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 (e-mail: {j.umenberger, i.manchester}@acfr.usyd.edu.au).*

[b]*Department of Information Technology, Uppsala University, Sweden, (e-mail: {johan.wagberg, thomas.schon}@it.uu.se}).*

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.

Keywords: System identification, expectation maximization, Lagrangian relaxation, convex optimization.

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 [14]. 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., [14, 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 [14, 15] (PEM) and the Metropolis-Hastings algorithm [19, 8].

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 [35, 12], and the Expectation Maximization (EM) algorithm [3, 25, 7, 26]. The augmentation together with appropriate priors also allows for closed form expressions in a Gibbs sampler [6, 36], (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 [18, 30]. An important technique employed in this approach is a type of Lagrangian relaxation, similar to a method widely applied in combinatorial optimization [13] and robust control, where it is referred to as the S-procedure [38, 22].

The problem of model stability has also seen considerable attention over the years. In subspace identification, a number of strategies have been proposed: [16] showed that stability can be guaranteed by augmenting the extended observability matrix with rows of zeros; in [34], regularization was used to constrain the spectral radius of the identified A to a user-specified value; [10] 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 [11] introduced an LMI parametrization of all stable models, $\mathcal{S}$; this approach was generalized in [20] 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. [16] is known to bias the estimated model, and even unconstrained subspace methods are generally considered to be less accurate than PEM [5]. Furthermore, [11] replaces the least-squares objective with a weighted projection which, as noted in [27] can produce substantial distortions.

As a middle ground between the conservatism of [10] and the distortions of [11], the authors of [27] proposed a constraint generation approach; c.f. also [1]. 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 [31, 30, 32]. 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 parameterization 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 parameterization 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 [33]. This paper extends that work in several significant ways. Foremost, we now incorporate model stability constraints into the more common latent states formulation, c.f. §4.1, as well as the latent disturbances case, c.f. §4.2. We also extend the proposed method to handle correlated disturbances and measurement noise, c.f. §4.3. In §4.2 we apply Lagrangian relaxation without resorting to Monte Carlo approximations, unlike [33]. 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 §5.2 and §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 $\{x_t\}^b_{t=a}$ 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 $r_{\text{sp}}(A)$.

2.2. The minorization-maximization principle

The minorization-maximization (MM) principle [21, 9] 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 optimization of f is challenging, but optimization of b is tractable (e.g. concave). In the following two subsections, we present EM and Lagrangian relaxation as special cases of the MM principle, for problems involving missing data. Each of these algorithms is predicated on the assumption that there exists latent variables, z, such that optimization of $f(\theta)$ would be more straightforward if z were known.

2.3. The Expectation Maximization algorithm

The EM algorithm [3] applies the MM principle to ML estimation, i.e., $f(\theta) = \log p_\theta(u_{1:T}, y_{1:T}) \triangleq L_\theta(y_{1:T})$. Each iteration of the algorithm consists of two steps: the expectation (E) step computes the *auxiliary* function

$$Q(\theta, \theta_k) \triangleq \int L_\theta(y_{1:T}, Z) p_{\theta_k}(Z|y_{1:T})\, dZ \tag{4a}$$

$$= \mathrm{E}_{\theta_k}\left[L_\theta(y_{1:T}, Z)|y_{1:T}\right], \tag{4b}$$

which is then maximized in lieu of the likelihood function during the maximization (M) step. The auxiliary function can be shown to satisfy the following inequality

$$L_\theta(y_{1:T}) - L_{\theta_k}(y_{1:T}) \geq Q(\theta, \theta_k) - Q(\theta_k, \theta_k) \tag{5}$$

and so the new parameter estimate θ_{k+1} obtained by maximization of $Q(\theta, \theta_k)$ is guaranteed to be of equal or greater likelihood than θ_k. In this sense, EM may be thought of as a specific MM recipe for building lower bounds $Q(\theta, \theta_k)$ to the objective $L_\theta(y_{1:T})$, in ML estimation problems involving latent variables.

Remark 1. Strictly speaking $Q(\theta, \theta_k)$ does not minorize $L_\theta(y_{1:T})$. Rather, the *change* in $Q(\theta, \theta_k)$ lower bounds the *change* in $L_\theta(y_{1:T})$; c.f. (5). Nevertheless, with some abuse of terminology, we will refer to $Q(\theta, \theta_k)$ as a lower bound, as shorthand for the relationship in (5).

2.4. Lagrangian relaxation

The method of Lagrangian relaxation proposed in [30] constructs convex upper bounds for constrained minimization problems of the form

$$\min_{\theta, z}\; J(\theta, z) \text{ s.t. } F(\theta, z) = 0, \tag{6}$$

i.e. $f(\theta) = \min_z J(\theta, z)$ s.t. $F(\theta, z) = 0$. Here $J(\theta, z)$ is a cost function assumed to be convex in θ for each z, and $F(\theta, z)$, assumed affine in θ, encodes the constraints between z and θ, e.g. (1a), (1b).

Unlike EM, in which we estimate z, Lagrangian relaxation *supremizes* over the latent variables to generate the bound. Specifically, the relaxation of (6) takes the form

$$\bar{J}_\lambda(\theta) = \sup_z\; J(\theta, z) - \lambda(z)' F(\theta, z), \tag{7}$$

where $\lambda(z)$ may be interpreted as a Lagrange multiplier. For arbitrary λ, the function $\bar{J}_\lambda(\theta)$ has two key properties:

1) It is convex in θ. Recall that J and F are convex and affine in θ, respectively. As such, $\bar{J}_\lambda(\theta)$ is the supremum of an infinite family of convex functions, and is therefore convex in θ; see §3.2.3 of [2].

2) It is an upper bound for the original problem (6). Given θ, let z^* be any z such that $F(\theta, z^*) = 0$. Then

$$J(\theta, z^*) + \lambda F(\theta, z^*) = J(\theta, z^*) \geq f(\theta),$$

which implies that the supremum over all z can be no smaller; i.e. $\bar{J}_\lambda(\theta)$ is an upper bound for $f(\theta)$.

The original optimization problem (6) may then be approximated by the convex program $\min_\theta\; \bar{J}_\lambda(\theta)$.

In this paper, we will show how to construct $\lambda(z)$ so that the convex bound is tight at a point θ_k, thus allowing us to use Lagrangian relaxation in the MM framework.

3. EM for linear dynamical systems

In the application of EM to the identification of dynamical systems, there are two natural choices of latent variables: systems states, $x_{1:T}$, and initial conditions and disturbances $\{x_1, w_{1:T}\}$. In this section, we recap the latent states case, detail the latent disturbances formulation, and elucidate the key differences between the two.

3.1. EM with latent states

Latent states are the *de facto* choice of latent variables in the identification of dynamical systems. Consequently, this formulation has been studied extensively, c.f. [7]. Here we recap the essential details, to pave the way for the introduction of stability guarantees in §4.1. Choosing latent states yields a joint likelihood function of the form

$$p_\theta(y_{1:T}, x_{1:T}) = \left[\prod_{t=1}^{T} p_\theta(y_t|x_t)\right]\left[\prod_{t=1}^{T-1} p_\theta(x_{t+1}|x_t)\right] p_\theta(x_1). \tag{8}$$

The E step computes the *auxiliary* function,

$$Q^{\mathrm{s}}(\theta, \theta_k) = \mathrm{E}_{\theta_k}\left[\log p_\theta(y_{1:T}, x_{1:T})|\, y_{1:T}\right] \tag{9}$$

which decomposes as

$$Q^{\mathrm{s}}(\theta, \theta_k) = \underbrace{\mathrm{E}_{\theta_k}\left[\log p_\theta(x_1)|\, y_{1:T}\right]}_{\propto -Q^{\mathrm{s}}_1(\theta,\theta_k)} + \tag{10}$$

$$\underbrace{\sum_{t=1}^{T} \mathrm{E}_{\theta_k}\left[\log p_\theta(y_t|x_t)|\, y_{1:T}\right]}_{\propto -Q^{\mathrm{s}}_2(\theta,\theta_k)} + \underbrace{\sum_{t=1}^{T} \mathrm{E}_{\theta_k}\left[\log p_\theta(x_{t+1}|x_t)|\, y_{1:T}\right]}_{\propto -Q^{\mathrm{s}}_3(\theta,\theta_k)}$$

Notice that $-Q^{\mathrm{s}} \propto Q^{\mathrm{s}}_1 + Q^{\mathrm{s}}_2 + Q^{\mathrm{s}}_3$. It is more convenient to discuss maximization of Q^{s} in terms of minimization of

$\sum_{i=1}^{3} Q_i^s$. As $-Q^{\mathrm{s}}$ is convex in θ, minimization is straightforward and reduces to linear least squares; c.f [7, Lemma 3.3]. Global minimizers of Q_1^{s}, Q_2^{s} and Q_3^{s} are given by

$$\mu = \hat{x}_{1|T}, \qquad \Sigma_1 = \widehat{\Sigma}_{1|T}, \tag{11a}$$

$$[C\ D] = \Phi_{yz}\Phi_{zz}^{-1}, \quad \Sigma_v = \Phi_{yy} - \Phi_{yz}\Phi_{zz}^{-1}\Phi_{yz}, \tag{11b}$$

$$[A\ B] = \Phi_{xz}\Phi_{zz}^{-1}, \quad \Sigma_w = \Phi_{xx} - \Phi_{xz}\Phi_{zz}^{-1}\Phi_{xz}, \tag{11c}$$

resp., where $z_t = [x_t', u_t']'$, $\Phi_{yy} = \frac{1}{T}\sum_{t=1}^{T} y_t y_t'$, and

$$\hat{x}_{1|T} = \mathrm{E}_{\theta_k}\left[x_1|\, y_{1:T}\right], \qquad \widehat{\Sigma}_{1|T} = \mathrm{Var}_{\theta_k}\left[x_1|y_{1:T}\right], \tag{12}$$

$$\Phi_{yz} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}_{\theta_k}\left[y_t z_t'|\, y_{1:T}\right], \ \Phi_{xz} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}_{\theta_k}\left[x_{t+1} z_t'|\, y_{1:T}\right]$$

$$\Phi_{zz} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}_{\theta_k}\left[z_t z_t'|\, y_{1:T}\right], \ \Phi_{xx} = \frac{1}{T}\sum_{t=2}^{T+1}\mathrm{E}_{\theta_k}\left[x_t x_t'|\, y_{1:T}\right].$$

The quantities in (12) can be computed by the RTS smoother [23]; c.f. [7, Lemma 3.2] for details. A numerically robust square-root implementation of the smoothing algorithm should be used for accuracy, e.g. [7, §4].

