Maximum-a-posteriori estimation of LTI state-space models via efficient Monte-Carlo sampling

Manas Mejari and Dario Piga
Dalle Molle Institute for Artificial Intelligence, USI-SUPSI,
Via la Santa 1, CH-6962 Lugano-Viganello, Switzerland.
Email: {manas.mejari,dario.piga}@supsi.ch

Abstract- *This paper addresses Maximum-A-Posteriori (MAP) estimation of Linear Time-Invariant State-Space (LTI-SS) models. The joint posterior distribution of the model matrices and the unknown state sequence is approximated by using Rao-Blackwellized Monte-Carlo sampling algorithms. Specifically, the conditional distribution of the state sequence given the model parameters is derived analytically, while only the marginal posterior distribution of the model matrices is approximated using a Metropolis-Hastings Markov-Chain Monte-Carlo sampler. From the joint distribution, MAP estimates of the unknown model matrices as well as the state sequence are computed. The performance of the proposed algorithm is demonstrated on a numerical example and on a real laboratory benchmark dataset of a hair dryer process.*

1 Introduction

Identification of LTI-SS models is a topic of interest to the system and control community, thanks to their ability to efficiently describe the behavior of multi-input multi-output dynamical systems. The seminal paper by Ho-Kalman [1] gave the first solution to the deterministic realization problem, *i.e.*, characterization of LTI-SS representations from input-output (IO) data. This led to the development of subspace identification algorithms based on IO and SS realization schemes, like, MOESP [2], N4SID [3]. Alternatively, in the *maximum likelihood* (ML) or *prediction error* methods [4], the SS matrices are parameterized and the optimal parameters are obtained by minimizing a cost function [5].

A *Bayesian* framework has been proposed in [6, 7] for identification of LTI models. The contribution [6] employs *Metropolis-Hastings Markov Chain Monte Carlo* (MH-MCMC) sampling to approximate the posterior density of the model parameters, while [7] uses *blocked Gibbs sampling*. The approach in [7], however, requires a specific form of the prior on the unknown parameters in order to obtain samples from its posterior easily. Particle or sequential MC methods have been proposed in [8, 9, 10] for non-linear SS models. Specifically, in [9, 10], marginalization and data augmentation strategies have been presented using sequential MC methods. The marginalization amounts to integrating out the state sequence, as done within the Expectation-Maximization algorithm [11] and viewing only the model parameters as unknown. On the other hand, data augmentation treats the states as latent variables which are estimated along with the unknown model parameters. To this end, a Gibbs sampler is presented in [9] which, however, requires to sample from a high dimensional distribution over the space of unknown state sequence. A dual Extended Kalman filter (EKF) is proposed in [12] which sequentially estimates state sequence and model parameters by running two EKFs concurrently.

Inspired from the data augmentation strategies [9, 12], in this paper, we develop a *batch* identification algorithm for LTI-SS representations and compute MAP estimates of the LTI model matrices as well as the unknown state sequence. To this end, we seek the mode of the *joint* posterior distribution over the model matrices and the unknown state-sequence. According to Rao-Blackwellized strategy [13], the joint posterior is approximated by deriving the conditional distribution of the state-sequence (given the model matrices) analytically, while the marginal posterior of model matrices is approximated via MH-MCMC [14] sampling. We remark that in [9], the MH-MCMC has been presented to approximate the marginal posterior over the model parameters while the state sequence is marginalized. On the other hand, in this work, we seek the *joint* posterior of model matrices and unknown state sequence by relying on Rao-Blackwellization technique. This allows us to avoid sampling from a high-dimensional distribution over the space of state sequence. From the joint posterior, MAP estimates of model matrices and the unknown state sequence are computed.

2 Problem formulation

We consider the following discrete-time LTI-SS model

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

where $x_t \in \mathbb{R}^{n_x}$, $u_t \in \mathbb{R}^{n_u}$ and $y_t \in \mathbb{R}^{n_y}$ denote the state, the input and the output of the model at time index t, respectively. The process noise $w_t \in \mathbb{R}^{n_x}$ and the measurement noise $v_t \in \mathbb{R}^{n_y}$ are assumed to be zero mean white Gaussian noises with known covariance matrices Q_w and Q_v, respectively, *i.e.*, $w_t \sim \mathcal{N}(0, Q_w)$ and $v_t \sim \mathcal{N}(0, Q_v)$. In order to compact the notation, we stack the LTI-SS model matrices A, B, C, D in a vector $\Theta \in \mathbb{R}^{n_\theta}$ (with $n_\theta = n_x^2 + n_x n_u + n_x n_y + n_y n_u$), *i.e.*,

$$\Theta = \left[(vec(A))^\top \ (vec(B))^\top \ (vec(C))^\top \ (vec(D))^\top \right]^\top.$$

Given a training data set $\{u_{1:T}, y_{1:T}\}$ consisting of T input-output pairs, we aim at computing (and then maximizing) the joint posterior distribution $p(\Theta, x_{1:T} | u_{1:T}, y_{1:T})$ of the model matrices A, B, C, D and of the state sequence $x_{1:T}$, under the following prior over the parameters Θ and the initial state x_1:

$$(\Theta, x_1) \sim p(\Theta, x_1) = \mathcal{N}((\Theta, x_1); 0, \sigma^2 I_{n_\theta + n_x}). \tag{2}$$

Through the rest of the paper, we drop the conditional dependence on the input sequence $u_{1:T}$ for brevity.

