated. In order to understand the basic ideas and motivations for the Bayesian approach, Section 5 introduces the related problem of compound estimation and recalls the notion of exchangeability, as a prerequisite to Section 6 where their role in the system identification problem will be discussed. In particular the theory of compound estimation provides, from a classical (read frequentist) perspective, a sound theoretical motivation for adopting a regularization/Bayesian point of view. Finally an overview of some recent research in which I have been personally involved, regarding the design of priors and their use in structure selection problems, will be given in Section 7 and 8, respectively. A brief outlook into the future will be provided in Section 9.

Of course this overview reflects the author's view and other people would have certainly provided a different one. Despite the long list of references I certainly have omitted many relevant ones; yet I still hope this paper can provide a starting point for anybody interested in digging a bit deeper into the roots of Bayesian Learning in System Identification.

2. Statement of the problem

Let $u(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$ be, respectively, the measured *input* and *output* signals in a dynamical system; the purpose of system identification is to find, from a finite collection of input-output data $\{u(t), y(t)\}_{t \in [1, N]}$, a "good" dynamical model which describes the phenomenon under observation. The candidate model will be searched for within a so-called "model set" denoted by $\mathcal{M}$. In this paper we shall use the symbol $\mathcal{M}_n(\theta)$ for parametric model classes where the subscript n denotes the model complexity, i.e. the number of free parameters.

In this paper we shall be concerned with identification of linear models for jointly stationary processes $\{y(t), u(t)\}_{t\in\mathbb{Z}}$, i.e. models described by a convolution

$$y(t) = \sum_{k=1}^{\infty} g_k u(t-k) + \sum_{k=0}^{\infty} h_k e(t-k) \quad t \in \mathbb{Z}. \tag{1}$$

where g and h are the so-called impulse responses of the system and $\{e_t\}_{t\in\mathbb{Z}}$ is a zero mean white noise process which under suitable assumptions is the one-step-ahead prediction error; a convenient description of the linear system (1) is given in terms of the transfer functions

$$G(q) := \sum_{k=1}^{\infty} g_k q^{-k} \quad H(q) := \sum_{k=0}^{\infty} h_k q^{-k}$$

The linear model (1) yields an "optimal" (in the mean square sense) output predictor which shall be denoted later on by $\hat{y}(t|t-1)$. As mentioned above, under suitable assumptions, the noise $e(t)$ in (1) is the so-called *innovation* process $e(t) = y(t) - \hat{y}(t|t-1)$.

In order to simplify the exposition, in this work we shall only deal with feedback free (i.e. assuming that there is no feedback from y to u, see Granger, 1963) Output Error (OE) systems; thus $H(q) = I_p$ will be postulated. All ideas can be extended to handling, without major difficulties (see e.g. Pillonetto, Chiuso, & De Nicolao, 2011a), more general situation involving colored noise (i.e. $H(q) \neq I_p$) as well as the case where feedback is present. This however would obscure the presentation and is thus omitted.

Therefore our focus will be on linear models of the form

$$\begin{aligned} y(t) &= \sum_{k=1}^{\infty} g_k u(t-k) + e(t) \\ &= \hat{y}(t|t-1) + e(t) \end{aligned} \tag{2}$$

where (second order) joint stationarity of $\{y(t), u(t)\}_{t\in\mathbb{Z}}$ implies that $G(q)$ has to be BIBO stable (i.e. analytic outside the open unit disc of the complex plane, $|q| \geq 1$). In turn BIBO stability requires that g_k decays to zero as $k \to \infty$, and therefore the infinite summation in (2) can be approximated by a finite summation

$$y(t) \simeq \sum_{k=1}^{T} g_k u(t-k) + e(t) \tag{3}$$

for a large enough integer T. Since this is always possible up to an arbitrarily small approximation error,[1] in this paper we shall always work with FIR models, assuming exact equality is satisfied in (3). This transforms the infinite dimensional model (2) into a finite dimensional one (3). All the results in this paper could indeed be formulated with reference to the infinite dimensional model (2), at the price of bringing so called Reproducing Kernel Hilbert Spaces (RKHS) (Aronszajn, 1950; Saitoh, 1988) into the picture. In our opinion this only entails additional difficulties for the reader and essentially no gain in terms of tools and results and will thus be avoided. The interested reader is referred to Pillonetto and De Nicolao (2010), Pillonetto et al. (2011a) and Pillonetto, Dinuzzo, Chen, Nicolao, and Ljung (2014). We would like to remind the reader that a detailed study of the asymptotic properties of Bayes procedures for infinite dimensional models is delicate and outside the scope of this paper, see for instance Knapik, van der Vaart, and van Zanten (2011).

In the following we shall use the notation $Y \in \mathbb{R}^{pN}$ to denote the (column) vector containing the stacked outputs $y(t)$, $t \in [1, N]$ and $\hat{Y}(g)$ the vector of stacked predictors $\hat{y}(t|t-1)$, $t \in [1, N]$ which is a linear function of the impulse response coefficients g_k:

$$\hat{Y}(g) = \Phi g$$

where $\Phi \in \mathbb{R}^{pN \times pmT}$ is a suitable matrix built with the input data $u(t)$, while the column vector $g \in \mathbb{R}^{pmT}$ contains the (vectorized) impulse response matrix coefficient $g_k \in \mathbb{R}^{p\times m}$, $k \in [1, T]$.

Performing identification of g (i.e. estimation of the impulse response from a finite set of input output data) can be thus framed as estimation of the unknown g in the linear model

$$Y = \Phi g + E \quad g \in \mathbb{R}^d \quad d := pmT \tag{4}$$

Unfortunately the dimension d of the unknown vector g may be very large (and possibly much larger that the length of the available data Y) so that the inverse problem of determining g from Y in (4), e.g. minimizing the square loss

$$\hat{g}_{LS} := \arg\min_g \|Y - \Phi g\|^2, \tag{5}$$

may be (very) ill conditioned. This may be due to the fact that the input process lives in a high dimensional space (pm large, so that many impulse responses need to be estimated) or simply because g_k decays very slowly to zero and thus many lags need to be included in the parameter vector g.

To face this problem one has to impose constraints on the structure of the vector g. One possibility is to parameterize $g_k = g_k(\theta)$ using a vector $\theta \in \mathbb{R}^n$, $n << pN$. In the remaining part of the paper we shall sometimes make explicit the dimension of the parameter vector using the notation θ_n. For instance one may assume

[1] Rigorously one should account for the transient effect, which can be beneficial for small data sets where N (and thus implicitly T) is necessarily small. This can be done rather easily estimating the free response for each output channel. Details are outside the scope of this paper and shall not be discussed here.

Please cite this article as: A. Chiuso, Regularization and Bayesian learning in dynamical systems: Past, present and future, Annual Reviews in Control (2016), http://dx.doi.org/10.1016/j.arcontrol.2016.04.013