3.2. EM with latent disturbances

In the latent disturbances formulation of EM, it is convenient to work with the more general parametrization

$$x_{t+1} = Ax_t + Bu_t + Gw_t, \tag{13}$$

of LGSS model dynamics. This permits identification of singular state-space models, in which $n_w < n_x$, as discussed in §5.1. When using latent disturbances, we set $\Sigma_w = I$ and $\theta = \{\mu, \Sigma_1, \Sigma_v, A, B, C, D, G\}$. To avoid cumbersome notation, we use the same variable θ to group parameters in both the latent states and disturbances formulations; the contents of θ can be easily inferred from the context. Choosing latent disturbances yields a joint likelihood function of the form

$$p_\theta(y_{1:T}, x_1, w_{1:T}) = \left[\prod_{t=1}^{T} p_\theta(y_t|w_{1:t-1}, x_1)\right] p_\theta(w_{1:T})p_\theta(x_1). \tag{14}$$

Analogously to (10), the auxiliary function

$$Q^{\mathrm{d}}(\theta, \theta_k) = \mathrm{E}_{\theta_k}\left[\log p_\theta(y_{1:T}, x_1, w_{1:T})|y_{1:T}\right] \tag{15}$$

conveniently decomposes as

$$Q^{\mathrm{d}}(\theta,\theta_k) = \underbrace{\mathrm{E}_{\theta_k}\left[\log p_\theta(x_1)|y_{1:T}\right]}_{\propto -Q_1^{\mathrm{d}}(\theta,\theta_k)} + \underbrace{\mathrm{E}_{\theta_k}\left[\log p_\theta(w_{1:T})|y_{1:T}\right]}_{\propto -Q_2^{\mathrm{d}}(\theta_k)} + \underbrace{\mathrm{E}_{\theta_k}\left[\log p_\theta(y_{1:T}|x_1, w_{1:T})|y_{1:T}\right]}_{\propto -Q_3^{\mathrm{d}}(\theta,\theta_k)}. \tag{16}$$

The following lemma details the computation of $Q^{\mathrm{d}}(\theta, \theta_k)$. For clarity of exposition, we introduce the following *lifted* form of the dynamics in (13),

$$Y = \bar{C}\bar{H}Z + (\bar{C}\bar{N} + \bar{D})U + V,$$

where $Y = \mathrm{vec}(y_{1:T})$, $U = \mathrm{vec}(u_{1:T})$, $V = \mathrm{vec}(v_{1:T})$, $Z = \mathrm{vec}([x_1, w_{1:T-1}])$,

$$\bar{H} = \begin{bmatrix} I & 0 & 0 & 0 & \dots & 0 \\ A & G & 0 & 0 & \dots & 0 \\ A^2 & A & G & 0 & \dots & 0 \\ \vdots & & & & \ddots & \vdots \\ A^{T-1} & A^{T-2}G & A^{T-3}G & & \dots & G \end{bmatrix},$$

$$\bar{N} = \begin{bmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ B & 0 & 0 & 0 & \dots & 0 \\ AB & B & 0 & 0 & \dots & 0 \\ \vdots & & & & \ddots & \vdots \\ A^{T-2}B & A^{T-3}B & \dots & AB & B & 0 \end{bmatrix},$$

$$\bar{C} = I_T \otimes C \text{ and } \bar{D} = I_T \otimes D. \tag{17}$$

Lemma 1. *The auxiliary function $Q^{\mathrm{d}}(\theta, \theta_k)$ defined in (15) is given by*

$$-Q^{\mathrm{d}}(\theta,\theta_k) \propto \log\det\Sigma_1 + |\hat{x}_{1|T} - \mu|^2_{\Sigma_1^{-1}} + \mathrm{tr}\big(\Sigma_1^{-1}\widehat{\Sigma}_{1|T}\big) + T\log\det\Sigma_v + \mathrm{tr}(\Sigma_Y^{-1}(\bar{C}\bar{H}\Omega\bar{H}'\bar{C}' + \hat{\Delta}\hat{\Delta}')),$$

where $\hat{x}_{1|T} = \mathrm{E}_{\theta_k}\left[x_1|y_{1:T}\right]$ and $\widehat{\Sigma}_{1|T} = \mathrm{Var}_{\theta_k}\left[x_1|y_{1:T}\right]$ as in (12), and

$$\hat{Z} = \mathrm{E}_{\theta_k}\left[Z|y_{1:T}\right], \tag{18a}$$

$$\Omega = \mathrm{Var}_{\theta_k}\left[Z|y_{1:T}\right], \tag{18b}$$

$$\mu_Y \triangleq \mathrm{E}_\theta\big[Y|Z\big] = \bar{C}\bar{H}Z + (\bar{C}\bar{N} + \bar{D})U, \tag{19a}$$

$$\Sigma_Y \triangleq \mathrm{Var}_\theta\big[Y|Z\big] = I_T \otimes \Sigma_v, \tag{19b}$$

$$\hat{\Delta} = \mathrm{E}_{\theta_k}\big[Y - \mu_Y|y_{1:T}\big] = Y - \bar{C}\bar{H}\hat{Z} - (\bar{C}\bar{N} + \bar{D})U. \tag{20}$$

Proof. Refer to Appendix A.1.

For the LGSS models considered in this work, $\hat{Z}$ and Ω can be computed in closed form by standard *disturbance* smoothers; see, e.g., [4, §4.5]. Once again, it is prudent to use square-root implementations of these smoothing algorithms, given in [4, §6.3], for numerical robustness (i.e. nonnegative definiteness of covariances).

We now turn our attention to the M step, i.e. minimization of $-Q^{\mathrm{d}} \propto Q_1^{\mathrm{d}} + Q_2^{\mathrm{d}} + Q_3^{\mathrm{d}}$. It is clear that Q_1^{d} and Q_1^{s} take the same form, so $\mu = \hat{x}_{1|T}$ and $\Sigma_1 = \widehat{\Sigma}_{1|T}$ globally minimize $Q_1^{\mathrm{d}}(\theta, \theta_k)$, as in (11a). Minimization of $Q_2^{\mathrm{d}}(\theta_k)$ is unnecessary, as it is constant w.r.t. θ. Minimization of Q_3^{d}, however, is a more challenging problem. Indeed, from Lemma 1, it is clear that the quantities $\bar{H}$ and $\bar{N}$ render $Q_3^{\mathrm{d}}(\theta, \theta_k)$ a nonconvex function of the model parameters.

To summarize, the computations involved in each iteration of the latent disturbances formulation of EM are straightforward, with the exception of minimization of $Q_3^{\mathrm{d}}(\theta, \theta_k)$, which is nonconvex.

A common heuristic for terminating the EM algorithm is to cease iterations once the change in likelihood falls below a certain tolerance δ, i.e.

$$L_{\theta_{k+1}}(y_{1:T}) - L_{\theta_k}(y_{1:T}) < \delta. \tag{21}$$

Alternatively, one can simply run the algorithm for a finite number of iterations, chosen so as to attain a model of sufficient quality; this is the approach taken, e.g., in [7, 37].

4. Convex M step with guaranteed model stability

In this section we incorporate model stability constraints into the EM framework. The set $\mathcal{S}$ of stable A matrices is nonconvex; however, by using Lagrangian relaxation we build convex bounds on $-Q(\theta,\theta_k)$, which we optimize over a convex parametrization of all stable linear models.

4.1. Ensuring stability with latent states

We begin with the latent states formulation. The global minimizers (11a) and (11b) of Q_1^{s} and Q_2^{s}, resp., remain unchanged, as they do not influence stability. To optimize Q_3^{s}, it is convenient to work with the representation:

$$Q_3^{\mathrm{s}}(\theta,\theta_k) = \sum_{t=1}^{M^{\mathrm{s}}} |\tilde{x}_{t+1} - A\tilde{x}_t - B\tilde{u}_t|^2_{\Sigma_w^{-1}} + T\log\det\Sigma_w \quad (22)$$

where $M^{\mathrm{s}} = 2n_x + n_u$, and $\tilde{x}$ and $\tilde{u}$ satisfy

$$\sum_{t=1}^{M^{\mathrm{s}}} \begin{bmatrix} \tilde{x}_{t+1} \\ \tilde{x}_t \\ \tilde{u}_t \end{bmatrix} \begin{bmatrix} \tilde{x}_{t+1} \\ \tilde{x}_t \\ \tilde{u}_t \end{bmatrix}' = T \begin{bmatrix} \Phi_{xx} & \Phi_{xz}' \\ \Phi_{xz} & \Phi_{zz} \end{bmatrix} = \Phi^{\mathrm{s}}. \quad (23)$$

Our goal is to minimize Q_3^{s} subject to model stability constraints. The main challenge is nonconvexity of the set $\mathcal{S}$ of (Schur) stable matrices, i.e., the Lyapunov condition $A'PA - P < 0$ is not jointly convex in A and $P \in \mathbb{S}^{n_x}_{++}$. One can circumvent this difficulty by introducing an equivalent *implicit* representation of the dynamics, e.g.,

$$Ex_{t+1} = Fx_t + Ku_t. \quad (24)$$

In what follows, let $\theta^s = \{E, F, K, \Sigma_w\}$.

Lemma 2. *A matrix $A \in \mathbb{R}^{n_x\times n_x}$ is Schur stable iff there exists $E \in \mathbb{R}^{n_x\times n_x}$ and $P \in \mathbb{S}^{n_x}_{++}$ such that the LMI*

$$S^{\mathrm{s}}(\theta^s) \triangleq \begin{bmatrix} E + E' - P - I & F' \\ F & P \end{bmatrix} \geq 0 \quad (25)$$

holds with $F = EA$.

Proof. This result is a trivial modification of Lemma 4 and Corollary 5 in [17, Section 3.2].

By Lemma 2, $\Theta^s = \{\theta^s : \exists P \in \mathbb{S}^{n_x}_{++},\ S^{\mathrm{s}}(\theta^s) \geq 0\}$ defines a convex parametrization of all stable linear systems. Note also that (25) implies $E + E' > 0$, which ensures that the implicit dynamics in (24) are well-posed, i.e. $A = E^{-1}F$.

