Automatica 96 (2018) 280–292

Contents lists available at ScienceDirect

Automatica

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

Maximum likelihood identification of stable linear dynamical systems ☆

Jack Umenberger [a,*], Johan Wågberg [b], Ian R. Manchester [a], Thomas B. Schön [b]

[a] *School of Aerospace, Mechanical and Mechatronic Engineering, University of Sydney, Australia*
[b] *Department of Information Technology, Uppsala University, Sweden*

ARTICLE INFO

Article history:
Received 28 March 2016
Received in revised form 16 March 2018
Accepted 8 June 2018

Keywords:
System identification
Expectation maximization
Lagrangian relaxation
Convex optimization

ABSTRACT

This paper concerns maximum likelihood identification of linear time invariant state space models, subject to model stability constraints. We combine Expectation Maximization (EM) and Lagrangian relaxation to build tight bounds on the likelihood that can be optimized over a convex parametrization of all stable linear models using semidefinite programming. In particular, we propose two new algorithms: EM with latent States & Lagrangian relaxation (EMSL), and EM with latent Disturbances & Lagrangian relaxation (EMDL). We show that EMSL provides tighter bounds on the likelihood when the effect of disturbances is more significant than the effect of measurement noise, and EMDL provides tighter bounds when the situation is reversed. We also show that EMDL gives the most broadly applicable formulation of EM for identification of models with singular disturbance covariance. The two new algorithms are validated with extensive numerical simulations.

1. Introduction

Linear time invariant (LTI) state space models provide a useful approximation of dynamical system behavior in a multitude of applications. In situations where models cannot be derived from first principles, some form of data-driven modeling, i.e. system identification, is appropriate (Ljung, 1999). This paper is concerned with identification of discrete-time linear Gaussian state space (LGSS) models,

$$x_{t+1} = Ax_t + Bu_t + w_t, \tag{1a}$$
$$y_t = Cx_t + Du_t + v_t, \tag{1b}$$

where $x_t \in \mathbb{R}^{n_x}$ denotes the system state, and $u_t \in \mathbb{R}^{n_u}$, $y_t \in \mathbb{R}^{n_y}$ denote the observed input and output, respectively (henceforth, resp.). The disturbances, $w_t \in \mathbb{R}^{n_w}$ and measurement noise, v_t, are modeled as zero mean Gaussian white noise processes, while the uncertainty in the initial condition x_1 is modeled by a Gaussian distribution, i.e.

$$w_t \sim \mathcal{N}(0, \Sigma_w),\ v_t \sim \mathcal{N}(0, \Sigma_v),\ x_1 \sim \mathcal{N}(\mu, \Sigma_1). \tag{2}$$

For convenience, all unknown model parameters are denoted by the variable $\theta = \{\mu, \Sigma_1, \Sigma_w, \Sigma_v, A, B, C, D\}$.

In this work, we seek the maximum likelihood (ML) estimate of the model parameters θ, given measurements $u_{1:T}$ and $y_{1:T}$, subject to model stability constraints, i.e.

$$\hat{\theta}^{\text{ML}} = \arg\max_{\theta} p_\theta(u_{1:T}, y_{1:T}) \text{ s.t. } A \in \mathcal{S}. \tag{3}$$

ML methods have been studied extensively and enjoy desirable properties, such as asymptotic efficiency; see, e.g., Ljung (1999, Chapters 7 and 9).

Identification of LTI systems is complicated by (at least) two factors: *latent variables* and *model stability*, the latter being an essential property in many applications. Typically, observed data consists of inputs and (noisy) outputs only; the internal states and/or exogenous disturbances are *latent* or 'hidden'. Bilinearity of (1) in x and θ means that the joint set of feasible states and parameters is nonconvex. Additionally, even if x is known, the set of Schur stable matrices, which we denote $\mathcal{S}$, is also nonconvex.

