<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *International Journal of Control*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

Valenzuela, P E., Schön, T B., Rojas, C R. (2019)
On model order priors for Bayesian identification of SISO linear systems
International Journal of Control, 92(7): 1645-1661
<https://doi.org/10.1080/00207179.2017.1406147>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-255303>

To appear in the *International Journal of Control*
Vol. 00, No. 00, Month 20XX, 1–33

On model order priors for Bayesian identification of SISO linear systems

Patricio E. Valenzuela^{a*}, Thomas B. Schön^{b†} and Cristian R. Rojas^{a‡}

^a*Department of Automatic Control and ACCESS Linnaeus Center, School of Electrical Engineering, KTH
Royal Institute of Technology, SE-100 44 Stockholm, Sweden.*

^b*Department of Information Technology, Uppsala University, SE-751 05 Uppsala, Sweden.*

(November 1st, 2017. 14392 words.)

*Corresponding author. Email: pva@kth.se

†Email: thomas.schon@it.uu.se

‡Email: crro@kth.se

This research was financially supported by the Swedish Foundation for Strategic Research (SSF) via the project ASSEMBLE and by the Swedish Research Council under grants 2016-06079 (NewLEADS), 621-2015-4393, 621-2013-5524, 621-2011-5890 and 621-2009-4017.

A method for the identification of single input, single output linear systems is presented. The method employs a Bayesian approach to compute the posterior distribution of the model parameters given the data set. Since this distribution is often unavailable in closed form, a Metropolis Hastings algorithm is implemented to draw samples from it. To implement the sampler, the inclusion of prior information regarding the model order of the identified system is discussed. As one of the main contributions of this work, a prior over the Hankel singular values of the model is imposed. Numerical examples illustrate the method.

Keywords: System identification; Bayesian estimation; Metropolis Hastings sampler; Model order prior.

1. Introduction

System identification is concerned with the problem of estimating a dynamical model given a set of measurements (Ljung, 1999; Söderström & Stoica, 1988). Most of the available methods in system identification can be categorized into one of the following three groups: (i) prediction error/maximum likelihood methods, where the model is obtained as the minimizer of a scalar cost function of the data set and the parameterized model (Ljung, 1999; Schön, Wills, & Ninness, 2011); (ii) Bayesian methods, where the model is estimated based on its posterior distribution, which is computed given the data set and a prior distribution over the model parameters (Fox, Sudderth, Jordan, & Willsky, 2011; Hjalmarsson & Gustafsson, 1995; Wills, Schön, Lindsten, & Ninness, 2012); and (iii) instrumental variables and subspace methods (Katayama, 2006; Söderström & Stoica, 1981, 1983; Viberg, 1995), which do not optimize a cost function to obtain the model.

1.1 Contribution

In this work we introduce a new approach for Bayesian identification of single input, single output (SISO) linear dynamical systems, where priors over the model order are imposed. The method computes samples that are approximately distributed according to the posterior probability density function (pdf) of the model parameters. To this end, a Metropolis Hastings sampler (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Robert & Casella, 2004) is implemented. An advantage of using a Metropolis Hastings sampler compared to other methods is that it gives more freedom to choose the prior distribution of the model parameters, which allows to define priors that do not necessarily lead to posterior distributions available in closed form. We exploit this feature by imposing a prior distribution over the Hankel singular values of the model, to account for prior knowledge or belief about the order of the model.

We remark that the aim of the proposed method is to provide a fully Bayesian framework to estimate dynamical models and encode prior information about the system in the estimation. Hence, we do not aim to prove that the technique outperforms the results achieved by existing methods, but instead to place in context the performance of the proposed approach with methods such as the prediction error method and subspace techniques. We also note that related issues, e.g. choice of the prior and proposal distributions and computational complexity, are only addressed by numerical simulations. Due to its extent, a thorough analysis of the proposed technique is beyond the scope of this article and it is left as future work.

