Automatica 79 (2017) 326–339

Contents lists available at ScienceDirect

Automatica

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

Maximum Entropy vector kernels for MIMO system identification☆

Giulia Prando, Alessandro Chiuso, Gianluigi Pillonetto

Department of Information Engineering, University of Padova, Padova, Italy

ARTICLE INFO

Article history:
Received 11 August 2015
Received in revised form
5 September 2016
Accepted 22 December 2016

Keywords:
Linear system identification
Kernel-based regularization
Hankel nuclear norm
Bayesian interpretation of regularization
Gaussian processes
Maximum entropy paradigm

ABSTRACT

Recent contributions have framed linear system identification as a nonparametric regularized inverse problem. Relying on ℓ_2-type regularization which accounts for the stability and smoothness of the impulse response to be estimated, these approaches have been shown to be competitive w.r.t. classical parametric methods. In this paper, adopting Maximum Entropy arguments, we derive a new ℓ_2 penalty; to do so we exploit the structure of the Hankel matrix, thus controlling at the same time complexity, measured by the McMillan degree, stability and smoothness of the identified models. As a special case, we recover the nuclear norm penalty on the squared block Hankel matrix. In contrast with the previous literature on reweighted nuclear norm penalties, our kernel is described by a small number of hyper-parameters, which are iteratively updated through marginal likelihood maximization; constraining the structure of the kernel acts as a (hyper)regularizer which helps controlling the effective degrees of freedom of our estimator. To optimize the marginal likelihood, we adapt a Scaled Gradient Projection (SGP) algorithm which is proved to be significantly computationally cheaper than other first and second order off-the-shelf optimization methods. The paper also contains an extensive comparison with many state-of-the-art methods on several Monte-Carlo studies, which confirms the effectiveness of our procedure.

1. Introduction

Although linear system identification is sometimes considered a mature field, with a wide and solid literature summarized in the well known textbooks (Ljung, 1999; Söderström & Stoica, 1989), the recent developments on regularization based methods have brought new insights and opened new avenues.

The most common "classical" approaches are parametric Prediction Error Methods (PEM) (Ljung, 1999; Söderström & Stoica, 1989) and subspace methods (Chiuso & Picci, 2003; Larimore, 1983; Lindquist & Picci, 2015; Van Overschee & De Moor, 1996; Verhaegen, 1994). These techniques require that a model complexity (the *order* hereon) is fixed, and thus estimated, first. As an alternative to the standard parametric approach, recent literature has proposed a Bayesian perspective, leading to a class of regularized methods (Chen, Andersen, Ljung, Chiuso, & Pillonetto, 2014; Chen, Ohlsson, & Ljung, 2012; Pillonetto, Chiuso, & De Nicolao, 2011; Pillonetto & De Nicolao, 2010; Pillonetto, Dinuzzo, Chen, De Nicolao, & Ljung, 2014; Zorzi & Chiuso, 2015, 2017). The use of Bayesian inference is not new in the field of identification and time-series estimation: early works on this topic appeared in the late '70, early '80 (Akaike, 1979; Doan, Litterman, & Sims, 1984; Goodwin, Gevers, & Ninness, 1992; Kitagawa & Gersh, 1985); see Chiuso (2016) for an overview.

The Bayesian paradigm considers the impulse response as a stochastic process whose prior distribution describes the model class in a flexible manner. This allows to face the so-called bias/variance trade-off by jointly performing estimation and model selection.

In Chen et al. (2012), Pillonetto and De Nicolao (2010) and Pillonetto et al. (2011) prior distributions, which can also be derived using Maximum Entropy arguments (Carli, Chen, & Ljung, 2014; Chen et al., 2016; De Nicolao, Ferrari-Trecate, & Lecchini, 1998; Pillonetto & De Nicolao, 2011), are designed to encode smoothness and stability of the impulse response to be estimated, leading to ℓ_2-type penalties so that closed-form solutions are available.

In this paper, we focus on the identification of Multi Input–Multi Output (MIMO) systems, where matrix impulse responses have to be identified. Similar problems are encountered in multi-task

☆ This work has been partially supported by the FIRB project "Learning meets time" (RBFR12M3AC) funded by MIUR and by Progetto di Ateneo CPDA147754/14 "New statistical learning approach for multi-agents adaptive estimation and coverage control". The material in this paper was partially presented at the 53rd IEEE Conference on Decision and Control, December 15–17, 2014, Los Angeles, CA, USA and at the 17th IFAC Symposium on System Identification, October 19–21, 2015, Beijing, China. This paper was recommended for publication in revised form by Associate Editor Martin Enqvist under the direction of Editor Torsten Söderström.