The challenge now becomes the optimization of Q_3^{s} with models in the implicit form (24). Simply solving

$$\min_{\theta^s\in\Theta^s} \sum_{t=1}^{M^{\mathrm{s}}} |E\tilde{x}_{t+1} - F\tilde{x}_t - K\tilde{u}_t|^2_{\Sigma_w^{-1}} + T\log\det\Sigma_w$$

is insufficient, as there is no guarantee that this will reduce Q_3^{s}. We proceed by using Lagrangian relaxation to build a convex upper bound on Q_3^{s}. For clarity of exposition, let us temporarily ignore the $\log\det\Sigma_w$ term, as well as the summation, in (22) and consider the nonconvex problem

$$\min_{A,B,\Sigma_w} \ |\tilde{x}_{t+1} - A\tilde{x}_t - B\tilde{u}_t|^2_{\Sigma_w^{-1}} \quad \text{s.t. } A \in \mathcal{S} \quad (26)$$

for some t. Problem (26) is completely equivalent to

$$\min_{x_{t+1},\theta^s\in\Theta^s} |\tilde{x}_{t+1} - x_{t+1}|^2_{\Sigma_w^{-1}} \text{ s.t. } Ex_{t+1} = F\tilde{x}_t + K\tilde{u}_t \quad (27)$$

as both (26) and (27) have the same objective and feasible set. Introducing $\Delta = \tilde{x}_{t+1} - x_{t+1}$ and $\epsilon_t = E\tilde{x}_{t+1} - F\tilde{x}_t - K\tilde{u}_t$, the Lagrangian relaxation of (27) is given by

$$\bar{J}^s_\lambda(\theta^s, t) = \sup_\Delta \ |\Delta|^2_{\Sigma_w^{-1}} - \lambda(\Delta)'\left(E\Delta - \epsilon_t\right) \quad (28)$$

for some multiplier $\lambda(\Delta)$, c.f. Section 2.4. As $\bar{J}^s_\lambda(\theta^s,t)$ upper bounds (26), we can construct a convex upper bound for Q_3^{s} by combining $\sum_t \bar{J}^s_\lambda(\theta^s,t)$ with a linear bound on the concave $\log\det\Sigma_w$ term.[1]

Lemma 3. *Consider the function*

$$\bar{Q}_3^{\mathrm{s}}(\theta^s) \triangleq \sum_{t=1}^{M^{\mathrm{s}}} \bar{J}^s_\lambda(\theta^s,t) + T\mathrm{tr}\left(\Sigma_{w_k}^{-1}\Sigma_w\right), \quad (29)$$

where Σ_{w_k} is the estimate of Σ_w stored in θ_k. $\bar{Q}_3^{\mathrm{s}}(\theta^s)$ is a convex upper bound on $Q_3^{\mathrm{s}}(\theta,\theta_k)$.

The function $\bar{Q}_3^{\mathrm{s}}$ is a convex upper bound on Q_3^{s} for any multiplier, λ. However, to be suitable for EM, i.e. for (5) to hold, we require $\bar{Q}_3^{\mathrm{s}}$ to be tight at θ_k, i.e. $\bar{Q}_3^{\mathrm{s}}(\theta_k^s) = Q_3^{\mathrm{s}}(\theta_k,\theta_k)$. The following lemma provides a choice of multiplier that ensures this property.

Lemma 4. *For each $\bar{J}^s_\lambda(\theta^s_k,t)$ in (29), $t = 1,\ldots,M^{\mathrm{s}}$, let $\lambda(\Delta) = 2H\Delta$ where $H = (P')^{-1}\Sigma_w^{-1}$ is such that $\theta^s_k = \{P, PA_k, PB_k, \Sigma_{w_k}\} \in \Theta^s$. Then $\bar{Q}_3^{\mathrm{s}}(\theta^s_k) = Q_3^{\mathrm{s}}(\theta_k,\theta_k)$, i.e. the bound is tight at θ_k.*

Proof. Refer to Appendix B.1.

Before leaving the latent states case we note that the upper bound $\bar{Q}_3^{\mathrm{s}}$ can be optimized as the following SDP:

$$\min_{R,\theta^s\in\Theta^s} \ \mathrm{tr}\left(R\Phi^{\mathrm{s}}\right) + T\mathrm{tr}\left(\Sigma_{w_k}^{-1}\Sigma_w\right) \quad (30)$$
$$\text{s.t. } \begin{bmatrix} R & E_K^{F\prime}H & 0 \\ H'E_K^F & H'E + E'H & I \\ 0 & I & \Sigma_w \end{bmatrix} \geq 0,$$

where $R \in \mathbb{S}^{2n_x+n_u}$ is a slack variable, $E_K^F = [E, -F, -K]$, and Φ^{s} is the empirical covariance matrix in (23). A complete summary of the approach is given in Algorithm 1.

[1] The linear bound on $\log\det\Sigma_w$ is $\mathrm{tr}\left(\Sigma_{w_k}^{-1}\Sigma_w\right) + \log\det\Sigma_{w_k} + n_x$, but we exclude the constant terms from (29) for brevity.

Remark 2. To ensure model stability, it is only necessary to solve (30) if the spectral radius of $A_{\rm ls}$ is too large, where $A_{\rm ls}$ is the least squares solution from (11c), i.e., if $r_{\rm sp}(A_{\rm ls}) > 1-\delta$ for some-user selected $\delta > 0$, c.f. (3.2) of Algorithm 1.

Algorithm 1 EM with latent States and Lagrangian relaxation (EMSL)

---

1. Set $k = 0$ and initialize θ_k such that $A_k \in \mathcal{S}$ and $L_{\theta_k}(y_{1:T})$ is finite; c.f. §3.
2. **Expectation (E) Step:** compute (12).
3. **Maximization (M) Step:**
   (3.1) Update θ_{k+1} with least squares, as in (11).
   (3.2) If $r_{\rm sp}(A_{k+1}) > 1-\delta$ for user-chosen $\delta > 0$, solve $\theta^s_{k+1} = \arg\min_{\theta^s\in\Theta^s} \bar{Q}^{\rm s}_3(\theta^s)$, and update θ_{k+1} with $A_{k+1} = E^{-1}_{k+1}F_{k+1}$, $B_{k+1} = E^{-1}_{k+1}K_{k+1}$, and $\Sigma_{w_{k+1}}$
4. Evaluate termination criteria, e.g. (21). If `false`, $k \leftarrow k+1$ and return to step 2.

4.2. Convex bounds with stability guarantees for latent disturbances

We now turn our attention to the latent disturbances formulation. The developments in this section follow the same pattern as §4.1, however, the computations are more involved. As in the latent states case, the global minimizer (11a) of $Q^{\rm d}_1$ remains unchanged, and so we concentrate on optimization of $Q^{\rm d}_3$ subject to a model stability constraint, $A \in \mathcal{S}$. It is convenient to conceptualize $Q^{\rm d}_3(\theta, \theta_k)$ in terms of *simulation error*, defined as

$$\mathcal{E}(\theta, U, Y, Z) \triangleq \sum_{t=1}^{T} |y_t - Cx_t - Du_t|^2_{\Sigma_v^{-1}}, \tag{31}$$

where $\mathrm{vec}(x_{1:T}) = \bar{N}U + \bar{H}Z$, i.e., the simulated states.

Lemma 5. *With simulation error defined as in (31), $Q^{\rm d}_3(\theta, \theta_k)$ in (16) is equivalent to:*

$$Q^{\rm d}_3(\theta, \theta_k) = \mathcal{E}(\theta, U, Y, \hat{Z}) + \sum_{j=1}^{M^{\rm d}} \mathcal{E}(\theta, 0, 0, Z^j) + T \log\det \Sigma_v \tag{32}$$

where $M^{\rm d} = n_x + (T-1)n_w$, and $Z^j \in \mathbb{R}^{n_x+(T-1)n_w}$ are such that $\sum_{j=1}^{M^{\rm d}} ZZ^{j'} = \Omega$.

Proof. Refer to Appendix B.2.

Our task $\min_\theta Q^{\rm d}_3$ s.t. $A \in \mathcal{S}$ is challenging due to non-convexity of both the objective and feasible set. As in §4.1, we circumvent the latter by introducing an implicit representation of the dynamics in (13),

$$Ex_{t+1} = Fx_t + Ku_t + Lw_t. \tag{33}$$

Setting $\theta^d = \{E, F, K, L, \Sigma_v\}$ we can define the convex set of stable models $\Theta^d = \{\theta^d : \exists P \in \mathbb{S}^{n_x}_{++},\ S^{\rm d}(\theta^d) \geq 0\}$, with

$$S^{\rm d}(\theta^s) \triangleq \begin{bmatrix} E + E' - P - \delta I & F' & C' \\ F & P & 0 \\ C & 0 & \Sigma_w \end{bmatrix} \geq 0, \tag{34}$$

for $\delta > 0$. We use the LMI (34), instead of (25), to ensure finiteness of the supremum in (37), c.f. Appendix B.4.

To optimize $Q^{\rm d}_3$ with models in the implicit form (33), we use Lagrangian relaxation to build a convex upper bound on $Q^{\rm d}_3$. For clarity of exposition, let us temporarily ignore the $\log\det\Sigma_v$ term, and summation, in (32) and concentrate on minimization of simulation error

$$\min_\theta \ \mathcal{E}(\theta, U, Y, Z) \text{ s.t. } A \in \mathcal{S}. \tag{35}$$

Problem (35) is completely equivalent to

$$\min_{\Delta, \theta^d\in\Theta^d} \ |Y - \bar{C}\Delta - \bar{D}U|^2_{\Sigma_Y^{-1}} \text{ s.t. } \bar{E}\Delta = \bar{\epsilon}, \tag{36}$$

as both problems have the same objective and feasible set. In (36), $\Delta \in \mathbb{R}^{Tn_x}$ denotes the states $x_{1:T}$ that we optimize over, $\bar{E} \in \mathbb{R}^{Tn_x\times Tn_x}$ and $\bar{\epsilon} \in \mathbb{R}^{Tn_x}$ are given by

$$\begin{bmatrix} E & 0 & \dots & & \\ -F & E & 0 & & \\ 0 & -F & E & 0 & \\ \vdots & & & \ddots & \ddots \end{bmatrix} \quad \& \quad \begin{bmatrix} Ex_1 \\ Ku_1 + Lw_1 \\ \vdots \\ Ku_{T-1} + Lw_{T-1} \end{bmatrix}$$

resp., and $\bar{C}$, $\bar{D}$, Σ_Y are defined in (17). The Lagrangian relaxation of (36) is given by

$$\bar{J}^d_\lambda(\theta^d, U, Y, Z) = \sup_\Delta \ |Y - \bar{C}\Delta - \bar{D}U|^2_{\Sigma_Y^{-1}} - \lambda(\Delta)'\left(\bar{E}\Delta - \bar{\epsilon}\right) \tag{37}$$

for some multiplier $\lambda(\Delta) \in \mathbb{R}^{Tn_x}$, c.f. §2.4. As $\bar{J}^d_\lambda$ upper bounds (26), we can construct a convex upper bound for $Q^{\rm d}_3$ by replacing each simulation error term in (32) with the appropriate bound:

Lemma 6. *Consider the following function*

$$\bar{Q}^{\rm d}_3(\theta^d) \triangleq \bar{J}_{\lambda^0}(\theta^d, U, Y, \hat{Z}) + \sum_{j=1}^{M^{\rm d}} \bar{J}_{\lambda^j}(\theta^d, 0, 0, Z^j) + T\mathrm{tr}\left(\Sigma^{-1}_{v_k}\Sigma_v\right), \tag{38}$$

where $\sum_{j=1}^{M^{\rm d}} Z^jZ^{j'} = \Omega$ as in Lemma 5, and Σ_{v_k} is the estimate of Σ_v stored in θ_k. $\bar{Q}^{\rm d}_3(\theta^d)$ is a convex upper bound for $Q^{\rm d}_3(\theta, \theta_k)$.

Proof. Refer to Appendix B.3.

