Automatica 48 (2012) 1553–1565

Contents lists available at SciVerse ScienceDirect

Automatica

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

A Bayesian approach to sparse dynamic network identification☆

Alessandro Chiuso [1], Gianluigi Pillonetto

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

ARTICLE INFO

Article history:
Received 21 February 2011
Received in revised form
7 November 2011
Accepted 6 February 2012
Available online 26 June 2012

Keywords:
Linear system identification
Sparsity inducing priors
Kernel-based methods
Lasso
Elastic Net
Gaussian processes

ABSTRACT

Modeling and identification of high dimensional systems, involving signals with many components, poses severe challenges to off-the-shelf techniques for system identification. This is particularly so when relatively small data sets, as compared to the number signal components, have to be used. It is often the case that each component of the measured signal can be described in terms of a few other measured variables and these dependences can be encoded in a graphical way via so called "Dynamic Bayesian Networks''. The problem of finding the interconnection structure as well as estimating the dynamic models can be posed as a system identification problem which involves variable selection. While this variable selection could be performed via standard selection techniques, computational complexity may however be a critical issue, being combinatorial in the number of inputs and outputs. In this paper we introduce two new nonparametric techniques which borrow ideas from a recently introduced kernel estimator called "stable-spline'' as well as from sparsity inducing priors which use ℓ_1-type penalties. Numerical experiments regarding estimation of large scale sparse (ARMAX) models show that this technique provides a definite advantage over a group LAR algorithm and state-of-the-art parametric identification techniques based on prediction error minimization.

1. Introduction

Black-box identification approaches are widely used to learn dynamic models from a finite set of input/output data (Ljung, 1999; Soderstrom & Stoica, 1989). In particular, in this paper we focus on the identification of *large scale* linear systems that involve a wide amount of variables and find important applications in many different domains such as chemical engineering, econometrics/finance, computer vision, systems biology, social networks and so on (Banbura, Giannone, & Reichlin, 2008; Kolaczyk, 2009; Mohammadpour & Grigoriadis, 2010).

In engineering applications, when data are collected from a physical plant, it is often the case that there is an underlying interconnection structure; for instance the overall system could be the interconnection via cascade, parallel, feedback and combinations thereof of many dynamical systems. In this scenario any given variable may be directly related to only a few other variables. This sort of structure, which may be self-evident in certain engineering domains where the system has been designed via interconnection, may also pop up when modeling phenomena involving a large number of variables. These interconnections, such as parallel, series or feedback, may be intrinsic (e.g. in economic and social "systems'') without being related to "physical'' control loops. One convenient and popular way of graphically describing these relations is via so called graphical models (Lauritzen, 1996).

In the static Gaussian case, the "relation'' between variables can be expressed in terms of conditional independence conditions between subsets of them, see e.g. Dempster (1972). Estimation of sparse graphical models has been the subject of intense research which is impossible to survey here; we only point the reader to the early paper (Meinshausen & Bühlmann, 2006) which proposes using the Lasso for this purpose. In the dynamic case, i.e. when observed data are trajectories of (possibly stationary) stochastic processes, one may consider several notions of conditional independence which can be encoded via the so-called time series correlation (TSC) graphs, Granger causality graphs and "partial correlation'' graphs, see Dahlhaus and Eichler (2003) for details.

When the number of measured variables is very large and possibly larger than the number of data available (i.e. the number of "samples'' available for statistical inference), even though there is no "physical'' underlying network, constructing meaningful models which are useful for prediction/monitoring/intepretation

☆ This research has been partially supported by the PRIN Project "New Methods and Algorithms for Identification and Adaptive Control of Technological Systems'', by the Progetto di Ateneo CPDA090135/09 funded by the University of Padova and by the European Community's Seventh Framework Programme [FP7/2007-2013] under agreement n. FP7-ICT-223866-FeedNetBack and under grant agreement n. 257462 HYCON2 Network of excellence. The material in this paper was partially presented at the 49th IEEE Conference on Decision and Control (CDC 2010), December 15–17, 2010, Atlanta, Georgia, USA and the 24th Neural Information Processing Symposium (NIPS), December 6–11, 2010, Vancouver. This paper was recommended for publication in revised form by Associate Editor Wolfgang Scherrer under the direction of Editor Torsten Söderström.

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

[1] Tel.: +39 0498277709; fax: +39 049 8277699.

0005-1098/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2012.05.054

requires trading off model complexity vs. fit. In a parametric setup this complexity depends on the number of parameters which is related to both the complexity of each "subsystem" (e.g. measured via its order) as well as to their number (i.e. the number of dynamical systems which are "non zero"). Estimation problems involving variable selection, which can be also framed as determining connectivity in a suitable graphical description, have been recently studied in the literature, see for instance Materassi and Innocenti (2010), Materassi and Salapaka (2010), Napoletani and Sauer (2008), Timme (2007) and references therein. In the paper Timme (2007) coupled nonlinear oscillators (Kuramoto type) are considered where the coupling strengths are to be estimated; in Napoletani and Sauer (2008) nonlinear dynamics are allowed and the attention is restricted to the linear term[2] in the state update equation, equivalent to a vector autoregressive (VAR) model of order one. In both cases it is assumed that the entire state space is measurable and an ℓ_1-penalized regression problem is solved for estimating the coupling strenghts/linear approximations. Sparse models under "smoothing" conditional independence relations, encoded by "partial correlation" graphs or equivalently via zeros in the inverse spectrum Brillinger (1981), have been recently studied in the literature. For instance, Songsiri and Vandeberghe (2010) considers VAR models and ℓ_1-type penalized regression while Avventi, Lindquist, and Wahlberg (2010) considers ARMA models; in Materassi and Innocenti (2010) and Materassi and Salapaka (2010) a methodology based on smoothing *a la* Wiener is proposed, where interconnections are found by putting a threshold on the estimated transfer functions.