1.2 Related work

The existing Bayesian methods for system identification are mainly implemented using the Gibbs sampler (Carter & Kohn, 1994; Fox et al., 2011; Geman & Geman, 1984; Lindsten, Jordan, & Schön, 2014; Wills et al., 2012). The Gibbs sampler is a particular implementation of the Metropolis Hastings algorithm (Robert & Casella, 2004), and hence it computes samples that are approximately

distributed according to the posterior pdf. However, the implementation of the Gibbs sampler requires that the conditional marginal distributions of the posterior pdf are available in closed form, and that they can be used to generate samples. This property is commonly ensured by restricting the prior to the set of conjugate distributions of the likelihood (Robert, 2007, Section 3.3). Prior distributions outside this set are often intractable for the Gibbs sampler, and other methods are needed. For example, this is the case if the user specifies a prior distribution for the order of the dynamical system (Robert & Casella, 2004, Chapter 11).

In the case of model order selection, a Bayesian framework has been considered in the literature (Andrieu, De Freitas, & Doucet, 1999; Cappé, Robert, & Rydén, 2003; Claeskens & Hjort, 2008). The main method employed is the reversible jump Markov chain Monte Carlo (RJMCMC) sampler (P. J. Green, 1995; Hastie & Green, 2012). The implementation of this algorithm requires the definition of a differentiable bijective transformation between model sets of different order, which is achieved by introducing auxiliary variables. While this approach is relatively easy to implement for small models, its complexity increases rapidly with the model order.

We remark that the Bayesian approach to system identification is not a recent development. In this line, Peterka (1981) argued already in 1981 that Bayesian techniques can provide an attractive solution to the system identification problem. More recently, Ninness & Henriksen (2010) presented a Markov chain Monte Carlo method to implement the Bayesian framework in system identification, where a Metropolis Hastings sampler is employed to compute posterior distributions. On the other hand, the role of sparse estimation in Bayesian system identification is gaining interest in the recent years. Indeed, in Prando, Chiuso, & Pillonetto (2014) the role of rank penalties in Bayesian system identification is discussed. The model order of the estimated system is implicitly taken into account by including a regularization term associated with a weighted quadratic function of the Hankel matrix of the estimated impulse response coefficients. The purpose of including such a penalty is to enforce sparseness on the singular values of the Hankel matrix.

1.3 Structure of the article

The remainder of this work is organized as follows. Section 2 introduces the problem formulation. Section 3 discusses the Metropolis Hastings sampler. The specification of prior distributions over the parameter set is considered in Section 4. The implementation of a random walk over the parameter space is presented in Section 5. Section 6 illustrates the method with a numerical example. Finally, Section 7 concludes this article.

Notation: $\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^{n \times m}$, and $\mathbb{N}$ stand for the real set, the set of n -dimensional vectors with real entries, the set of $n \times m$ matrices with real entries, and the set of natural numbers, respectively. $\mathbb{R}_+^n$ contains the n -dimensional vectors with nonnegative real entries. The complex set is denoted by $\mathbb{C}$ and $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$. Given $z \in \mathbb{C}$, $|z|$ denotes its magnitude. The i -th eigenvalue of a matrix $A \in \mathbb{R}^{n \times n}$ is written $\lambda_i(A)$, where $|\lambda_i(A)| \geq |\lambda_{i+1}(A)|$. The transpose operator is denoted by $(\cdot)^\top$. $\mathcal{N}(x, \Sigma)$ is the normal distribution with mean $x \in \mathbb{R}^n$, and covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$, and hence Σ is symmetric and positive definite. Given $a, b \in \mathbb{R}$, $a < b$, $\mathcal{U}[a, b]$ stands for the uniform distribution over the interval $[a, b]$. A subscript is sometimes added to the probability density function to emphasize the random variable over which it is defined. The time shift operator is denoted by q , i.e., $q x_t = x_{t+1}$. Given a scalar, rational function G without poles on the circle $\{z \in \mathbb{C} : |z| = 1\}$, we define its 2-norm as

$$\|G\|_2 := \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} |G(e^{i\omega})|^2 d\omega}. \quad (1)$$