3 Rao-Blackwellized Monte-Carlo Sampling Algorithm

The posterior distribution of interest $p(\Theta, x_{1:T} | y_{1:T})$ can not be computed analytically. In order to overcome this problem, $p(\Theta, x_{1:T} | y_{1:T})$ can be approximated via Markov Chain Monte Carlo sampling. The idea of MCMC algorithms is to draw a set of M samples, $\{\Theta[k], x_1[k], \cdots x_T[k]\}_{k=1}^M$, from an irreducible and aperiodic Markov chain whose stationary distribution is the target distribution of interest. In this way, the posterior distribution can be approximated by the empirical point-mass distribution as: $p(\Theta, x_{1:T} | y_{1:T}) \approx \frac{1}{M} \sum_{k=1}^M \delta_{\Theta[k], x_1[k], \cdots x_T[k]}(\Theta, x_{1:T})$. This *naive* MCMC approach requires to draw the samples from a high-dimensional parameter space $(\Theta, x_{1:T})$, which grows with the number of data points T. This may require to draw a large number of samples M to obtain a reasonable approximation of the target distribution. In order to overcome this problem, the joint posterior distribution is factorized as

$$p(\Theta, x_{1:T} | y_{1:T}) = p(x_{2:T} | \Theta, x_1, y_{1:T})\, p(\Theta, x_1 | y_{1:T}),$$

and the structure of the LTI-SS model (1) is exploited to compute an *analytical* expression of the conditional joint distribution of the unknown state sequence $p(x_{2:T} \mid \Theta, x_1, y_{1:T})$. Only the marginal posterior $p(\Theta, x_1 \mid y_{1:T})$ of the model matrices and the initial state is approximated with the empirical point-mass distribution $p(\Theta, x_1 | y_{1:T}) \approx \frac{1}{M} \sum_{k=1}^M \delta_{\Theta[k], x_1[k]}(\Theta, x_1)$ through MCMC simulation. This avoids sampling over the state parameters $x_{2:T}$. Approximating only the marginal of the target joint distribution of interest with MCMC sampling, is referred to as *Rao-Blackwellized* MCMC [13].

3.1 Computation of $p(x_{2:T} \mid \Theta, x_1, y_{1:T})$

In this section, we derive the analytic expression for the conditional distribution $p(x_{2:T} \mid \Theta, x_1, y_{1:T})$ of the unknown state sequence $x_{2:T}$ given the initial state x_1 and the model parameters Θ. We remark that the standard formulation of the Kalman smoother can not be directly used here as it only provides the marginal distribution $p(x_t \mid \Theta, y_{1:T})$, while we seek the joint distribution over the state sequence $p(x_{2:T} \mid \Theta, x_1, y_{1:T})$. To this end, we derive the joint distribution using Bayes rule, exploiting linear dynamical structure of (1) and by using multivariate Gaussian identities.

From the Bayes' rule, the distribution $p(x_{2:T} | \Theta, x_1, y_{1:T})$ can be written as

$$p(x_{2:T} | \Theta, x_1, y_{1:T}) = \frac{p(y_{1:T} | x_{1:T}, \Theta) p(x_{2:T} | \Theta, x_1)}{\int p(y_{1:T} | x_{1:T}, \Theta) p(x_{2:T} | \Theta, x_1) dx_{2:T}}. \tag{3}$$

In the following, we derive the expressions of likelihood $p(y_{1:T} | x_{1:T}, \Theta)$ and the conditional $p(x_{2:T} | \Theta, x_1)$ appearing in (3) to obtain the analytical expression for the joint state posterior $p(x_{2:T} \mid \Theta, x_1, y_{1:T})$. From (1a), the following linear prediction equation for the state x_t can be easily derived:

$$x_t = A^{t-1} x_1 + \sum_{k=1}^{t-1} A^{t-k-1} B u_k + \sum_{k=1}^{t-1} A^{t-k-1} w_k. \tag{4}$$

In order to compact the notation, equations (1b) and (4) are written in the matrix form as follows

$$X = \bar{A} x_1 + \bar{B} U_1 + G W, \tag{5a}$$
$$Y = \bar{C} X + \bar{D} U_2 + V, \tag{5b}$$

where $Y = y_{2:T}$, $X = x_{2:T}$, $U_1 = u_{1:T-1}$, $U_2 = u_{2:T}$, $W = w_{1:T-1}$, $V = v_{2:T}$ and $\bar{C} = \text{blkdiag}\{C, \cdots, C\} \in \mathbb{R}^{(T-1)n_y \times (T-1)n_x}$, $\bar{D} = \text{blkdiag}\{D, \cdots, D\} \in \mathbb{R}^{(T-1)n_y \times (T-1)n_u}$

$$\bar{A} = \begin{bmatrix} A \\ A^2 \\ \vdots \\ A^{T-1} \end{bmatrix}, \bar{B} = \begin{bmatrix} B & 0 & \cdots & 0 \\ AB & B & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 \\ A^{T-2}B & A^{T-3}B & \cdots & B \end{bmatrix}, G = \begin{bmatrix} I_{n_x} & 0 & \cdots & 0 \\ A & I_{n_x} & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 \\ A^{T-2} & A^{T-3} & \cdots & I_{n_x} \end{bmatrix}.$$

From (5) and because of the white noise assumption on the noise sequences W and V, the following identities hold:

$$p(x_{2:T} \mid \Theta, x_1) = \mathcal{N}(X; \bar{A} x_1 + \bar{B} U_1, G \bar{Q}_w G^\top), \tag{6a}$$
$$p(y_{1:T} \mid x_{1:T}, \Theta) = \mathcal{N}(y_1; C x_1 + D u_1, Q_v) \mathcal{N}(Y; \bar{C} X + \bar{D} U_2, \bar{Q}_v) \tag{6b}$$

where $\bar{Q}_w = \text{blkdiag}\{Q_w, \cdots, Q_w\} \in \mathbb{R}^{(T-1)n_x \times (T-1)n_x}$, $\bar{Q}_v = \text{blkdiag}\{Q_v, \cdots, Q_v\} \in \mathbb{R}^{(T-1)n_y \times (T-1)n_y}$.