E-mail addresses: prandogi@dei.unipd.it (G. Prando), chiuso@dei.unipd.it (A. Chiuso), giapi@dei.unipd.it (G. Pillonetto).

http://dx.doi.org/10.1016/j.automatica.2017.01.020
0005-1098/

learning where one would like to simultaneously estimate multiple functions while also exploiting their mutual information. To this aim Caruana (1997), Bakker and Heskes (2003), Micchelli and Pontil (2005), Evgeniou, Micchelli, and Pontil (2005) and Pillonetto and Dinuzzo (2010) have considered vector-valued kernels which account for the smoothness of the functions to be estimated. In the identification of finite dimensional linear MIMO systems, the coupling between different input–output channels is captured by the Hankel matrix, which has finite rank equal to the McMillan degree of the system.

The Hankel matrix and its properties have already been thoroughly exploited in subspace methods, where also Vector AutoRegressive Models (VARX) estimated under the PEM framework play a fundamental role; in fact, it has been shown in Chiuso (2007b) (see also Chiuso, 2007a, 2010) that certain subspace methods can be seen as estimation of a long (i.e. "nonparametric" in the context of this paper) VARX model followed by a suitable (data based) model reduction step. This paper goes one step further, by merging these two steps in one. While subspace methods reduce the order of the estimated VARX model via a model reduction step, in this paper regularization takes care of both stability and "complexity" (in terms of McMillan degree) at once, while estimating the VARX model itself.

Within this framework, our recent works (Prando & Chiuso, 2015; Prando, Chiuso, & Pillonetto, 2014; Prando, Pillonetto, & Chiuso, 2015) have attempted to merge the benefits of accounting for both stability/smoothness and complexity when building prior models. The main contributions of this work, w.r.t. the above referenced papers are as follows: (i) development, by means of MaxEnt arguments, of a new kernel encoding both complexity, as measured by its McMillan degree, as well as smoothness and stability; (ii) a new tailored Scaled Gradient Projection (SGP) method for marginal likelihood optimization, inspired by the one introduced in Bonettini, Chiuso, and Prato (2015); (iii) an extensive simulation study, where the proposed identification algorithm is compared with classical and state-of-the art identification methods, including PEM (Ljung, 1999), N4SID (Van Overschee & De Moor, 1996), Stable Spline (Pillonetto et al., 2011), reweighted nuclear norm-based algorithms (Mohan & Fazel, 2010) and regularized subspace methods (Verhaegen & Hansson, 2014).

The connection of the procedure introduced in this paper with the existing literature is now briefly discussed. Further details can be found in the Appendices. The prior distribution here introduced leads, as a special case, to an Hankel nuclear norm penalty, an heuristic related to that proposed in Fazel, Hindi, and Boyd (2001) as a convex surrogate to the rank function. In the system identification literature, the nuclear norm heuristic has also been applied in the context of subspace identification (Hansson, Liu, & Vandenberghe, 2012; Liu & Vandenberghe, 2009; Verhaegen & Hansson, 2014), even in the presence of incomplete datasets (Liu, Hansson, & Vandenberghe, 2013), to control the order of the estimated model. PEM methods equipped with nuclear norm penalties on the Hankel matrix built with the Markov parameters have also been considered (Grossmann, Jones, & Morari, 2009; Hjalmarsson, Welsh, & Rojas, 2012). Refer to (Prando et al., 2015) for a brief survey on the topic.

However, direct use of nuclear norm (or atomic) penalties may lead to undesired behaviour, as suggested and studied in Pillonetto, Chen, Chiuso, Ljung, and De Nicolao (2016), due to the fact that nuclear norm is not able alone to guarantee stability and smoothness of the estimated impulse responses. To address this limitation, Chiuso, Chen, Ljung, and Pillonetto (2013) already suggested the combination of the stability/smoothness penalty with the nuclear norm one; differently from the prior presented in this paper, the formulation given in Chiuso et al. (2013) did not allow to adopt marginal likelihood maximization to estimate the regularization parameters.