Notice, from (38), that $\bar{Q}_3^{\mathrm{d}}$ depends on $M^{\mathrm{d}}+1$ multipliers, $\{\lambda^j\}_{j=0}^{M^{\mathrm{d}}}$, unlike $\bar{Q}_3^{\mathrm{s}}$. Although $\bar{Q}_3^{\mathrm{d}}$ upper bounds Q_3^{d} for any choice of λ, as in §4.1 we require $\bar{Q}_3^{\mathrm{d}}$ to be a tight bound such that (5) holds, i.e., we need $\bar{Q}_3^{\mathrm{d}}(\theta_k^d) = Q_3^{\mathrm{d}}(\theta_k, \theta_k)$. To obtain such a set of multipliers $\{\lambda^j\}_{j=0}^{M^{\mathrm{d}}}$, we propose the following two-stage approach. At the k^{th} iteration,

i. For each of the $j = 0, \ldots, M^{\mathrm{d}}$ bounds $\bar{J}_{\lambda^j}(\theta^d)$ that comprise $\bar{Q}_3^{\mathrm{d}}(\theta^d)$, c.f. (38), solve the convex program

$$E_k^j = \arg\min_E \quad \bar{J}_{\lambda_\Delta^j}(\theta^d) \quad \text{s.t.} \quad \theta^d = \{E, EA_k, EB_k, EG_k, \Sigma_{v_k}\} \in \Theta^d$$

where $\lambda_\Delta^j = 2\Delta$.

ii. Set $\lambda^j = 2\left(\Delta + h^j\right)$ with $h^j \in \mathbb{R}^{Tn_x}$ given by

$$h^j = (\bar{E}_k^{j'})^{-1}\left(\Psi_k^j (\bar{E}_k^j)^{-1}\bar{\epsilon}_k^j + \bar{\epsilon}_k^j - \bar{C}_k' \bar{\Sigma}_{Y,k}^{-1}(Y - \bar{D}_k U)\right) \tag{39}$$

where $\Psi_k^j = \bar{C}_k'\bar{\Sigma}_{Y,k}^{-1}\bar{C}_k - \bar{E}_k^j - \bar{E}_k^{j'}$. Here $\bar{E}_k^j$ and $\bar{\epsilon}_k^j$ denote $\bar{E}$ and $\bar{\epsilon}$, resp., built with $E = E_k^j$, $F = E_k^j A_k$, $K = E_k^j B_k$, and $L = E_k^j G_k$. $\bar{C}_k$ and $\bar{D}_k$ denote $\bar{C}$ and $\bar{D}_k$, resp., built with $C = C_k$, $D = D_k$.

The following lemma guarantees that the multipliers generated by this two-stage procedure give a 'tight' bound:

Lemma 7. *Given $\theta_k^d \in \Theta^d$, let $\{\lambda^j\}_{j=0}^{M^{\mathrm{d}}}$ in (38) take the form $\lambda(\Delta)^j = 2(\Delta + h^j)$ with h^j given by (39). Then $\bar{Q}_3^{\mathrm{d}}(\theta_k^d) = Q_3^{\mathrm{d}}(\theta_k, \theta_k)$, i.e. the bound is tight at θ_k.*

Proof. Refer to Appendix B.4.

A complete summary of the latent disturbances approach to EM with stability constraints is given in Algorithm 2.

Remark 3. This EMDL formulation includes, as a special case, models in innovations form, c.f. [14, §4.3]. For such models, innovations replace disturbances in (1a) and the latent variables reduce to the initial state, x_1. EM in this setting was studied in [37]. The difference between [37] and our approach is the M step: in [37] $Q(\theta, \theta_k)$ is optimized directly with a quasi-Newton method; we optimize a convex upper bound on $-Q(\theta, \theta_k)$ over a convex parametrization of stable models.

4.3. Correlated disturbances and measurement noise

For clarity of exposition, we have considered models in which there is no correlation between disturbances and measurement noise. However, the methods we have presented readily extend to the correlated case, i.e.

$$\begin{bmatrix} w_t \\ v_t \end{bmatrix} \sim \mathcal{N}(0, \Sigma_{s_c}), \quad \Sigma_{s_c} = \begin{bmatrix} \Sigma_w & \Sigma_{wv} \\ \Sigma_{wv}' & \Sigma_v \end{bmatrix}. \tag{40}$$

Algorithm 2 EM with latent Disturbances and Lagrangian relaxation (EMDL)

1. Set $k = 0$ and initialize θ_k such that $A_k \in \mathcal{S}$ and $L_{\theta_k}(y_{1:T})$ is finite; c.f. §3.
2. **Expectation (E) Step:**
   (2.1) Compute $\hat{x}_{1|T}$ and $\widehat{\Sigma}_{1|T}$ as in (12).
   (2.2) Compute $\hat{Z}$ and Ω as in (18).
3. **Maximization (M) Step:**
   (3.1) Update $\{\mu, \Sigma_1\}_{k+1} = \{\hat{x}_{1|T}, \widehat{\Sigma}_{1|T}\}$.
   (3.2) Assemble $\{\lambda^j\}_{j=0}^{M^{\mathrm{d}}}$ of the form $\lambda^j = 2(\Delta + h^j)$ by computing $\{h^j\}_{j=0}^{M^{\mathrm{d}}}$ with (39).
   (3.3) Obtain $\theta_{k+1}^d = \arg\min_{\theta^d \in \Theta^d} \bar{Q}_3^{\mathrm{d}}(\theta^d)$.
   (3.4) Update θ_{k+1} with $A_{k+1} = E_{k+1}^{-1}F_{k+1}$, $B_{k+1} = E_{k+1}^{-1}K_{k+1}$, $G_{k+1} = E_{k+1}^{-1}L_{k+1}$, and $\Sigma_{v_{k+1}}$.
4. Evaluate termination criteria, e.g. (21). If `false`, $k \leftarrow k+1$ and return to step 2.

With latent states, the joint likelihood becomes

$$p_\theta(y_{1:T}, x_{1:T}) = \left[\prod_{t=1}^{T} p_\theta(y_t, x_{t+1}|x_t)\right] p_\theta(x_1),$$

with $-Q^{\mathrm{s}}(\theta, \theta_k) \propto Q_1^{\mathrm{s}}(\theta, \theta_k) + Q_{\mathrm{c}}^{\mathrm{s}}(\theta, \theta_k)$ where

$$-Q_{\mathrm{c}}^{\mathrm{s}}(\theta, \theta_k) \propto \sum_{t=1}^{M_{\mathrm{c}}^{\mathrm{s}}} \left\| \begin{bmatrix} \tilde{x}_{t+1} \\ \tilde{y}_t \end{bmatrix} - \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \tilde{x}_t \\ \tilde{u}_t \end{bmatrix} \right\|_{\Sigma_{s_c}^{-1}}^2 + T \log\det \Sigma_{s_c}. \tag{41}$$

Here $M_{\mathrm{c}}^{\mathrm{s}} = 2n_x + n_y + n_u$, and $\tilde{x}$, $\tilde{y}$, $\tilde{u}$ satisfy

$$\sum_{t=1}^{M_{\mathrm{c}}^{\mathrm{s}}} \tilde{\zeta}_t \tilde{\zeta}_t' = \sum_{t=1}^{T} \mathrm{E}_{\theta_k}\left[\zeta_t \zeta_t' | \, y_{1:T}\right],$$

where $\tilde{\zeta}_t = \left[\tilde{x}_{t+1}', \tilde{y}_t', \tilde{x}_t', \tilde{u}_t'\right]'$ and $\zeta_t = \left[x_{t+1}', y_t', x_t', u_t'\right]'$. Clearly, (41) has the same form as (22), and so the Lagrangian relaxation of §4.1 is applicable.

Similarly, in the latent disturbances formulation the joint likelihood can be factorized as

$$p_\theta(y_{1:T}, x_1, w_{1:T}) = \prod_{t=1}^{T} p_\theta(y_t, w_t | x_1, w_{1:t-1}) p(x_1),$$

with $p(y_t, w_t | x_1, w_{1:t-1})$ given by

$$\mathcal{N}\left(\begin{bmatrix} Cx_t + Du_t \\ 0 \end{bmatrix}, \Sigma_{s_d}\right), \quad \Sigma_{s_d} = \begin{bmatrix} \Sigma_v & \Sigma_{wv} \\ \Sigma_{wv}' & I \end{bmatrix},$$

where $\mathrm{vec}(x_{1:T}) = \bar{N}U + \bar{H}Z$. Introducing $y_t^c = [y_t', w_t']'$,

$C^c = [C', 0']'$, and $D^c = [D', 0']'$ we have

$$\log p_\theta(y_{1:T}, w_{1:T}|x_1) \propto \sum\nolimits_{t=1}^{T} |y_t^c - C^c x_t - D^c u_t|^2_{\Sigma_{s_d}^{-1}} + T \log\det \Sigma_{s_d}. \quad (42)$$

The resemblance to (31) is apparent. Computing the expected value of (42) leads to a quantity that takes the same form as Q_3^{d} in (32), to which the Lagrangian relaxation of §4.2 is applicable.

5. On the choice of latent variables

5.1. Singular state space models

In applications it may be known *a priori* that the dimension of the disturbance is less than that of the state variable, i.e. $n_w < n_x$. For example, consider a mechanical system in which the disturbances are forces or torques. There are typically fewer disturbance forces than state variables (as force directly affects acceleration, but not position or velocity, in continuous time dynamics), and so G is rank-deficient. As the transition density $p_\theta(x_{t+1}|x_t) = \mathcal{N}(Ax_t + Bu_t, GG')$ no longer admits a closed form representation, when GG' is singular, it is well known that the standard EM algorithms based on latent states are no longer directly applicable.

Modifications of the EM algorithm have been proposed to circumvent this difficulty. The work of [28] introduced a perturbation model with full-rank process noise covariance, and proved that the EM iterations remain well behaved when the perturbation is set to zero. However, this approach was restricted to models in which the disturbances and measurement noise are uncorrelated. A subsequent paper [29] addressed the case of correlated state and measurement noise, but only considered models in innovations form; extension to the case of models in general form was left to future work. Furthermore, this method requires the variance of the initial state (i.e. Σ_1) to be excluded from the estimated parameters, θ.

The latent disturbances formulation of EM, c.f. §3.2, provides the most general solution to the difficulties associated with singular state space models. Specifically, with latent disturbances we can handle rank deficient G, with the possibility of correlated state and measurement noise (c.f. §4.3), as well as unknown initial conditions (μ, Σ_1), for models not necessarily in innovations form, c.f. Remark 3. When using latent disturbances we work with the joint likelihood function $p_\theta(y_{1:T}, x_1, w_{1:T})$, given in (14). Comparing (14) to (8), we observe that the problematic transition density is replaced by the joint distribution of disturbances $p_\theta(w_{1:T})$ which remains well-defined in the singular case, $n_w < n_x$.

5.2. Absence of disturbances or measurement noise

In this section, we study the auxiliary function $Q(\theta, \theta_k)$ in the limit cases of $G = 0$ and $\Sigma_v = 0$, for different choices of latent variables. These results will offer insight into the behavior of the EM algorithm as a function of disturbance magnitude, which is explored in §5.3.