Proposition 1. *Let us define the matrices* $b_1 = Y - \bar{D} U_2$, $b_2 = \bar{A} x_1 + \bar{B} U_1$, $\bar{G} = G \bar{Q}_w G^\top$, $H = \left[\bar{C}^\top \bar{Q}_v^{-1} \bar{C} + \bar{G}^{-1} \right]^{-1}$, $F = \left[\bar{C}^\top \bar{Q}_v^{-1} b_1 + \bar{G}^{-1} b_2 \right]$ *and* $\mu = HF$. *Then, the conditional joint posterior distribution of the state sequence*

$p(x_{2:T} \mid \Theta, x_1, y_{1:T})$ in (3) *is Gaussian with mean μ and covariance H,* i.e.*,*

$$p(x_{2:T} \mid \Theta, x_1, y_{1:T}) = \mathcal{N}(X; \mu, H). \tag{7}$$

Proof. From equations (6) and the properties $|G| = 1$, $\left|\bar{G}\right| = \left|G\bar{Q}_w G^\top\right| = \left|\bar{Q}_w\right| = |Q_w|^{T-1}$, the numerator and the denominator in (3) can be written as

$$\begin{aligned} p(y_{1:T} \mid x_{1:T}, \Theta) p(x_{2:T} \mid \Theta, x_1) &= \mathcal{N}(y_1; Cx_1 + Du_1, Q_v) \times \\ &\quad \mathcal{N}(Y; \bar{C}X + \bar{D}U_2, \bar{Q}_v) \mathcal{N}(X; \bar{A}x_1 + \bar{B}U_1, \bar{G}) \\ &= \mathcal{N}(y_1; Cx_1 + Du_1, Q_v) (2\pi)^{\frac{-(T-1)n_y}{2}} \times \\ &\quad \left|\bar{Q}_v\right|^{\frac{-1}{2}} (2\pi)^{\frac{-(T-1)n_x}{2}} \left|\bar{Q}_w\right|^{\frac{-1}{2}} (2\pi)^{\frac{(T-1)n_x}{2}} \times \\ &\quad |H|^{\frac{1}{2}} e^{-\frac{1}{2}\left[b_1^\top \bar{Q}_v^{-1} b_1 + b_2^\top \bar{G}^{-1} b_2 - F^\top \mu\right]} \mathcal{N}(X; \mu, H), \end{aligned} \tag{8a}$$

and

$$\begin{aligned} &\int p(y_{1:T} \mid x_{1:T}, \Theta) p(x_{2:T} \mid \Theta, x_1) dx_{2:T} \\ &= \mathcal{N}(y_1; Cx_1 + Du_1, Q_v) \times \\ &\quad \int \mathcal{N}(Y; \bar{C}X + \bar{D}U_2, \bar{Q}_v) \mathcal{N}(X; \bar{A}x_1 + \bar{B}U_1, \bar{G}) dX \\ &= \mathcal{N}(y_1; Cx_1 + Du_1, Q_v) \times \\ &(2\pi)^{\frac{-(T-1)n_y}{2}} \left|\bar{Q}_v\right|^{\frac{-1}{2}} (2\pi)^{\frac{-(T-1)n_x}{2}} \left|\bar{Q}_w\right|^{\frac{-1}{2}} \times \\ &(2\pi)^{\frac{(T-1)n_x}{2}} |H|^{\frac{1}{2}} e^{-\frac{1}{2}\left[b_1^\top \bar{Q}_v^{-1} b_1 + b_2^\top \bar{G}^{-1} b_2 - F^\top \mu\right]}, \end{aligned} \tag{8b, 8c}$$

where the above equations can be derived using the Gaussian identities in Appendix (see also [15]). Taking the ratio between (8a) and (8c), the proposition follows. □

3.2 Computation of $p(\Theta, x_1 \mid y_{1:T})$

The following proposition provides the expression of the marginal posterior distribution $p(\Theta, x_1 \mid y_{1:T})$, which is then approximated through Metropolis-Hastings MCMC (MH-MCMC) sampling.

Proposition 2. *The marginal posterior distribution $p(\Theta, x_1 \mid y_{1:T})$ of the model parameters Θ and the initial state x_1, given the sequence of output observations $y_{1:T}$, is given by*

$$p(\Theta, x_1 \mid y_{1:T}) = p(\Theta, x_1)\, k_\theta \times \frac{e^{-\frac{1}{2}\left[(b_1 - \bar{C}b_2)^\top \Sigma^{-1} (b_1 - \bar{C}b_2) + (y_1 - Cx_1 - Du_1)^\top Q_v^{-1} (y_1 - Cx_1 - Du_1)\right]}}{\int p(y_{1:T} \mid \Theta, x_1) p(\Theta, x_1) d\Theta dx_1}, \tag{9}$$

with $k_\theta = (2\pi)^{\frac{-n_y T}{2}} |Q_v|^{-\frac{1}{2}} \left|\bar{Q}_v\right|^{-\frac{1}{2}} \left|\bar{Q}_w\right|^{-\frac{1}{2}} |H|^{\frac{1}{2}}$, *and* $\Sigma = \bar{Q}_v + \bar{C}\bar{G}\bar{C}^\top$.