2. Problem formulation

Consider the discrete time, single input, single output (SISO) linear time invariant system defined as

$$y_t = G_o(q) u_t + H_o(q) e_t, \quad t \in \mathbb{N}, \quad (2)$$

where G_o, H_o are scalar real rational functions in the time shift operator q , $u_t \in \mathbb{R}$ is the input, $e_t \in \mathbb{R}$ is a white noise process with pdf p_e , and $y_t \in \mathbb{R}$ is the measured output. In addition, we consider the following:

Assumption 1: *The pair (G_o, H_o) of scalar rational functions in q has all its poles in the set $\mathbb{D}$. Also, H_o^{-1} has all its poles in $\mathbb{D}$ and*

$$\lim_{q \rightarrow \infty} H_o(q) = 1. \quad (3)$$

Assumption 1 is required to guarantee that the transients due to initial conditions decay asymptotically to zero, thus permitting their outputs to reach asymptotic stationarity (Ljung, 1999; Priestley, 1981).

Our aim is to estimate the parameters $\theta \in \Theta \subseteq \mathbb{R}^{n_\theta}$ in the model structure

$$y_t = G_\theta(q) u_t + H_\theta(q) e_t^\theta, \quad (4)$$

given the observed data set $\mathbf{Z}_N := \{(y_t, u_t)\}_{t=1}^N$. Note that the white noise sequence e_t^θ is parameterized through its pdf, denoted by p_e^θ . In the following, we assume that the model structure (4) satisfies Assumption 1 for all $\theta \in \Theta$ (with (G_θ, H_θ) instead of (G_o, H_o)). In addition, we assume that there exists a parameter $\theta_0 \in \Theta$ such that the model (4) describes the system (2) when $\theta = \theta_0$, i.e., there is no undermodelling (Ljung, 1999).

In this work we are interested in exploring a Bayesian approach to estimate $\hat{\theta}_N \in \Theta$ describing the system (2). In this setting, we require the specification of a prior distribution over Θ , which reflects the available knowledge (or belief) that the user has about θ_0 before observing the data set $\mathbf{Z}_N$. In the following, we denote by p_Θ the pdf associated with the prior distribution over Θ .

Given p_Θ and the data set $\mathbf{Z}_N$, the posterior distribution over Θ can be computed using Bayes' rule as

$$p(\theta|\mathbf{Z}_N) = \frac{1}{K} p(y_{1:N}|u_{1:N}, \theta) p_\Theta(\theta), \quad (5)$$

where

$$K := \int_{\Theta} p(y_{1:N}|u_{1:N}, \theta) p_\Theta(\theta) d\theta \quad (6)$$

is a normalizing constant, and $x_{1:N} := (x_1, \dots, x_N)$. In (5), the expression $p(y_{1:N}|u_{1:N}, \theta)$ is the likelihood that the input $u_{1:N}$, and the system (2) with parameter θ have generated the output $y_{1:N}$.

Once the pdf (5) is computed, there are several options to obtain an estimate of the model parameters $\hat{\theta}_N$. The most commonly employed ones are

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta \in \Theta} p(\theta|\mathbf{Z}_N), \quad (7)$$

referred to as the maximum a posteriori (MAP) estimator, and

$$\hat{\theta}_m = \mathbf{E} \{ \theta | \mathbf{Z}_N \}, \quad (8)$$

the conditional mean of the model parameters given the observed data set $\mathbf{Z}_N$.

Remark 1: We note that the maximum a posteriori estimator (7) can be obtained without explicit knowledge of the posterior pdf. Indeed, the optimization in (7) can be performed directly in a numerical fashion. However, this approach only delivers the maximizer of $p(\theta | \mathbf{Z}_N)$ and not the uncertainty of this estimator.

The key step to estimate the model parameters in a Bayesian setting is the computation of (5). Unfortunately, obtaining such an expression in closed form is often difficult or intractable in practice, which has been the main limitation of this approach for a long time.