In this work we shall focus on stationary stochastic Gaussian processes described via Granger causality concepts, where conditional independence conditions encode the fact that the prediction of (the future of) one variable (which we shall call "output variable") may require only the past history of few other variables (which we shall call "inputs") plus possibly its own past. This can be represented with a graph where nodes are variables and (directed) edges are (non zero) transfer functions, self-loops encoding dependence on the "output" own past.[3] The concept of Granger causality is, in fact, the most "natural" description in the context of causal control systems since, under rather mild conditions (Gevers & Anderson, 1982), the model obtained in this way is the unique internally stable feedback interconnection which describes the second order statistics of the stationary process under study. In general both the dynamical systems and the interconnection structure are unknown and have to be inferred from data. Without loss of generality we shall address the problem of modeling the relation between one node in this graph (the "output" variable) and all the other measured variables (the "inputs"). Beyond linearity, we shall not make any assumption on each subsystem (e.g. no knowledge of system orders). Our focus is both on finding the underlying connection structure (if any) as well as obtaining reliable and easily interpretable models which can be used, e.g. for prediction/monitoring etc.

Note that one may not be interested at all in building a complete "network of dependences" for the joint process (u, y) but just in modeling an "output" y as a function of the inputs u. This of course involves selecting the most relevant variables and therefore the identification procedure should be sparsity-favoring. Such sparsity principle permeates many well known techniques in machine learning and signal processing such as feature selection, selective shrinkage and compressed sensing (Donoho, 2006; Hastie & Tibshirani, 1990). Recently proposed estimation techniques which induce sparse models include the well known Lasso (Tibshirani, 1996) and Least Angle Regression (LAR) (Efron, Hastie, Johnstone, & Tibshirani, 2004) where variable selection is performed exploiting the ℓ_1 norm. Extensions of this procedure for group selection include Group Lasso and Group LAR (GLAR) (Yuan & Lin, 2006) where the sum of the Euclidean norms of each group (in place of the absolute value of the single components) is used. Theoretical analysis of these approaches and connections with the multiple kernel learning problem can be found in Bach (2008), Micchelli and Pontil (2005), Zou (2006) and Zhao and Yu (2006). We warn the reader that one should not take "sparse" estimators as *panacea*; it is for instance shown in Leeb and Pötscher (2008) that sparse estimators which possess some sort of "Oracle property" (Fan & Li, 2001) have unbounded (normalized) maximal risk as the sample size increases.

Most of the work available in the literature addresses the "static" (i.e. with no "time-dependence") scenario while very little, with some exception (Hsu, Hung, & Chang, 2008; Wang, Li, & Tsai, 2007), can be found regarding the identification of dynamic systems.

In this paper we adopt a Bayesian point of view to prediction and identification of sparse linear systems. Our starting point is the new identification paradigm developed in Pillonetto and De Nicolao (2010) that relies on nonparametric estimation of impulse responses (see also Pillonetto, Chiuso, and De Nicolao (2011) for extensions to predictor estimation). Expanding on our recent works Chiuso and Pillonetto (2010a,b) we extend this nonparametric paradigm to the design of optimal linear predictors so as to jointly perform identification and variable selection. Without loss of generality, analysis is restricted to MISO systems, where the variable to be predicted is called "output variable" and all the other (say $m - 1$) available variables are called "inputs". In this way we interpret the predictor as a system with m inputs (given by the past outputs and inputs) and one output (output predictions).

We consider two approaches: the first, which we shall call *Stable-Spline GLAR* (SSGLAR), is based on the GLAR algorithm in Yuan and Lin (2006) and can be seen as a variation of the so-called "elastic net" (Zou & Hastie, 2005); the second, which we shall call *Stable-Spline Exponential Hyperprior* (SSEH) assigns an exponential hierarchical hyperprior with a common hypervariance to the scale factors. This second approach has connections with the so-called *Relevance Vector Machine* in Tipping (2001). The hierarchical hyperprior favors sparsity through an ℓ_1 penalty on kernel hyperparameters. Inducing sparsity by hyperpriors is an important feature of our second approach. In fact, this permits to obtain the marginal posterior of the hyperparameters in closed form and hence also their estimates in a robust way. Once the kernels are selected, the impulse responses are obtained in closed form via the Representer Theorem (Aronszajn, 1950). As we shall see, however, SSEH requires solving a non-linear optimization problem which may benefit from a "good" initialization. We shall argue that a forward-selection type of procedure provides a robust and computationally attractive way of initializing SSEH.

Numerical experiments involving sparse ARMAX systems show that this approach provides a definite advantage over both the standard GLAR (applied to ARX models) and PEM (equipped with AIC or BIC) in terms of predictive capability on new output data while also effectively capturing the "structural" properties of the dynamic network, i.e. being able to identify correctly, with high probability, the absence of dynamic links between certain variables.

The paper is organized as follows: Section 2 contains the problem formulation while Section 3 contains some background material on the nonparametric approach to system identification introduced in Pillonetto and De Nicolao (2010) and

[2] Thinking of a first order Taylor expansion around the trajectory.

[3] In the language of classical System Identification, dependence of the predictor on the past outputs will result in ARMAX models, lack of dependence in Output Error (OE) models.