Proof. Using Bayes' rule, $p(\Theta, x_1 \mid y_{1:T})$ can be factorized as

$$p(\Theta, x_1 \mid y_{1:T}) = \frac{p(y_{1:T} \mid \Theta, x_1) p(\Theta, x_1)}{\int p(y_{1:T} \mid \Theta, x_1) p(\Theta, x_1) d\Theta dx_1}. \tag{10}$$

By substituting (5a) into (5b), we obtain the following expression of the output equation:

$$\begin{aligned} Y &= \bar{C}(\bar{A}x_1 + \bar{B}U_1) + \bar{D}U_2 + \bar{C}GW + V, \\ &= \bar{C}b_2 + \bar{D}U_2 + \bar{C}GW + V. \end{aligned} \tag{11}$$

From (11) and independence of W and V, it follows that

$$\begin{aligned} p(y_{2:T} \mid \Theta, x_1) &= \mathcal{N}(Y; \bar{C}b_2 + \bar{D}U_2, \Sigma) \\ &= (2\pi)^{\frac{-n_y(T-1)}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}\left[(b_1 - \bar{C}b_2)^\top \Sigma^{-1} (b_1 - \bar{C}b_2)\right]}, \end{aligned} \tag{12}$$

where we recall, $b_1 = Y - \bar{D}U_2$ and $\Sigma = \bar{Q}_v + \bar{C}G\bar{Q}_w G^\top \bar{C}^\top = \bar{Q}_v + \bar{C}\bar{G}\bar{C}^\top$. From Weinstein–Aronszajn determinant identity and properties of determinants, it follows that

$$\begin{aligned} |\Sigma| &= \left|\bar{Q}_v + \bar{C}\bar{G}\bar{C}^\top\right| = \left|\bar{Q}_v\right| \left|I_{(T-1)n_y} + \bar{Q}_v^{-1}\bar{C}\bar{G}\bar{C}^\top\right| \\ &= \left|\bar{Q}_v\right| \left|I_{(T-1)n_x} + \bar{C}^\top \bar{Q}_v^{-1}\bar{C}\bar{G}\right| = \left|\bar{Q}_v\right| \left|(\bar{G}^{-1} + \bar{C}^\top \bar{Q}_v^{-1}\bar{C})\bar{G}\right| \\ &= \left|\bar{Q}_v\right| \left|(\bar{G}^{-1} + \bar{C}^\top \bar{Q}_v^{-1}\bar{C})\right| \left|\bar{G}\right| \\ &= \left|\bar{Q}_v\right| \left|H^{-1}\right| \left|\bar{Q}_w\right|. \end{aligned} \tag{13}$$

From (12) and (13), the term $p(y_{1:T} \mid \Theta, x_1)$ appearing in the numerator of (10) can be written as

$$\begin{aligned} &p(y_{1:T} \mid \Theta, x_1) \\ &= \mathcal{N}(y_1; Cx_1 + Du_1, Q_v) \times \mathcal{N}(Y; \bar{C}b_2 + \bar{D}U_2, \Sigma) \\ &= (2\pi)^{\frac{-n_y T}{2}} |Q_v|^{-\frac{1}{2}} |\Sigma|^{-\frac{1}{2}} \times \\ &\quad e^{-\frac{1}{2}\left[(b_1 - \bar{C}b_2)^\top \Sigma^{-1} (b_1 - \bar{C}b_2) + (y_1 - Cx_1 - Du_1)^\top Q_v^{-1} (y_1 - Cx_1 - Du_1)\right]} \\ &= (2\pi)^{\frac{-n_y T}{2}} |Q_v|^{-\frac{1}{2}} \left|\bar{Q}_v\right|^{-\frac{1}{2}} \left|\bar{Q}_w\right|^{-\frac{1}{2}} |H|^{\frac{1}{2}} \times \\ &\quad e^{-\frac{1}{2}\left[(b_1 - \bar{C}b_2)^\top \Sigma^{-1} (b_1 - \bar{C}b_2) + (y_1 - Cx_1 - Du_1)^\top Q_v^{-1} (y_1 - Cx_1 - Du_1)\right]} \\ &= k_\theta \times e^{-\frac{1}{2}\left[(b_1 - \bar{C}b_2)^\top \Sigma^{-1} (b_1 - \bar{C}b_2) + (y_1 - Cx_1 - Du_1)^\top Q_v^{-1} (y_1 - Cx_1 - Du_1)\right]}. \end{aligned}$$

This completes the proof. □

We remark that the integral in the denominator of (9) can not be solved analytically. In the next subsection, we apply MH-MCMC sampling to approximate the conditional posterior $p(\Theta, x_1 \mid y_{1:T})$ using the results in Proposition 2.

3.3 Approximation of $p(\Theta, x_1 \mid y_{1:T})$ through MCMC

The main idea behind MH-MCMC sampler, outlined in Algorithm 1, is to simulate a Markov Chain whose stationary distribution is equal to the desired marginal posterior $p(\Theta, x_1 \mid y_{1:T})$ given in (9). The initial samples $\Theta[0]$ and $x_1[0]$, a *proposal distribution* $q(\Theta^*, x_1^* \mid \Theta[k], x_1[k])$ and the length of the Markov Chain M (*i.e.*, the number of iterations) have to be specified by the user. At each iteration $k \geq 1$, a proposal (Θ^*, x_1^*) is drawn from the proposal distribution $q(\Theta^*, x_1^* \mid \Theta[k], x_1[k])$ (step 1.1), which is accepted with probability $\mathcal{A}((\Theta^*, x_1^*), (\Theta[k], x_1[k]))$ (steps 1.2 and 1.3). If the proposal (Θ^*, x_1^*) is accepted, we set the next sample $\Theta[k+1]$ to Θ^* and $x_1[k+1]$ to x_1^* (step 1.3.1)), otherwise $\Theta[k+1]$ and $x_1[k+1]$ are set to the current samples

$\Theta[k]$ and $x_1[k]$ (step 1.3.2). The outputs of the algorithm are samples $\{\Theta[k], x_1[k]\}_{k=1}^{M}$, that can be proven to be generated by a Markov chain with stationary distribution $p(\Theta, x_1 \mid y_{1:T})$ (see [14]). Algorithm 1 requires to compute the acceptance probability $\mathcal{A}((\Theta^*, x_1^*), (\Theta[k], x_1[k]))$ at step 1.2, which depends on the ratio of the distributions $\frac{p(\Theta^*, x_1^* | y_{1:T})}{p(\Theta[k], x_1[k] | y_{1:T})}$.