Various strategies have been developed to deal with the problem of latent variables. *Marginalization*, for instance, involves integrating out (i.e. marginalizing over) the latent variables, leaving θ as the only quantity to be estimated. This approach is adopted by prediction error methods (Ljung, 1999, 2002) (PEM)

☆ This research was financially supported by the project *Learning flexible models for nonlinear dynamics* (contract number: 2017-03807), funded by the Swedish Research Council and by the Swedish Foundation for Strategic Research (SSF) via the project *ASSEMBLE* (contract number: RIT15-0012). The material in this paper was partially presented 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.

* Corresponding author.
E-mail addresses: j.umenberger@acfr.usyd.edu.au (J. Umenberger), johan.wagberg@it.uu.se (J. Wågberg), i.manchester@acfr.usyd.edu.au (I.R. Manchester), thomas.schon@it.uu.se (T.B. Schön).

https://doi.org/10.1016/j.automatica.2018.06.036
0005-1098/

and the Metropolis–Hastings algorithm (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953).

Alternatively, one may treat the latent variables as additional quantities to be estimated together with the model parameters. Such a strategy is termed *data augmentation*, and examples include subspace methods (Larimore, 1983; Van Overschee & De Moor, 1994), and the Expectation Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; Gibson & Ninness, 2005; Schön, Wills, & Ninness, 2011; Shumway & Stoffer, 1982). The augmentation together with appropriate priors also allows for closed form expressions in a Gibbs sampler (Geman & Geman, 1984; Wills, Schön, Lindsten, & Ninness, 2012), (as a special case of the Metropolis–Hastings algorithm).

Recently, a new family of methods have been developed in which one *supremizes over* the latent variables, with an appropriate multiplier, to obtain convex upper bounds for quality-of-fit cost functions, such as output error (Megretski, 2008; Tobenkin, Manchester, & Megretski, 2017). An important technique employed in this approach is a type of Lagrangian relaxation, similar to a method widely applied in combinatorial optimization (Lemaréchal, 2001) and robust control, where it is referred to as the S-procedure (Pólik & Terlaky, 2007; Yakubovich, 1971).

The problem of model stability has also seen considerable attention over the years. In subspace identification, a number of strategies have been proposed: Maciejowski (1995) showed that stability can be guaranteed by augmenting the extended observability matrix with rows of zeros; in Van Gestel, Suykens, Van Dooren, and De Moor (2001), regularization was used to constrain the spectral radius of the identified A to a user-specified value; Lacy and Bernstein (2002) constrained the largest singular value of A to be less than unity, using a linear matrix inequality (LMI), yielding sufficient albeit conservative conditions for stability; the follow-up work of Lacy and Bernstein (2003) introduced an LMI parametrization of all stable models, $\mathcal{S}$; this approach was generalized in Miller and De Callafon (2013) to constrain the eigenvalues of A to arbitrary convex regions of the complex plane. However, these subspace methods do not fall within, nor inherit the desirable properties of, the ML framework; e.g. Maciejowski (1995) is known to bias the estimated model, and even unconstrained subspace methods are generally considered to be less accurate than PEM (Favoreel, De Moor, & Van Overschee, 2000). Furthermore, Lacy and Bernstein (2003) replace the least-squares objective with a weighted projection which, as noted in Siddiqi, Boots, and Gordon (2007) can produce substantial distortions.

As a middle ground between the conservatism of Lacy and Bernstein (2002) and the distortions of Lacy and Bernstein (2003), the authors of Siddiqi et al. (2007) proposed a constraint generation approach; cf. also Boots (0000). The method takes as its starting point an unconstrained least squares problem, such as those arising in subspace identification or EM with latent states, and then iteratively introduces linear constraints until a stable model is identified. This leaves the desired cost function undistorted; however, the resulting polytopic approximation of $\mathcal{S}$ excludes many stable systems from consideration.

In output-error (a.k.a. simulation-error) identification, which can be interpreted as a special case of ML with no disturbances, convex optimization approaches have been developed based on LMI parameterizations of all stable models and convex bounds on output error, including the Lagrangian relaxation mentioned above (Tobenkin et al., 2017; Tobenkin, Manchester, Wang, Megretski, & Tedrake, 2010; Umenberger & Manchester, 2016). However, due to the approximation of output error these "one-shot" convex optimization methods will generally be biased and will not produce true ML estimates.