Exploiting the structure of the prior distribution used in this paper, we design an iterative procedure which alternatively updates the impulse response estimate and the hyper-parameters defining the prior. Our algorithm, which is related to iterative reweighted methods used in compressed sensing and signal processing (Candes, Wakin, & Boyd, 2008; Chartrand & Yin, 2008; Daubechies, Devore, Fornasier, & Güntürk, 2010; Fornasier, Rauhut, & Ward, 2011; Mohan & Fazel, 2012) and so-called *Sparse Bayesian Learning* (SBL) (Tipping & Smola, 2001; Wipf & Nagarajan, 2010), differs from the previous literature in that the regularization matrix takes on a very special structure, described by few hyper-parameters. With this special structure, the weights update does not admit a closed-form solution and thus direct optimization of the marginal likelihood needs to be performed.

While a clear-cut conclusion in terms of relative performance cannot be drawn at the moment, it is fair to say that (a) the new method developed in this paper outperforms the classical "Stable Spline" (Pillonetto et al., 2011), especially when dealing with MIMO systems; (b) the new method outperforms a Reweighted Nuclear Norm algorithm in certain scenarios (e.g. a "mildly-resonant" fourth order system) while performing comparably in others (e.g. randomly generated "large" MIMO systems).

The paper is organized as follows. Section 2 introduces the problem and Section 3 briefly frames system identification in the context of Bayesian estimation. In Section 4, Maximum Entropy arguments are used to derive a family of prior distributions. Section 5 illustrates our algorithm while Section 6 describes the adaptation of a Scaled Gradient Projection method, which is used to solve the marginal likelihood optimization problem. An extensive experimental study is conducted in Section 7, while some concluding remarks are drawn in Section 8.

Notation: In the following, $\mathbb{R}$, $\mathbb{R}_+ := [0, \infty)$, $\mathbb{Z}$ and $\mathbb{N}$, respectively, denote the set of real, positive real, integers and natural numbers. $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$, respectively, indicate the set of n-dimensional real vectors, and $m \times n$ real matrices. The transpose of $A \in \mathbb{R}^{m\times n}$ is denoted as $A^\top$. 0_n, $0_{m\times n}$ and I_n, respectively, represent the zero vector in $\mathbb{R}^n$, the zero matrix in $\mathbb{R}^{m\times n}$ and the $n \times n$ identity matrix. The symbol $\otimes$ denotes the Kronecker product, $\mathcal{N}(\mu, \sigma)$ the Gaussian distribution with mean μ and variance σ. Given $v \in \mathbb{R}^n$, diag(v) is a diagonal matrix of size $n\times n$ with the diagonal given by v. Given matrices $V_i \in \mathbb{R}^{m_i\times n_i}$, $i = 1, .., n$, $blkdiag(V_1, \ldots, V_n)$ denotes the block-diagonal matrix of size $(m_1 + \cdots + m_n) \times (n_1 + \cdots + n_n)$ with the V_i's as diagonal blocks. $\mathbb{E}[\cdot]$ and Tr $\{\cdot\}$ will, respectively, indicate the expectation and trace operators.

2. Problem formulation

We consider the following linear, causal and time-invariant (LTI) Output-Error (OE) system:

$$y(t) = H(q)u(t) + e(t) \tag{1}$$

where $y(t) = [y_1(t) \cdots y_p(t)]^\top \in \mathbb{R}^p$ is the p-dimensional output signal, $u(t) = [u_1(t) \cdots u_m(t)]^\top \in \mathbb{R}^m$ is the m-dimensional input signal, $e(t)$ is the additive noise and

$$H(q) = \sum_{k=1}^{\infty} h(k)q^{-k} \tag{2}$$

is the system transfer function with q^{-1} being the backward shift operator: $q^{-1}u(t) = u(t-1)$. For simplicity, we will assume the presence of a delay in $H(q)$, i.e. $h(0) = H(\infty) = 0$. In addition, we assume $e(t) \sim \mathcal{N}(0_p, \Sigma)$, $\Sigma = \text{diag}(\sigma)$, $\sigma = [\sigma_1 \cdots \sigma_p]^\top$.

The objective is to estimate, from a finite set of input–output data $\mathcal{D}_N = \{u(t), y(t);\ t = 1, \ldots, N\}$, the impulse response coefficients $\left\{h(k) \in \mathbb{R}^{p\times m};\ k = 1, \ldots, \infty\right\}$.