From Proposition 2 (eq. (9)), it follows that, for given pairs (Θ^*, x_1^*) and $(\Theta[k], x_1[k])$, the ratio $\frac{p(\Theta^*, x_1^* | y_{1:T})}{p(\Theta[k], x_1[k] | y_{1:T})}$ is given by

$$\frac{p(\Theta^*, x_1^* \mid y_{1:T})}{p(\Theta[k], x_1[k] \mid y_{1:T})} = \frac{e^{-\frac{1}{2}\left[(b_1^* - \bar{C}^* b_2^*)^\top (\Sigma^*)^{-1} (b_1^* - \bar{C}^* b_2^*)\right]}}{e^{-\frac{1}{2}\left[(b_1[k] - \bar{C}[k] b_2[k])^\top (\Sigma[k])^{-1} (b_1[k] - \bar{C}[k] b_2[k])\right]}} \times \frac{e^{-\frac{1}{2}\left[(y_1 - C^* x_1^* - D^* u_1)^\top Q_v^{-1} (y_1 - C^* x_1^* - D^* u_1)\right]}}{e^{-\frac{1}{2}\left[(y_1 - C[k] x_1[k] - D[k] u_1)^\top Q_v^{-1} (y_1 - C[k] x_1[k] - D[k] u_1)\right]}} \times \frac{p(\Theta^*)}{p(\Theta[k])} \frac{|H^*|^{\frac{1}{2}}}{|H[k]|^{\frac{1}{2}}}.$$

Note that, using the samples generated by Algorithm 1, we can also generate samples of the output signal for a new set of inputs u and thus, statistical properties of the estimated output can be inferred numerically.

Remark 1. *The main tuning parameter in MH-MCMC sampling algorithm is the proposal distribution $q(\cdot)$. If the chosen proposal distribution has small-variance, the proposal samples explore the space slowly, with slow convergence of the stationary distribution of the Markov chain to the target. On the other hand, for high-variance proposal distributions, the proposal samples are likely to belong to regions with low probability density leading to a very low acceptance rate and again convergence to the target distribution can be slow. Determining the best proposal for a specific target distribution is a very challenging problem in MCMC methods. In the case studies discussed in Section 5, we use an isotropic Gaussian proposal with variance chosen via cross-validation.*

4 Maximum-A-Posteriori (MAP) estimate

Once the joint posterior distribution $p(\Theta, x_{1:T} | y_{1:T})$ is computed, the *Maximum-A-Posteriori* (MAP) estimate of the parameters $(\Theta, x_{1:T})$ can be also derived. The samples $\{\Theta[k], x_1[k]\}_{k=1}^{M}$ obtained by running MH-MCMC Algorithm 1 can be used to approximate the MAP estimate which is the argument solving the optimization problem:

$$\begin{aligned} \max_{\Theta, x_1, X} p(\Theta, x_1, X | y_{1:T}) &= \max_{\Theta, x_1} p(\Theta, x_1 \mid y_{1:T}) p(X | \Theta, x_1, y_{1:T}) \\ &\approx \max_{\{\Theta[k], x_1[k]\}_{k=1}^{M}} p(\Theta[k], x_1[k] \mid y_{1:T}) \max_X p(X | \Theta[k], x_1[k], y_{1:T}) \end{aligned} \tag{14}$$

For given sample $(\Theta[k], x_1[k])$, the term $\max_X p(X \mid \Theta[k], x_1[k], y_{1:T})$ in (14) can be computed analytically using the expression of the conditional joint posterior of the state sequence $p(X \mid \Theta[k], x_1[k], y_{1:T})$ given in Proposition 1 (eq. (7)). Indeed, this distribution is Gaussian, and its maximum is achieved at its mean $X = \mu[k]$. Thus,

$$\max_X p(X \mid \Theta[k], x_1[k], y_{1:T}) = (2\pi)^{-\frac{(T-1)n_x}{2}} |H[k]|^{-\frac{1}{2}}. \tag{15}$$

Algorithm 1 MH-MCMC sampler for $p(\Theta, x_1 \mid y_{1:T})$

Input: initial conditions $\Theta[0]$, $x_1[0]$; proposal distribution $q(\Theta^*, x_1^* \mid \Theta[k], x_1[k])$, number of iterations M.

1. **for** $k = 0, \cdots, M-1$ **do**
   - 1.1. **draw** proposal (Θ^*, x_1^*) from the distribution $q(\Theta^*, x_1^* \mid \Theta[k], x_1[k])$;
   - 1.2. **set** acceptance probability

     $\mathcal{A}((\Theta^*, x_1^*), (\Theta[k], x_1[k])) \leftarrow$

     $\min\left\{1, \frac{p(\Theta^*, x_1^* \mid y_{1:T}) q(\Theta[k], x_1[k] \mid \Theta^*, x_1^*)}{p(\Theta[k], x_1[k] \mid y_{1:T}) q(\Theta^*, x_1^* \mid \Theta[k], x_1[k])}\right\}$;

   - 1.3. **accept** proposal (Θ^*, x_1^*) with probability $\mathcal{A}((\Theta^*, x_1^*), (\Theta[k], x_1[k]))$;
     - 1.3.1. **if** the proposal is accepted
       set $\Theta[k+1] \leftarrow \Theta^*$; **set** $x_1[k+1] \leftarrow x_1^*$;
     - 1.3.2. **else**
       set $\Theta[k+1] \leftarrow \Theta[k]$; **set** $x_1[k+1] \leftarrow x_1[k]$;
2. **end for**;