Proposition 8. *Consider a model of the form (1), and let θ be such that $\Sigma_w = 0$, i.e. disturbances are omitted from the model. The auxiliary function built on* latent states*, $Q^{\mathrm{s}}(\theta, \theta_k)$, is undefined when $A \neq A_k$ or $B \neq B_k$.*

PROOF. Refer to Appendix C.1.

Proposition 9. *Consider a model of the form (1), with dynamics of the form (13), and let θ be such that $G = 0$, i.e. disturbances are omitted from the model. Furthermore, suppose $\Sigma_1 = 0$; i.e. the initial conditions $x_1 = \mu$ are modeled without uncertainty. Then $L_\theta(y_{1:T}, x_1) = Q^{\mathrm{d}}(\theta, \theta_k)$ for all θ, θ_k; i.e., the auxiliary function built on* latent disturbances*, $Q^{\mathrm{d}}(\theta, \theta_k)$, reduces to the log likelihood.*

PROOF. Refer to Appendix C.2.

Proposition 10. *Consider a* first order *model of the form (1), and let θ be such that $\Sigma_v = 0$, i.e. output noise is omitted from the model. The auxiliary function built on* latent disturbances*, $Q^{\mathrm{d}}(\theta, \theta_k)$, is undefined for $\theta \neq \theta_k$, i.e. the domain of $Q(\theta, \theta_k)$ collapses to a single point, $\theta = \theta_k$.*

PROOF. Refer to Appendix C.3.

Proposition 11. *Consider a* first order *model of the form (1), and let θ be such that $\Sigma_v = 0$, i.e. output noise is omitted from the model. Let $Q^{\mathrm{s}}(\theta, \theta_k)$ denote the auxiliary function built on* latent states*, then:*

i. $Q^{\mathrm{s}}(\theta, \theta_k)$ is undefined for all θ such that $C \neq C_k$ or $D \neq D_k$.

ii. $Q^{\mathrm{s}}(\theta, \theta_k) = L_\theta(y_{1:T})$ for all θ such that $C = C_k$ and $D = D_k$.

PROOF. Refer to Appendix C.4.

5.3. Influence of disturbance magnitude on bound fidelity

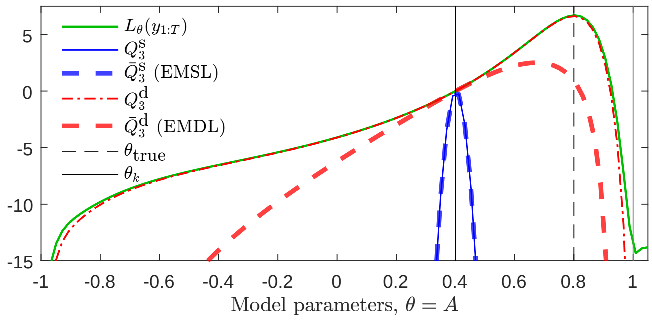
In this section, we empirically investigate the fidelity of $Q(\theta, \theta_k)$ as a bound on $L_\theta(y_{1:T})$, as a function of the magnitude of the disturbances, $w_{1:T}$, and the choice of latent variables. The results are presented in Figure 1, which depicts Q^{s}, Q^{d} and $L_\theta(y_{1:T})$ for a first order ($n_x = 1$) LGSS model, with $A = 0.7$, $B = 0.3$, $C = 0.1$, $D = 0.01$, and $GG' = \Sigma_w$. Each bound, Q^{s} and Q^{d}, is plotted as a function of the single unknown scalar parameter $\theta = A$. Note that $\mathcal{S} = \{A : -1 < A < 1\}$ is convex for $n_x = 1$; this is not true for $n_x > 1$.

We begin with the case of 'small' disturbances (i.e. $\Sigma_w \ll \Sigma_v$) as depicted in Figure 1(a), and observe the following: $Q^{\mathrm{d}}(\theta, \theta_k)$ represents $L_\theta(y_{1:T})$ with high fidelity, whereas $Q^{\mathrm{s}}(\theta, \theta_k)$ is localized about θ_k. Such an observation is not without precedent. For instance, in the latent states formulation of [25, Section 10] it was noted that an initial disturbance covariance estimate $\Sigma_w = 0$ results in $\theta_k = \theta_0$ for all k; i.e. the model parameters are not improved. Proposition 8 makes this observation more precise: in the 1D case of Figure 1(a), when $\Sigma_w = 0$, $Q^{\mathrm{s}}(\theta, \theta_k)$ is undefined for $A \neq A_k$. Taken together, Figure 1(a) and Proposition 8 suggest that as Σ_w becomes smaller (relative to Σ_v) the bound $Q^{\mathrm{s}}(\theta, \theta_k)$ becomes more localized about θ_k; the domain collapses to a single point, $\theta = \theta_k$, when $\Sigma_w = 0$. Conversely, as Σ_w (and Σ_1) decrease, $Q^{\mathrm{d}}(\theta, \theta_k)$ becomes an increasingly accurate representation of the log likelihood, eventually reproducing $L_\theta(y_{1:T})$ *exactly*, when Σ_w (and Σ_1) are identically zero, as in Proposition 9.

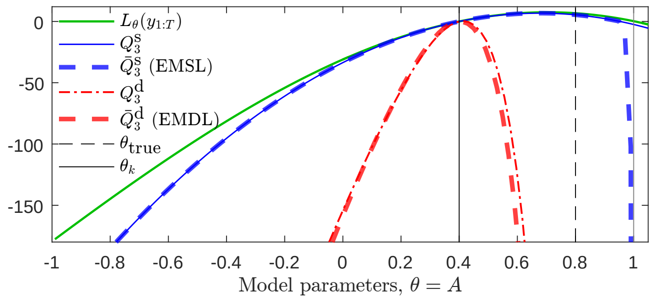
Turning our attention to the case of 'large' disturbances (i.e. $\Sigma_w \gg \Sigma_v$) as depicted in Figure 1(b), we observe the opposite behavior: $Q^{\mathrm{s}}(\theta, \theta_k)$ faithfully represents the log likelihood, whereas $Q^{\mathrm{d}}(\theta, \theta_k)$ appears to be localized about θ_k. Once more, studying the limiting case $\Sigma_v = 0$ offers insight into this behavior: Proposition 10 states that when $\Sigma_v = 0$, $Q^{\mathrm{d}}(\theta, \theta_k)$ is undefined for $A \neq A_k$. Taken together, Figure 1(b) and Proposition 10 suggest that as Σ_v decreases (i.e. as Σ_w increases relative to Σ_v), the bound $Q^{\mathrm{d}}(\theta, \theta_k)$ becomes more localized about θ_k; the domain collapses to a single point, $\theta = \theta_k$, when $\Sigma_v = 0$. Conversely, for this 1D experiment with $\theta = A$, Proposition 11 states that $Q^{\mathrm{s}}(\theta, \theta_k)$ will reproduce $L_\theta(y_{1:T})$ *exactly*, when Σ_v is identically zero. Indeed, in Figure 1(b) with $\Sigma_v \ll \Sigma_w$, we observe $Q^{\mathrm{s}}(\theta, \theta_k)$ representing the likelihood faithfully.

To summarize: in the case of 'small disturbances' (i.e. $\Sigma_w \ll \Sigma_v$), $Q^{\mathrm{d}}(\theta, \theta_k)$ will tend to bound $L_\theta(y_{1:T})$ with greater fidelity, compared to $Q^{\mathrm{s}}(\theta, \theta_k)$. In the case of 'large disturbances' (i.e. $\Sigma_w \gg \Sigma_v$) the converse is true.

We conclude this section by drawing attention to the fidelity of the bounds from Lagrangian relaxation, i.e. $\bar{Q}_3^{\mathrm{d}}$ and $\bar{Q}_3^{\mathrm{s}}$. In Figure 1(a), $\bar{Q}_3^{\mathrm{d}}$ provides an effective bound on the likelihood, despite $L_\theta(y_{1:T})$ not being concave in the neighborhood of θ_k. In Figure 1(b), $\bar{Q}_3^{\mathrm{s}}$ almost perfectly reproduces Q^{s}, except at the boundary of the feasible set, $\mathcal{S}$, where it tends towards $-\infty$ as desired, unlike Q^{s}, which remains finite for unstable models ($A > 1$).



(a) 'Small' disturbances: $\Sigma_w = 1 \times 10^{-3}$ and $\Sigma_v = 1 \times 10^{-2}$.



(b) 'Large' disturbances: $\Sigma_w = 10$ and $\Sigma_v = 1 \times 10^{-2}$.

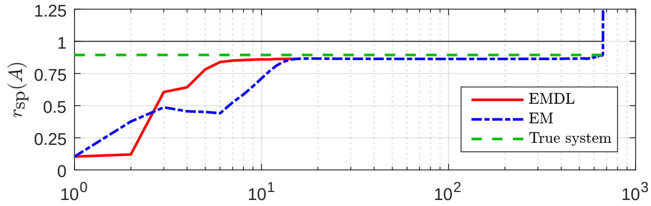
Figure 1: Bounds on the log likelihood $L_\theta(y_{1:T})$ of a 1^{st} order system with a single unknown scalar parameter, A. Q^{s} and Q^{d} denote the bounds based on latent states and disturbances resp., c.f. (10) and (16). $\bar{Q}_3^{\mathrm{s}}$ and $\bar{Q}_3^{\mathrm{d}}$ denote the bounds from Lagrangian relaxation, using latent states and disturbances resp., c.f. (29) and (38).

6. Numerical experiments

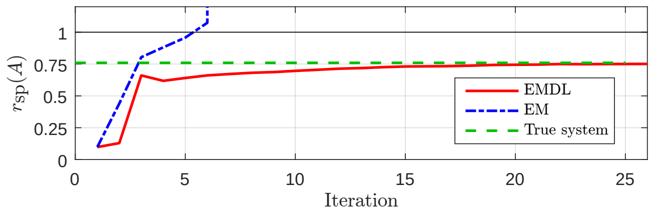
6.1. Stability of the identified model

This section provides empirical evidence of the value of the model stability constraints introduced in §4. We present examples of model instability arising in the standard unconstrained latent states formulation, during identification of models depicted in Figure 4. The exact numerical specification of these systems is avaiable at [24]. Figure 2 considers the case where measurement noise is more significant than the disturbances, for singular state space models; Figure 3 treats the 'significant disturbances' case with full rank covariance. In all examples, both the true and identified models are fourth order. We make the following observations. First, model instability can arise even when the true spectral radius is far from unity; c.f. Figure 2(b) and 3(b) concerning identification of the overdamped System 2 in Figure 4. Secondly, the consequences of instability are varied; e.g. in Figure 2(b), model instability leads to failure of the latent states algorithm due to poor numerical conditioning, whereas in Figure 3(a) the spectral radius of the identified model hovers above unity for thousands of iterations. Such a model may achieve adequate performance on training data, yet behave unreliably should the unstable modes be excited during validation.