In contrast to the above approaches, in this paper we maximize the true likelihood over a complete convex parametrization of all stable models. We do so by leveraging the underlying similarities between EM and Lagrangian relaxation to incorporate model stability constraints into the ML framework. The EM algorithm is an iterative approach to ML estimation, in which estimates of the latent variables are used to construct tractable lower bounds to the likelihood. We use Lagrangian relaxation to derive alternative bounds on the likelihood, that have advantage of being able to be optimized over a convex parametrization of all stable linear models, using standard techniques such as semidefinite programming (SDP).

In this paper, we treat both the latent states and latent disturbances formulation of EM, leading to two algorithms: EM with latent States & Lagrangian relaxation (EMSL), and EM with latent Disturbances & Lagrangian relaxation (EMDL). The former represents the *de facto* choice of latent variables; however, we show that the latter can lead to higher fidelity bounds on the likelihood, when the effect of measurement noise is more significant than that of the disturbances. We also show that latent disturbances lead to the most broadly applicable formulation of EM for identification of singular state space models.

We first introduced the basic idea of combining Lagrangian relaxation with a formulation of EM over latent disturbances in our conference paper (Umenberger, Wågberg, Manchester, & Schön, 2015). This paper extends that work in several significant ways. Foremost, we now incorporate model stability constraints into the more common latent states formulation, cf. Section 4.1, as well as the latent disturbances case, cf. Section 4.2. We also extend the proposed method to handle correlated disturbances and measurement noise, cf. Section 4.3. In Section 4.2 we apply Lagrangian relaxation without resorting to Monte Carlo approximations, unlike (Umenberger et al., 2015). Furthermore, the Lagrangian relaxation detailed in this paper makes use of a more effective multiplier, which improves fidelity of the bound. Finally, a new study of the behavior of the EM algorithm for large and small disturbances is presented in Section 5.2 and Section 5.3, offering insights to guide the practitioner as to the best choice of latent variables for a given problem.

2. Preliminaries

2.1. Notation

The cone of real, symmetric nonnegative (positive) definite matrices is denoted by $\mathbb{S}^n_+$ ($\mathbb{S}^n_{++}$). The $n \times n$ identity matrix is denoted I_n. Let vec : $\mathbb{R}^{m\times n} \mapsto \mathbb{R}^{mn}$ denote the function that stacks the columns of a matrix to produce a column vector. The Kronecker product is denoted $\otimes$. The transpose of a matrix A is denoted A'. For a vector a, $|a|^2_Q$ is shorthand for $a'Qa$. Time series data $\{x_t\}_{t=a}^b$ is denoted $x_{a:b}$ where $a, b \in \mathbb{N}$. A random variable x distributed according to the multivariate normal distribution, with mean μ and covariance Σ, is denoted $x \sim \mathcal{N}(\mu, \Sigma)$. We use $a(\theta) \propto b(\theta)$ to mean $b(\theta) = c_1 a(\theta) + c_2$ where c_1, c_2 are constants that do not affect the minimizing value of θ when optimizing $a(\theta)$. The log likelihood function is denoted $L_\theta(y_{1:T}) \triangleq \log p_\theta(u_{1:T}, y_{1:T})$. The spectral radius (magnitude of largest eigenvalue) of a matrix A is $r_{\mathrm{sp}}(A)$.

2.2. The minorization–maximization principle

The minorization–maximization (MM) principle (Hunter & Lange, 2004; Ortega & Rheinboldt, 1970) is an iterative approach to optimization problems of the form $\max_\theta f(\theta)$. Given an objective function $f(\theta)$ (not necessarily a likelihood), at each iteration of an MM algorithm we first build a *tight* lower bound $b(\theta, \theta_k)$ satisfying

$$f(\theta) \geq b(\theta, \theta_k)\ \forall\, \theta \text{ and } f(\theta_k) = b(\theta_k, \theta_k),$$

i.e. we *minorize* f by b. Then we optimize $b(\theta, \theta_k)$ w.r.t. θ to obtain θ_{k+1} such that $f(\theta_{k+1}) \geq f(\theta_k)$. The principle is useful when direct