Output: Samples $\{\Theta[k], x_1[k]\}_{k=1}^{M}$

By substituting (15) and (9) into (14) and omitting the terms that do not depend on Θ and x_1, we obtain the MAP estimate of the model parameters $\hat{\Theta}$ and initial state $\hat{x}_1$, *i.e.*,

$$\begin{aligned} \hat{\Theta}, \hat{x}_1 &\approx \arg \max_{\{\Theta[k], x_1[k]\}_{k=1}^{M}} p(\Theta[k], x_1[k] \mid y_{1:T}) |H[k]|^{-\frac{1}{2}} \\ &\approx \arg \max_{\{\Theta[k], x_1[k]\}_{k=1}^{M}} |H[k]|^{\frac{-1}{2}} |H[k]|^{\frac{1}{2}} p(\Theta[k], x_1[k]) \times \\ &\quad e^{\frac{-1}{2}\left[(b_1[k] - \bar{C}[k] b_2[k])^\top (\Sigma[k])^{-1} (b_1[k] - \bar{C}[k] b_2[k])\right]} \times \\ &\quad e^{\frac{-1}{2}\left[(y_1 - C[k] x_1[k] - D[k] u_1)^\top Q_v^{-1} (y_1 - C[k] x_1[k] - D[k] u_1)\right]} \\ &\approx \arg \max_{\{\Theta[k], x_1[k]\}_{k=1}^{M}} p(\Theta[k], x_1[k]) \times \\ &\quad e^{\frac{-1}{2}\left[(b_1[k] - \bar{C}[k] b_2[k])^\top (\Sigma[k])^{-1} (b_1[k] - \bar{C}[k] b_2[k])\right]} \times \\ &\quad e^{\frac{-1}{2}\left[(y_1 - C[k] x_1[k] - D[k] u_1)^\top Q_v^{-1} (y_1 - C[k] x_1[k] - D[k] u_1)\right]} \end{aligned} \tag{16}$$

The MAP estimate $\hat{X}$ of the state sequence is then obtained using the MAP estimates $\hat{\Theta}$ and $\hat{x}_1$. Specifically, the MAP estimate $\hat{X}$ is equal to $\mu[k]$, with $\mu[k]$ computed for $\Theta[k] = \hat{\Theta}$ and $x_1[k] = \hat{x}_1$ (see Proposition 1 for the definition of μ).

5 Case studies

The effectiveness of the proposed identification algorithm is shown via a numerical example and through a real laboratory benchmark dataset of a hair dryer process. In both examples, the prior $p(\Theta, x_1)$ over the unknown model matrices and the initial state is set to be Gaussian with unitary covariance matrix $I_{n_\theta + n_x}$. The match between the estimated and the true output is quantified on a noise-free validation data of length N_{val} in terms of the *Best Fit Rate* $\text{BFR}_i = \max\left\{1 - \sqrt{\frac{\sum_{t=1}^{N_{\text{val}}} (y_t^i - \hat{y}_t^i)^2}{\sum_{t=1}^{N_{\text{val}}} (y_t^i - \bar{y}_t^i)^2}}, 0\right\} \times 100\%$, and the *Vari-*

Table 1: Example 1: BFR and VAF on validation data.

	Output 1	Output 2
BFR / VAF	94.52 % / 99.70 %	90.18 % / 99.14 %

ance Accounted For $\text{VAF}_i = \max\left\{1 - \frac{\text{var}(y^i - \hat{y}^i)}{\text{var}(y^i)}, 0\right\} \times 100\%$, criterion defined for each output channel $i = 1, \cdots, n_y$, with $\hat{y}^i$ being the i-th open-loop simulated model output and $\bar{y}^i$ is the sample mean of the i-th output over the validation set. The operator $\text{var}(\cdot)$ denotes the variance of its argument.

5.1 Example 1: example with synthetic data

The training dataset of length $N = 100$ and validation dataset of length $N_{\text{val}} = 200$ are gathered from a (stable) randomly generated state-space model, with $n_y = 2$ outputs, $n_u = 1$ inputs, and $n_x = 4$ state variables. The covariance matrices of the process and measurement noise are $Q_w = (0.01)^2 I_{n_x}$ and $Q_v = (0.3)^2 I_{n_y}$. This corresponds to a *Signal-to-Noise Ratio* $\text{SNR}_i = 10\log\frac{\sum_{t=1}^{N}\left(y_t^i - v_t^i\right)^2}{\sum_{t=1}^{N}(v_t^i)^2}$ of 14.2 dB and 12.2 dB on the first and second output channel respectively, with v_t^i being the noise on the i-th output y_t^i.

The approximated conditional posterior $p(\Theta, x_1 \mid y_{1:T})$ of the model matrices and the initial state is obtained by running Algorithm 1 with $M = 2000$ iterations and randomly generated initial guess $\Theta[0], x_1[0]$. An isotropic Gaussian distribution $q(\Theta^*, x_1^* \mid \Theta[k], x_1[k])$ with mean $(\Theta[k], x_1[k])$ and diagonal covariance matrix $(0.01)^2 I_{n_\theta + n_x}$ is used as a proposal distribution. The first 500 samples of the Markov chain are discarded, according to the *burn-in* strategy. The execution time to run Algorithm 1 is 47 sec. The samples generated by Algorithm 1 are also used to simulate the output signal $\hat{y}$ and thus to numerically approximate its distribution. Fig. 1 shows the mean of the simulated output $\pm$ the standard deviation, along with the true output. From the approximated posterior distribution, the MAP estimates of the model parameters are computed as described in Section 4 and are used to simulate a single trajectory of the output $\hat{y}$. The achieved BFR and VAF measures for the two output channels on the validation dataset are reported in Table 1. The Markov parameters representing the impulse response of the estimated and the true system are shown in Fig. 2. Specifically, the impulse response h_t for output channel 1 is given by $h_0 = D_1$ and $h_t = C_1 A^{t-1} B$ for $t > 0$, with D_1 and C_1 denoting the rows of D and C corresponding to the first output. Similarly, for output channel 2, the impulse response is $h_0 = D_2$ and $h_t = C_2 A^{t-1} B$ for $t > 0$. The obtained results show a good match between the estimated and true output.