To supplement the results in Figures 2 and 3, we randomly generated 1500 stable SISO systems, of varying order, with Matlab's `drss` function, and report instability of the identified models in Table 1. Specifically, to generate problem data each model was simulated for $T = 2n_\theta$ time steps (where n_θ is the number of parameters in the model) with Σ_v set to give a SNR of 20dB, and GG' of rank 1 with eigenvalue 10^{-4}. Each identified model was of the same order as the system used to generate the training data. The latent states algorithm [28] was then run for 60 seconds, randomly initialized with `drss`. The proportion of trials for which the identified model was unstable for at least one iteration is recorded in Table 1.

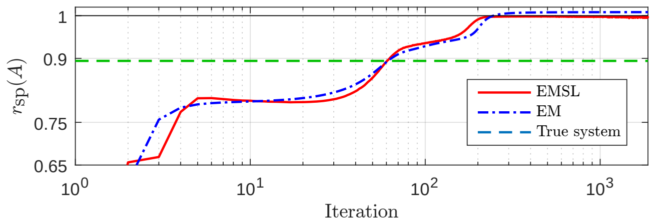


(a) System 1, underdamped: $\Sigma_v = 1.7 \times 10^{-3}$.

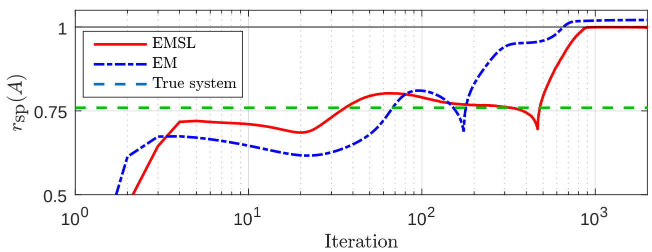


(b) System 2, overdamped: $\Sigma_v = 0.012$.

Figure 2: Spectral radius of identified models at each iteration of two methods: our latent disturbances method (EMDL), and the latent states method [28] (EM); c.f. Figure 4 for Bode plots of true systems. Both models (true and identified) are 4^{th} order. For each model, the disturbance covariance is singular with $G = [0, \sqrt{10^{-5}}, 0, 0]'$. In each case, both algorithms were initialized with the same randomly generated model from **drss**. Models were trained with $T = 75$ and 100 datapoints, in (a) and (b), resp.



(a) System 1, underdamped: $\Sigma_w = I_4$, $\Sigma_v = 1.7 \times 10^{-3}$.



(b) System 2, overdamped: $\Sigma_w = I_4$, $\Sigma_v = 0.012$.

Figure 3: Spectral radius of identified models at each iteration of two methods: our stable latent states method (EMSL), and the latent states method [7] (EM); c.f. Figure 4 for Bode plots of true systems. Both models (true and identified) are 4^{th} order. For each model, the disturbance covariance is full rank. In each case, both algorithms were initialized with the same randomly generated model from **drss**. Models were trained with $T = 75$ datapoints.

Table 1: Proportion of trials for which the identified model was unstable for at least one iteration, using the latent states algorithm [28]. 300 trials were conducted for each model order. The true SISO models were generated with **drss**. Each identified model was of the same order as the true model; c.f. Section 6.1 for details.

Model size, n_x	2	4	6	8	10
Unstable model	32%	32%	36%	48%	47%

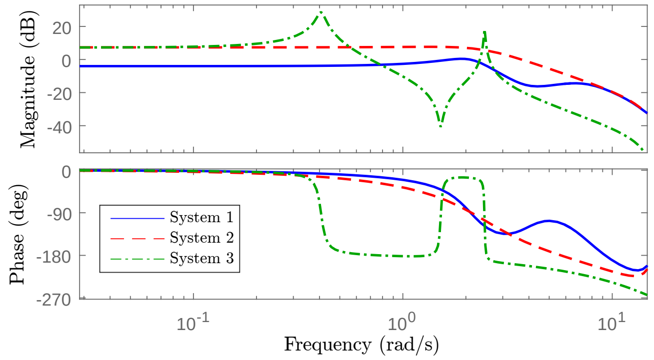


Figure 4: Bode plots of 4^{th} order systems used in the numerical experiments of Section 6. Refer to [24] for exact specifications.

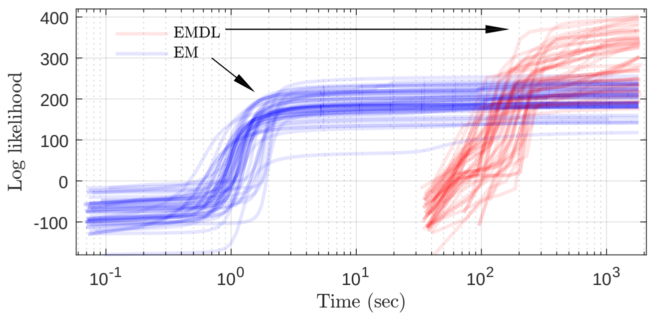
6.2. Convergence rate and computation time

In this section, we demonstrate that although the per-iteration complexity of our latent disturbances formulation (EMDL), c.f. Algorithm 2, is much greater, total computation time remains competitive with conventional latent states methods in cases where measurement noise dominates disturbances. This is due to the higher fidelity bounds on likelihood achieved using latent disturbances, e.g. Figure 1(a), meaning fewer iterations are required to significantly improve the likelihood. To illustrate this, we identify three 4^{th} order linear models, the Bode plots for which are given in Figure 4. Each of these systems has GG' of rank 1, with eigenvalue 10^{-5} ($G = [0, \sqrt{10^{-5}}, 0, 0]'$). We set Σ_v to give a signal-to-noise ratio (SNR) of approx. 20dB, which means Σ_v is two to three orders of magnitude larger than GG'. The experiment consists of 50 trials; in a single trial we repeat the following process for each system in Figure 4. First we simulate the system for $T = 250$ time steps, excited by $u_t \sim \mathcal{N}(0,1)$, to generate problem data $u_{1:T}$ and $y_{1:T}$. We then run EMDL and [28] for 30 minutes. Each algorithm is initialized with the same model (randomly generated for each trial), with system matrices close to zero.

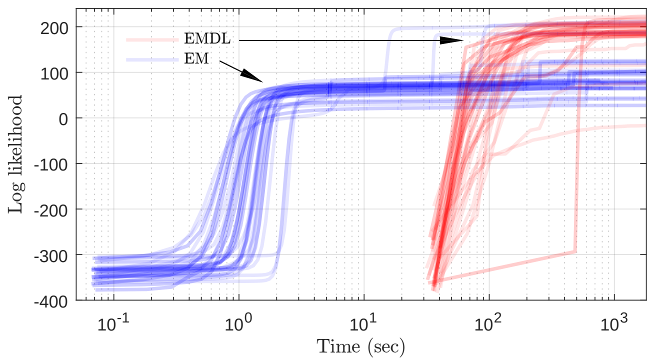
Figure 5 presents the log likelihood as a function of computation time. The longer per-iteration time of EMDL is immediately apparently; tens of seconds elapse before the first iteration completes. After 200 seconds of computation, EMDL is approx. equal to (or greater than) EM, and after 30 minutes EMDL has surpassed EM in almost all cases. These higher likelihoods correspond to more accurate models, as revealed by Figure 6, which plots the H_∞ and prediction error (on validation data) of the systems identified in Figure 5.

For the highly resonant System 3, both formulations of EM exhibited poor performance due to apparent capture in local maxima or stationary points. To illustrate this phenomenon, we compare performance with two different initialization strategies, and show the results in Figure 7.

The "cold start" results are as above, with the algorithms initialized with random system matrices close to

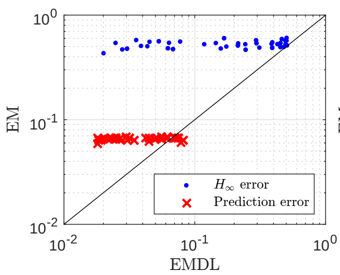


(a) System 1, smooth resonant peaks.

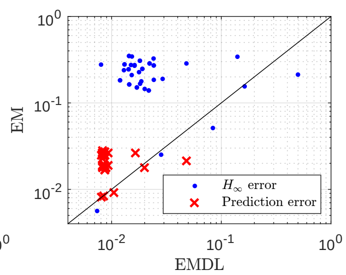


(b) System 2, overdamped.

Figure 5: Log likelihood as a function of computation time for EMDL and the latent states formulation of [28] (EM). 50 trials were carried out for each system, and $T = 250$ datapoints were used for fitting. Bode plots for System 1 and 2 are depicted in Figure 4.



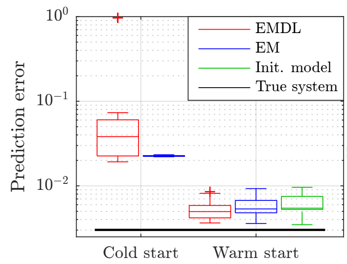
(a) System 1.



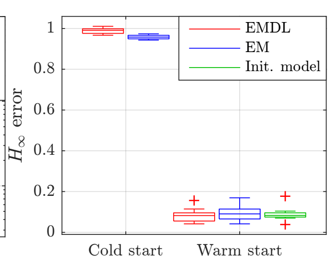
(b) System 2.

Figure 6: Prediction error (on validation data) and H_∞ error for EMDL and the latent states formulation of [28] (EM). The systems used are those reached at the conclusion of the trials depicted in Figure 5.

zero. In Figure 7(b) we see that the normalized H_∞ error is close to unity for both choices of latent variables, indicating only marginal improvement over a model with all matrices zero. The "warm start" results follow the common approach of initializing EM with a model produced by subspace identification. Here we used the modified subspace method of [32] which uses Lagrangian relaxation in place of least-squares for the second stage in order to guarantee model stability. Again we observe little improvement over the initial model in terms of prediction and H_∞ error, c.f. Figures 7(a) and (b), respectively. The results in Figure 7 suggest that for highly resonant systems, it



(a) Prediction error.



(b) H_∞ error.

Figure 7: Prediction error (on validation data) and H_∞ error for (EMDL) and the latent states formulation of [28] (EM), after identification of the highly resonant System 3, c.f. Figure 4. 'Cold start' and 'warm start' denote initialization with a random model (with system matrices close to zero) and a model from the subspace-Lagrangian method of [32], respectively.

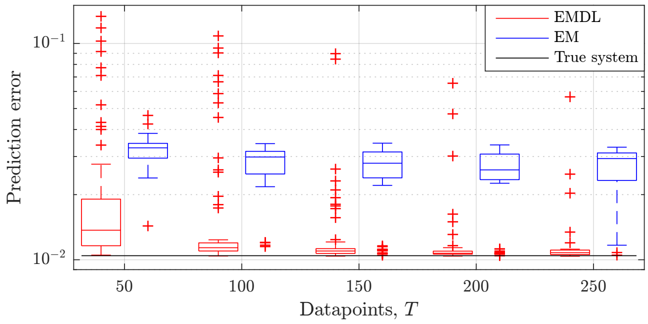
can be difficult for EM algorithms to make significant improvements to the initial model estimate, regardless of the choice of latent variables.