In an effort to compute expressions like (5), Markov chain Monte Carlo (MCMC) algorithms for approximating pdfs have received considerable attention, and methods to obtain estimates for the posterior pdf (5) have been developed and analyzed. In particular, the Gibbs sampler has been mainly employed to generate an estimate of the posterior distribution (Carter & Kohn, 1994; Fox et al., 2011; Wills et al., 2012). The main limitation of this approach is that the prior has to be carefully specified, since it must allow to sample from conditional pdfs derived from the posterior distribution. To achieve this, the prior is usually chosen as the conjugate of the likelihood out of mathematical convenience (Schön & Lindsten, 2016). In case one wants to specify a prior that is not conjugate to the likelihood, then the Gibbs sampler approach in Fox et al. (2011); Wills et al. (2012) cannot be directly employed to approximate the posterior distribution. For example, this is typically the case if one specifies a prior over the order of the model structure (4).

An alternative to the Gibbs sampler is provided by its root, the Metropolis Hastings method, where the assumption of conjugate distributions in Wills et al. (2012) can be relaxed. In the next section we describe the Metropolis Hastings algorithm and its application to compute the posterior pdf (5).

3. The Metropolis Hastings sampler

The Metropolis Hastings (MH) sampler is an MCMC method that generates samples approximately distributed according to a prescribed pdf π , called the target distribution. The samples are obtained by running a Markov chain with appropriate transition kernels, ensuring that π is the stationary pdf associated with the chain (Schön & Lindsten, 2016). In the context of this work, the target distribution is $\pi(\theta) = p(\theta | \mathbf{Z}_N)$, i.e., the posterior pdf of the parameter vector.

The idea behind the MH sampler is as follows: starting from an initial state $\theta^{(1)} \in \Theta$, we sample a new candidate $\theta^* \in \Theta$ from a user defined conditional pdf $f(\theta | \theta^{(1)})$, called the proposal distribution. Then, with probability

$$a(\theta^*, \theta^{(1)}) := \min \left(1, \frac{\pi(\theta^*) f(\theta^{(1)} | \theta^*)}{\pi(\theta^{(1)}) f(\theta^* | \theta^{(1)})} \right) \quad (9)$$

we define the next state as $\theta^{(2)} = \theta^*$, otherwise we keep the previous state of the chain, i.e., $\theta^{(2)} = \theta^{(1)}$. This procedure is repeated until the required number of samples have been generated.

Algorithm 1 presents the details of the MH sampler. Under the assumption that the support of the proposal pdf f includes the support of the target distribution π , it can be proved that Algorithm 1 has π as a stationary distribution of the chain (Robert & Casella, 2004).

By using $\pi(\theta) = p(\theta | \mathbf{Z}_N)$ and (5), we see that the acceptance probability for the MH sampler to

Algorithm 1 Metropolis Hastings samplerINPUTS: Target pdf π and proposal distribution f .OUTPUT: $\{\theta^{(i)}\}_{i=1}^{N_s}$ (samples distributed according to π).

-
- 1: Set the initial state of the Markov chain $\theta^{(1)}$.
 - 2: **for** $i = 1$ to N_s **do**
 - 3: Sample $\theta^* \sim f(\theta|\theta^{(i)})$.
 - 4: Sample $u \sim \mathcal{U}[0, 1]$.
 - 5: Compute the acceptance probability
-

$$a(\theta^*, \theta^{(i)}) = \min \left(1, \frac{\pi(\theta^*) f(\theta^{(i)}|\theta^*)}{\pi(\theta^{(i)}) f(\theta^*|\theta^{(i)})} \right). \quad (10)$$

- 6: Set the next state $\theta^{(i+1)}$ of the Markov chain according to

$$\theta^{(i+1)} = \begin{cases} \theta^*, & u \leq a(\theta^*, \theta^{(i)}) \\ \theta^{(i)}, & \text{otherwise.} \end{cases} \quad (11)$$

- 7: **end for**

generate samples from $p(\theta|\mathbf{Z}_N)$ is given by

$$a(\theta^*, \theta^{(i)}) = \min \left(1, \frac{p(y_{1:N}|u_{1:N}, \theta^*) p_{\Theta}(\theta^*) f(\theta^{(i)}|\theta^*)}{p(y_{1:N}|u_{1:N}, \theta^{(i)}) p_{\Theta}(\theta^{(i)}) f(\theta^*|\theta^{(i)})} \right). \quad (12)$$

From (12), we see that the acceptance probability only needs the knowledge of the posterior pdf up to a normalization constant.