5.2 Example 2: laboratory hair dryer process

As a second case study, we consider a real laboratory data set of a hair dryer process [16]. Air is blown through a tube and heated at the inlet. The input is the voltage of the heating device at the inlet and the output is the measured thermocouple voltage representing the outlet air temperature. From the available dataset, the first $N = 300$ samples are used for training and the remaining $N_{\text{val}} = 300$ samples are used for validation. We choose $n_x = 3$ as the dimension of the state-space model. The process and measurement noise covariances are set to $Q_w = (0.01)^2 I_{n_x}$ and $Q_v = (0.01)^2$.

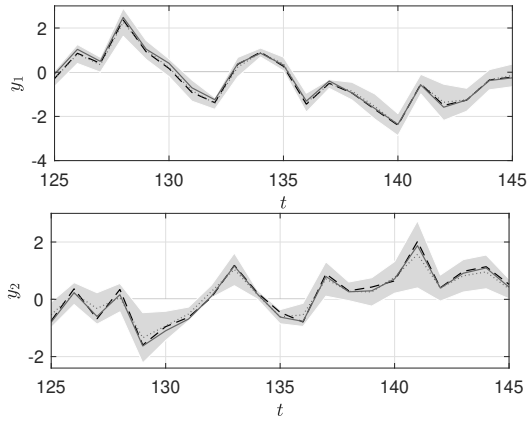

Fig. 1: Example 1: true (dashed) *vs* simulated output with MAP estimates (solid); mean value of the simulated output (dotted) $\pm$ 2 standard deviation (shaded gray region).

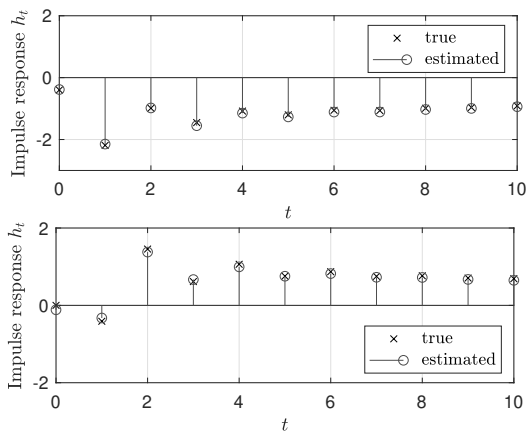

Fig. 2: Example 1: true *vs* estimated impulse response for output 1 (top) and output 2 (bottom).

Algorithm 1 is run for $M = 5000$ iterations to approximate the conditional posterior $p(\Theta, x_1 \mid y_{1:T})$ of the model matrices and the initial state, with a randomly generated initial guess $\Theta[0], x_1[0]$. The proposal distribution $q(\Theta^*, x_1^* \mid \Theta[k], x_1[k])$ is Gaussian with mean $(\Theta[k], x_1[k])$ and a diagonal covariance matrix $(0.0033)^2 I_{n_\theta + n_x}$. According to the *burn-in* strategy, the first 1000 samples of the Markov chain are discarded. The execution time to run Algorithm 1 is 925.7 seconds. Based on the samples generated by Algorithm 1, the MAP estimates of the unknown model parameters are computed and used to simulate the output $\hat{y}$ of the identified LTI-SS model. Figure 3 depicts the simulated output $\hat{y}_t$ w.r.t to the measured output y_t of the process. The distribution of the simulated output is also computed based on the samples generated by Algorithm 1. The mean $\pm$ the standard deviation is plotted in Fig. 3. The obtained BFR and VAF measures are 83.7% and 97.5%, respectively, which indicate a good match between estimated and true data.

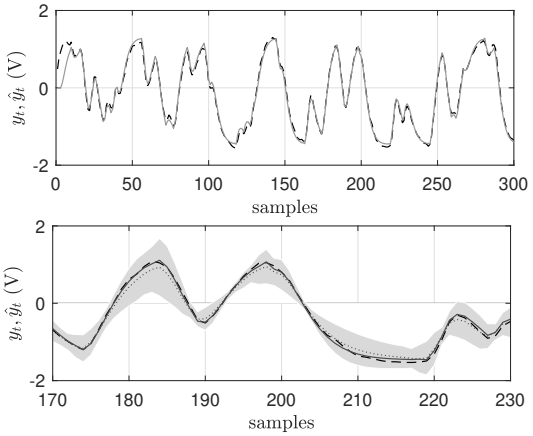

Fig. 3: Example 2: Top panel: true (dashed) *vs* simulated output with MAP estimates (solid). Bottom panel: true (dashed) *vs* simulated (solid), mean value of the simulated output (dotted) $\pm$ 2 standard deviation (shaded gray region).

6 Conclusions

In this paper, we have addressed maximum-a-posteriori estimation of LTI-SS models. A Rao-Blackwellized MCMC sampling algorithm has been proposed to approximate the joint posterior distribution of the model parameters and the unknown state sequence, and thus to compute the MAP estimate. The main advantage of using a Rao-Blackwellized algorithm is that the sampling space is greatly reduced, requiring the MCMC algorithm to sample only over the model parameters and the initial state. Future research direction involves developing a recursive version of the presented algorithm by propagating the posterior distributions of the model parameters over time using particle filters.