Finally, we analyse the performance of EMDL as a function of the number of datapoints used for training, T. Figure 8 presents H_∞ and prediction error for increasing T, for identification of System 2. System 2 was selected because capture in local maxima is less common, allowing us to study the asymptotic behavior of the global maximum more reliably. Both EMDL and the latent states algorithm [28] were run for 30 minutes in each trial. For EMDL, we observe an increase in accuracy (i.e., a decrease in both H_∞ and prediction error) for increasing T. In fact, for $T \geq 200$ the prediction error of the identified model is approximately equal to that of the true model. For the latent states algorithm, this trend is much less pronounced. The weak performance of latent states, along with the latent disturbances outliers, appears to be due to capture in local maxima, as model quality fails to improve in these cases, even after many additional iterations.

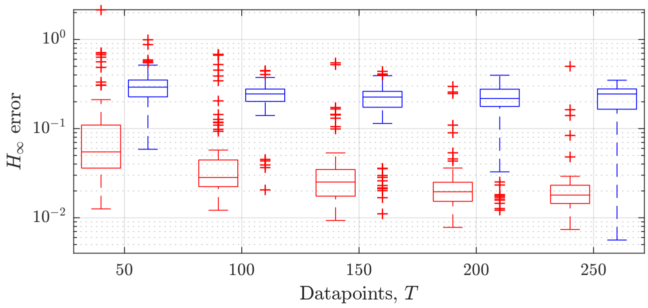
Table 2 records the mean computation time for a single iteration of the experimental trials carried out in Figure 8. Latent states methods, including EMSL, scale linearly with T, as the cost is dominated by the filtering and smoothing operations in the E step. In principle, EMDL, is $O(T^2)$, as optimization of each of the $M^{\mathrm{d}} + 1$ bounds in (38) requires $O(T)$, [32]. In practice, all M^{d} singular values of Ω are not typically required for accurate approximation of $Q_3(\theta, \theta_k)$.

Table 2: Mean per-iteration computation time (in seconds, to 3 sig. fig.) for the trials in Figure 8. EMSL and EMDL denote Algorithm 1 and Algorithm 2, respectively. EMSL is included for reference.

Data length, T	50	100	150	200	250
EM [28]	0.028	0.0541	0.08	0.103	0.126
EMSL	0.197	0.214	0.235	0.254	0.271
EMDL	37.5	40.7	48.9	54.6	65.6



(a) Prediction error, on validation data



(b) H_∞ error.

Figure 8: Prediction error (on validation data) and H_∞ error for EMDL and the latent states formulation of [28] (EM). The true model is the overdamped System 2, c.f. Figure 4. For each T, 50 trials were carried out, and each algorithm was run for 30 minutes.

7. Conclusion

This paper has incorporated model stability constraints into the maximum likelihood identification of linear dynamical systems. By combining the EM algorithm and Lagrangian relaxation, we construct tight convex bounds on the (negative) likelihood, that can be optimized over a convex parametrization of all stable linear systems, with semidefinite programming. The key practical outcomes of this work are as follows. The *de facto* choice of latent states leads to the simplest algorithms, as well as higher fidelity bounds on the likelihood when disturbances are more significant than measurement noise (i.e. $\Sigma_w \gg \Sigma_v$). Concerning software implementation, incorporating stability constraints into standard latent *states* algorithms is straightforward: if the identified model becomes unstable, simply replace the usual M step (11c) with the convex program (30), to continue the search over a convex set of stable models. On the other hand, when measurement noise is more significant than disturbances it may be advisable to formulate EM with latent *disturbances*. Although the per-iteration computational complexity of the ensuing algorithm is greater, the improved fidelity of the bounds on the likelihood can lead to faster convergence and more accurate models (e.g. by avoiding local maxima). Furthermore, latent disturbances lead to the most broadly applicable formulation of EM for identification of singular state space models.

Acknowledgments

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).

References

[1] Byron Boots. Learning Stable Linear Dynamical Systems. https://www.ml.cmu.edu/research/dap-papers/dap_boots.pdf. Accessed: 20-02-2018.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[3] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[4] James Durbin and Siem Jan Koopman. *Time series analysis by state space methods*. Number 38 in Oxford Statistical Science Series. Oxford University Press, 2012.

[5] Wouter Favoreel, Bart De Moor, and Peter Van Overschee. Subspace state space system identification for industrial processes. *Journal of Process Control*, 10(2):149–155, 2000.

[6] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, November 1984.

[7] Stuart Gibson and Brett Ninness. Robust maximum-likelihood estimation of multivariable dynamic systems. *Automatica*, 41(10):1667–1682, 2005.

[8] Wilfred K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[9] David R. Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[10] Seth L Lacy and Dennis S Bernstein. Subspace identification with guaranteed stability using constrained optimization. In *Proceedings of the American Control Conference (ACC)*, volume 4, pages 3307–3312, Anchorage, USA, 2002.

[11] Seth L Lacy and Dennis S Bernstein. Subspace identification with guaranteed stability using constrained optimization. *Automatic Control, IEEE Transactions on*, 48(7):1259–1263, 2003.

[12] Wallace E. Larimore. System identification, reduced-order filtering and modeling via canonical variate analysis. In *Proceedings of the American Control Conference (ACC)*, pages 445–451, San Francisco, USA, 1983.

[13] Claude Lemaréchal. Lagrangian relaxation. In *Computational Combinatorial Optimization*, pages 112–156. Springer, 2001.

[14] Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, 2 edition, January 1999.

[15] Lennart Ljung. Prediction error estimation methods. *Circuits, Systems and Signal Processing*, 21(1):11–21, 2002.

[16] Jan M Maciejowski. Guaranteed stability with subspace methods. *Systems & Control Letters*, 26(2):153–156, 1995.

[17] Ian Manchester, Mark M Tobenkin, and Alexandre Megretski. Stable nonlinear system identification: Convexity, model class, and consistency. In *Proceedings of the 16th IFAC Symposium on System Identification (SYSID)*, Brussels, Belgium, 2012.

[18] Alexandre Megretski. Convex optimization in robust identification of nonlinear feedback. In *Proceedings of the 47th IEEE Conference on Decision and Control (CDC)*, pages 1370–1374, Cancun, Mexico, 2008.

[19] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[20] Daniel N Miller and Raymond A De Callafon. Subspace identification with eigenvalue constraints. *Automatica*, 49(8):2468–2473, 2013.

[21] James M. Ortega and Werner C. Rheinboldt. *Iterative solution of nonlinear equations in several variables.* SIAM, 1970.

[22] Imre Pólik and Tamás Terlaky. A survey of the s-lemma. *SIAM review*, 49(3):371–418, 2007.

[23] Herbert E. Rauch, C.T. Striebel, and F. Tung. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450, 1965.

[24] Thomas B. Schön. System specifications. http://user.it.uu.se/~thosc112/umenbergerwmssystemspec.zip. 2017.

[25] Thomas B. Schön, Adrian Wills, and Brett Ninness. System identification of nonlinear state-space models. *Automatica*, 47(1):39–49, 2011.

[26] Robert H. Shumway and David S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982.

[27] Sajid Siddiqi, Byron Boots, and Geoffrey J. Gordon. A constraint generation approach to learning stable linear dynamical systems. In *Proceedings of Advances in Neural Information Processing Systems 20 (NIPS)*, 2007.

[28] Victor Solo. An EM algorithm for singular state space models. In *Proceedings of the 42nd IEEE Conference on Decision and Control (CDC)*, pages 3457–3460, Maui Maui, USA, 2003.

[29] Victor Solo. An EM algorithm for singular state space models II. In *Proceedings of the 43rd IEEE Conference on Decision and Control (CDC)*, pages 3611–3612, The Bahamas, 2004.

[30] Mark M Tobenkin, Ian R Manchester, and Alexandre Megretski. Convex parameterizations and fidelity bounds for nonlinear identification and reduced-order modelling. *IEEE Transactions on Automatic Control*, 62(7):3679–3686, 2017.

[31] Mark M. Tobenkin, Ian R. Manchester, Jennifer Wang, Alexandre Megretski, and Russ Tedrake. Convex optimization in identification of stable non-linear state space models. In *Proceedings of the 49th IEEE Conference on Decision and Control, CDC, Atlanta, USA*, pages 7232–7237, 2010.

[32] Jack Umenberger and Ian R. Manchester. Specialized algorithm for identification of stable linear systems using lagrangian relaxation. In *Proceedings of the American Control Conference (ACC)*, pages 930–935, Boston, USA, 2016.

[33] Jack Umenberger, Johan Wågberg, Ian R. Manchester, and Thomas B. Schön. On identification via EM with latent disturbances and Lagrangian relaxation. In *Proceedings of the 17th IFAC Symposium on System Identification (SYSID)*, Beijing, China, 2015.

[34] Tony Van Gestel, Johan AK Suykens, Paul Van Dooren, and Bart De Moor. Identification of stable models in subspace identification by using regularization. *Automatic Control, IEEE Transactions on*, 46(9):1416–1420, 2001.

[35] Peter Van Overschee and Bart De Moor. N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, 30(1):75–93, 1994.

[36] Adrian Wills, Thomas B Schön, Fredrik Lindsten, and Brett Ninness. Estimation of linear systems using a gibbs sampler. *IFAC Proceedings Volumes*, 45(16):203–208, 2012.

[37] Adrian Wills, Thomas B. Schön, and Brett Ninness. Estimating state-space models in innovations form using the expectation maximisation algorithm. In *Proceedings of the 49th IEEE Conference on Decision and Control (CDC)*, pages 5524–5529, Atlanta, USA, 2010.

[38] Vladmir Yakubovich. S-procedure in nonlinear control theory. *Vestnik Leningrad University*, 1:62–77, 1971.