References

[1] Ho, B., and Kalman, R. E., 1965. "Effective construction of linear state-variable models from input/output data". In Proc. 3rd Annual Allerton Conf. on Circuit and System Theory, pp. 449–459.

[2] Verhaegen, M., and Dewilde, P., 1992. "Subspace model identification part 1. The output-error state-space model identification class of algorithms". *International Journal of Control,* **56**(5), pp. 1187–1210.

[3] Overschee, P., and Moor, B., 1994. "N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems". *Automatica,* **30**(1), pp. 75 – 93.

[4] Ljung, L., 2003. "Aspects and experiences of user choices in subspace identification method". In Proc. of the 13th IFAC Symposium on System Identification.

[5] Bergboer, N., Verdult, V., and Verhaegen, M., 2002. "An efficient implementation of maximum likelihood identification of LTI state-space models by local gradient search". In Proc. of Conf. on Decision and Control.

[6] Ninness, B., and Henriksen, S., 2010. "Bayesian system identification via Markov chain Monte Carlo techniques". *Automatica,* **46**(1), pp. 40 – 51.

[7] Wills, A., Schön, T. B., Lindsten, F., and Ninness, B., 2012. "Estimation of linear systems using a gibbs sampler". In Proc. of the 16th IFAC Symposium on System Identification.

[8] Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J., and Chopin, N., 2015. "On particle methods for parameter estimation in state-space models". *Statist. Sci.,* **30**(3), 08, pp. 328–351.

[9] Schön, T. B., Lindsten, F., Dahlin, J., Wågberg, J., Naesseth, C. A., Svensson, A., and Dai, L., 2015. "Sequential monte carlo methods for system identification". In Proc. of the 17th IFAC Symposium on System Identification.

[10] Schön, T. B., Svensson, A., Murray, L., and Lindsten, F., 2018. "Probabilistic learning of nonlinear dynamical systems using sequential Monte Carlo". *Mechanical Systems and Signal Processing,* **104**, pp. 866–883.

[11] Roweis, S., and Ghahramani, Z., 2001. *Learning Nonlinear Dynamical Systems Using the Expectation–Maximization Algorithm*. John Wiley & Sons, Ltd, ch. 6, pp. 175–220.

[12] Wan, E. A., and Nelson, A. T., 2001. *Dual Extended Kalman Filter Methods*. John Wiley & Sons, Ltd, ch. 5, pp. 123–173.

[13] Casella, G., and Robert, C., 1996. "Rao-Blackwellisation of sampling schemes". *Biometrika,* **83**(1), pp. 81–94.

[14] Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I., 2003. "An introduction to MCMC for machine learning". *Machine Learning,* **50**(1), Jan, pp. 5–43.

[15] Piga, D., Bemporad, A., and Benavoli, A., 2020. "Rao-Blackwellized sampling for batch and recursive Bayesian inference of Piecewise Affine models". *Automatica,* **117**.

[16] Moor, B. D., Gersem, P. D., Schutter, B. D., and Favoreel, W., 1997. DaISy: Database for the Identification of Systems. Tech. rep.

A Gaussian Identities

Given the vectors $Y \in \mathbb{R}^N$, $X \in \mathbb{R}^{n_x}$, and $\hat{X} \in \mathbb{R}^{n_x}$, a matrix $C \in \mathbb{R}^{N \times n_x}$, and two symmetric positive definite matrices $Q_v \in \mathbb{R}^{N \times N}$, $Q_x \in \mathbb{R}^{n_x \times n_x}$, the following identity holds:

$$
\begin{aligned}
&e^{-\frac{1}{2}\left[(Y-CX)^\top Q_v^{-1}(Y-CX)+\left(X-\hat{X}\right)^\top Q_x^{-1}\left(X-\hat{X}\right)\right]} \\
&= (2\pi)^{\frac{n_x}{2}} |H|^{\frac{1}{2}} \mathcal{N}(X;\mu,H) e^{-\frac{1}{2}\left[Y^\top Q_v^{-1} Y+\hat{X}^\top Q_x^{-1}\hat{X}-F^\top \mu\right]}, \quad (17)
\end{aligned}
$$

with $H = \left(C^\top Q_v^{-1} C + Q_x^{-1}\right)^{-1}$, $F = \left(C^\top Q_v^{-1} Y + Q_x^{-1}\hat{X}\right)$, $\mu = HF$.

Proof. Let us consider the exponent in (17) up to the constant $-\frac{1}{2}$ and let us complete the square as follows

$$
\begin{aligned}
&(Y-CX)^\top Q_v^{-1}(Y-CX) + (X-\hat{X})^\top Q_x^{-1}(X-\hat{X}) \\
&= X^\top H^{-1} X - 2X^\top H^{-1} HF + Y^\top Q_v^{-1} Y + \hat{X}^\top Q_x^{-1}\hat{X} \\
&= (X-\mu)^\top H^{-1}(X-\mu) - F^\top \mu + Y^\top Q_v^{-1} Y + \hat{X}^\top Q_x^{-1}\hat{X}.
\end{aligned}
$$

Using the above equation, Eq. (17) can be rewritten as

$$
\begin{aligned}
&e^{-\frac{1}{2}(X-\mu)^\top H^{-1}(X-\mu)} e^{-\frac{1}{2}\left(Y^\top Q_v^{-1} Y+\hat{X}^\top Q_x^{-1}\hat{X}-F^\top\mu\right)} \\
&= (2\pi)^{\frac{n_x}{2}} |H|^{\frac{1}{2}} \mathcal{N}(X;\mu,H) e^{-\frac{1}{2}\left[Y^\top Q_v^{-1} Y+\hat{X}^\top Q_x^{-1}\hat{X}-F^\top\mu\right]}.
\end{aligned}
$$

This proves the identity in (17).