Appendix A. Proofs for Section 3

Appendix A.1. Proof of Lemma 1

The first term in (16), Q_1^{d}, is identical to Q_1^{s}, so we focus on Q_3^{d}. The p.d.f. $p_\theta(y_{1:T}|x_1, w_{1:T})$ is given by $p_\theta(Y|Z) = \mathcal{N}(Y;\mu_Y,\Sigma_Y)$, where μ_Y and Σ_Y are given in (19). Q_3^{d} may then be expressed as

$$\begin{aligned}&\mathrm{E}_{\theta_k}\left[\log\mathcal{N}(Y;\mu_Y,\Sigma_Y)|y_{1:T}\right]\\&\quad=-\frac{Tn_y}{2}\log 2\pi-\log\det\Sigma_Y-\mathrm{E}_{\theta_k}\left[|Y-\mu_Y|^2_{\Sigma_Y^{-1}}|y_{1:T}\right].\end{aligned}$$

Letting $\hat{Z}$ and Ω, defined in (18), denote the mean and covariance (respectively) of $p_{\theta_k}(x_1, w_{1:T-1}|y_{1:T})$, gives

$$Q_3(\theta,\theta_k)\propto -T\log\det\Sigma_v-\mathrm{tr}(\Sigma_Y^{-1}(\bar{C}\bar{H}\Omega\bar{H}'\bar{C}'+\hat{\Delta}\hat{\Delta}')),\tag{A.1}$$

where $\hat{\Delta}=\mathrm{E}_{\theta_k}\left[Y-\mu_Y|y_{1:T}\right]$ is defined in (20). □

Appendix B. Proofs for Section 4

Appendix B.1. Proof of Lemma 4

Evaluating the supremum in (28) yields

$$\bar{J}^s_\lambda(\theta^s,t)=\epsilon_t'H\left(H'E+E'H-\Sigma_w^{-1}\right)^{-1}H'\epsilon_t.\tag{B.1}$$

Let $e_t=\tilde{x}_{t+1}-A\tilde{x}_t-B\tilde{u}_t$, such that $\epsilon_t=Ee_t$. Then substituting $H=(E_k')^{-1}\Sigma_w^{-1}$ into (B.1) gives $\bar{J}^s_\lambda(\theta^s_k,t)=e_t'\Sigma_w^{-1}e_t$, i.e. (28) is tight to (26) at θ_k. □

Appendix B.2. Proof of Lemma 5

First consider the $\mathrm{tr}(\Sigma_Y\hat{\Delta}\hat{\Delta}')$ term in (A.1). From (20), $\hat{\Delta}$ is clearly the difference between the measured output $y_{1:T}$ and the simulated output of the model with the expected value of the latent disturbances, i.e. $\hat{Z}$. Therefore, $\mathrm{tr}(\Sigma_Y\hat{\Delta}\hat{\Delta}')=\sum_{t=1}^T|y_t-Cx_t-Du_t|^2_{\Sigma_v^{-1}}$ where $\mathrm{vec}(x_{1:T})=\bar{N}U+\bar{H}\bar{Z}$. Next, consider $\mathrm{tr}(\Sigma_Y^{-1}\bar{C}\bar{H}\Omega\bar{H}'\bar{C}')$. Decomposing $\Omega=\sum_{j=1}^{M^{\mathrm{d}}}Z^jZ^{j'}$ leads to

$$\mathrm{tr}(\Sigma_Y^{-1}\bar{C}\bar{H}\Omega\bar{H}'\bar{C}')=\sum_{j=1}^{M^{\mathrm{d}}}|\bar{C}\bar{H}w_j|^2_{\Sigma_Y^{-1}}=\sum_{j=1}^{M^{\mathrm{d}}}\sum_{t=1}^{T}|Cx_t^j|^2_{\Sigma_v^{-1}}$$

where $\mathrm{vec}(x^j_{1:T})=\bar{H}Z^j$., i.e. the sum of M^{s} simulation error problems with $Y=0$, $U=0$ and $Z=Z^j$. □

Appendix B.3. Proof of Lemma 6

As $\bar{Q}_3^{\mathrm{d}}(\eta)$ is defined by a summation of convex functions, it is itself a convex function. Summation of the following inequalities

$$\begin{aligned}&\bar{J}_{\lambda^0}(\eta,u_{1:T},y_{1:T},\hat{x}_{1|T},\hat{w}_{1:T})\geq\mathcal{E}(\eta,u_{1:T},y_{1:T},\hat{x}_{1|T},\hat{w}_{1:T}),\\&\bar{J}_{\lambda^j}(\eta,0,0,x_1^j,w^j_{1:T})\geq\mathcal{E}(\eta,0,0,x_1^j,w^j_{1:T}),\quad j=1,\ldots,T,\\&\mathrm{tr}(\Sigma_{v_k}^{-1}\Sigma_v)+\log\det\Sigma_{v_k}+n_y\geq\log\det\Sigma_v,\end{aligned}$$

gives $\bar{Q}_3^{\mathrm{d}}(\eta)\geq -Q_3(\beta,\theta_k)$. Notice that $n_y+\log\det\Sigma_{v_k}+\mathrm{tr}(\Sigma_{v_k}^{-1}\Sigma_v)$ is an affine upper bound on the concave term $\log\det\Sigma_v$, which is tight at our current best estimate of the covariance, Σ_{v_k}. □

Appendix B.4. Proof of Lemma 7

For $\theta_k^d \in \Theta^d$, we have $\Psi_k < 0$, c.f. [30, Theorem 6]. The Lagrangian in (37) is then concave in Δ, so the supremizing Δ satisfies $\Psi\Delta = \bar{E}_k' h^j + \bar{C}_k'(Y - \bar{D}_k U) - \bar{\epsilon}_k$. Substituting h^j from (39) into the above yields $\Psi\Delta = \Psi\bar{E}_k^{-1}\bar{\epsilon}_k$. As Ψ is full rank, this implies $\bar{E}_k\Delta = \bar{\epsilon}_k$. Then $J(\theta_k^d, \Delta) - \lambda F(\theta_k^d, \Delta) = J(\theta_k^d, \Delta) = \mathcal{E}(\theta_k^d)$. □

Appendix C. Proofs for Section 5

Appendix C.1. Proof of Proposition 8

When $\Sigma_w = 0$, $p_{\theta_k}(x_{1:T}|y_{1:T})$ is supported on the set

$$\mathcal{X}(\theta_k) = \{x_{1:T} : \text{vec}(x_{1:T}) = \bar{N}U + \bar{H}\tilde{Z} \;\forall\; \xi_1 \in \mathbb{R}^{n_x}\}$$

where $\tilde{Z} = [\xi_1', 0]'$ and $G = 0$. Then,

$$Q^{\text{s}}(\theta, \theta_k) = \int_{\mathcal{X}(\theta_k)} \log p_\theta(x_{1:T}, y_{1:T}) p_{\theta_k}(x_{1:T}|y_{1:T}) dx_{1:T}.$$

As $\Sigma_w = 0$, $p_\theta(x_{2:T}|x_1)$ is deterministic. When $A \neq A_k$ or $B \neq B_k$, $\log p_\theta(x_{2:T}|x_1) = 0$ for all $x_{1:T} \in \mathcal{X}(\theta_k)$, and so $\log p_\theta(x_{1:T}, y_{1:T})$ is undefined. As a consequence, $Q^{\text{s}}(\theta, \theta_k)$ is undefined.

When $A = A_k$ and $B = B_k$, $p_\theta(x_{2:T}|x_1) = 1$ for all $x \in \mathcal{X}(\theta_k)$ and so $Q^{\text{s}}(\theta, \theta_k)$ can be evaluated as usual. □

Appendix C.2. Proof of Proposition 9

As $G = 0$, $\Sigma_1 = 0$ the p.d.f. $p_{\theta_k}(x_1, w_{1:T}|y_{1:T})$ is trivially deterministic, evaluating to unity when $x_1 = \mu$ and $w_{1:T} \equiv 0$, and evaluating to zero otherwise. Therefore

$$Q^{\text{d}}(\theta, \theta_k) = \log p_\theta(y_{1:T}, \mu, 0) = \log p_\theta(y_{1:T}|\mu).$$

The log likelihood can be decomposed as

$$\begin{aligned} L_\theta(y_{1:T}) &= \log \int p_\theta(y_{1:T}, x_1) dx_1 \\ &= \log \int p_\theta(y_{1:T}|x_1) p_\theta(x_1) dx_1 = \log p_\theta(y_{1:T}|\mu), \end{aligned}$$

where the final equality follows from the fact that $p_\theta(x_1)$ is a δ-function, at $x_1 = \mu$. □

Appendix C.3. Proof of Proposition 10

For a given θ, let $x_{1:T}^\theta$ denote the unique state sequence that is 'consistent' with the data, i.e. $x_{1:T}^\theta \triangleq \{x_{1:T} : y_t = Cx_t + Du_t, t = 1, \ldots, T\}$. There is also a corresponding unique disturbance sequence, denoted $w_{1:T}^\theta = \{w_{1:T} : x_{t+1}^\theta = Ax_t^\theta + Bu_t + Gw_t, t = 1, \ldots, T\}$.

As $\Sigma_v = 0$, the p.d.f. $p_{\theta_k}(x_1, w_{1:T}|y_{1:T})$ is a δ-function at $x_1 = x_1^{\theta_k}$ and $w_{1:T} = w_{1:T}^{\theta_k}$. The auxiliary function is then given by $Q^{\text{d}}(\theta, \theta_k) = \log p_\theta(y_{1:T}, x_1^{\theta_k}, w_{1:T}^{\theta_k})$. We can decompose $p_\theta(y_{1:T}, x_1^{\theta_k}, w_{1:T}^{\theta_k})$ as in (14). As $\Sigma_v = 0$, the p.d.f. $p_\theta(y_{1:T}|x_1, w_{1:T})$ is also a δ-function at $x_1 = x_1^\theta$ and $w_{1:T} = w_{1:T}^\theta$. If $\theta \neq \theta_k$ then $x_1^\theta \neq x_1^{\theta_k}$ and $w_{1:T}^\theta \neq w_{1:T}^{\theta_k}$. In this case $p_\theta(y_{1:T}|x_1^{\theta_k}, w_{1:T}^{\theta_k}) = 0$ and so $Q^{\text{d}}(\theta, \theta_k)$ is undefined.

When $\theta = \theta_k$, $p_\theta(y_{1:T}|x_1^{\theta_k}, w_{1:T}^{\theta_k}) = 1$ and $Q^{\text{d}}(\theta, \theta_k)$ can be evaluated as usual. □

Appendix C.4. Proof of Proposition 11

For a given θ, let $x_{1:T}^\theta$ denote the unique state sequence that is 'consistent' with the data, i.e. $x_{1:T}^\theta \triangleq \{x_{1:T} : y_t = Cx_t + Du_t, t = 1, \ldots, T\}$. As $\Sigma_v = 0$, given $y_{1:T}$ both $p_{\theta_k}(x_{1:T}|y_{1:T})$ and $p_\theta(y_{1:T}|x_{1:T})$ are δ-functions at $x_{1:T} = x_{1:T}^\theta$. The auxiliary function is then given by

$$Q^{\text{s}}(\theta, \theta_k) = \log p_\theta(y_{1:T}, x_{1:T}^{\theta_k}).$$

Let us now consider the two cases:

i. When $C \neq C_k$ or $D \neq D_k$, $x_{1:T}^\theta \neq x_{1:T}^{\theta_k}$ and so $p_\theta(y_{1:T}|x_{1:T}^{\theta_k}) = 0$. Therefore, $Q^{\text{s}}(\theta, \theta_k)$ is undefined.

ii. When $C = C_k$ and $D = D_k$, $x_{1:T}^\theta = x_{1:T}^{\theta_k}$ and so

$$Q^{\text{s}}(\theta, \theta_k) = \log p_\theta(y_{1:T}|x_{1:T}^\theta) p_\theta(x_{1:T}^\theta) = \log p_\theta(x_{1:T}^\theta).$$

The likelihood can be expressed as

$$\begin{aligned} L_\theta(y_{1:T}) &= \log \int p_\theta(y_{1:T}|x_{1:T}) p_\theta(x_{1:T}) dx_{1:T} \\ &= \log p_\theta(x_{1:T}^\theta), \end{aligned}$$

where the second inequality comes from the fact that $p_\theta(y_{1:T}|x_{1:T})$ is a δ-function. Therefore, $L_\theta(y_{1:T}) = Q^{\text{s}}(\theta, \theta_k)$. □