On the other hand, the likelihood $p(y_{1:N}|u_{1:N}, \theta)$ can be rewritten as

$$p(y_{1:N}|u_{1:N}, \theta) = p(y_1|u_1, \theta) \prod_{t=2}^N p(y_t|y_{1:t-1}, u_{1:t}, \theta), \quad (13)$$

where $p(y_t|y_{1:t-1}, u_{1:t}, \theta)$ denotes the pdf of the measurement y_t given the tuple $(y_{1:t-1}, u_{1:t}, \theta)$. The pdf $p(y_t|y_{1:t-1}, u_{1:t}, \theta)$ can be obtained from the model (4) as

$$p(y_t|y_{1:t-1}, u_{1:t}, \theta) = p_e^{\theta}(\varepsilon_t(\theta)), \quad (14)$$

where

$$\varepsilon_t(\theta) := H_{\theta}^{-1}(q)(y_t - G_{\theta}(q)u_t) \quad (15)$$

is the one step ahead prediction error associated with the model (4) at time t . Inserting (13)–(15) into (12) we finally obtain

$$a(\theta^*, \theta^{(i)}) = \min \left(1, \frac{p(y_1|u_1, \theta^*) p_{\Theta}(\theta^*) f(\theta^{(i)}|\theta^*) \prod_{t=2}^N p_e^{\theta^*}(\varepsilon_t(\theta^*))}{p(y_1|u_1, \theta^{(i)}) p_{\Theta}(\theta^{(i)}) f(\theta^*|\theta^{(i)}) \prod_{t=2}^N p_e^{\theta^{(i)}}(\varepsilon_t(\theta^{(i)}))} \right). \quad (16)$$

From (16) we see that the acceptance probability can be computed using the model structure (4) for every $\theta \in \Theta$. However, to implement the MH sampler over the parameter set Θ we need to address two aspects:

- (1) the specification of the prior p_{Θ} over Θ , and
- (2) the implementation of a “random walk over Θ ” using the proposal distribution f .

The first point will be addressed in the next section, while the second point will be discussed in Section 5.

4. Specifying priors over the parameter set

As mentioned in Section 2, the prior pdf p_Θ reflects the available knowledge (or belief) the user has before the data set $\mathbf{Z}_N$ is obtained. In the literature, most of the methods using Gibbs and MH samplers specify an explicit prior over Θ (Dahlin, Lindsten, Schön, & Wills, 2012; Fox et al., 2011; Wills et al., 2012), i.e., the pdf p_Θ is an explicit function of θ . In addition, the requirement of a conjugate prior is generally imposed, to analytically compute posterior distributions. In case one wants to use more general priors, then the methods in Dahlin et al. (2012); Fox et al. (2011); Wills et al. (2012) cannot be employed, and hence other alternatives are needed.

4.1 Model order priors

In this work we are interested in exploring more general priors than those introduced by Fox et al. (2011); Wills et al. (2012). Extending the class of priors will allow us to encode particular properties of the model structure in the problem.

Here we are interested in defining priors over the order of the model structure (4). To proceed, we assume that the model set

$$\mathcal{M} := \left\{ (G_\theta, H_\theta, p_e^\theta) : \theta \in \Theta \right\} \quad (17)$$

contains all rational functions in q for the pair (G_θ, H_θ) that are of order at most $n_{\mathcal{M}}$, where $n_{\mathcal{M}}$ is a fixed number. This assumption is similar to the one introduced in Wills et al. (2012) in the sense that prior knowledge about an upper bound on the order of the system is required.

At this point we may consider employing one of the available methods for Bayesian model order selection (P. J. Green, 1995). However, the assumption of a differential bijective mapping between model structures of different order limits its applicability in the present problem. In particular, the main difficulty of applying the method of P. J. Green (1995) is the requirement of a closed form expression for the bijective mapping and its Jacobian, which are needed in order to run an MH sampler targeting the posterior distribution. To circumvent this issue, we work instead with priors defined over the Hankel singular values of the rational function G_θ . The definition of Hankel singular values is given next (M. Green & Limebeer, 2012):

Definition 1: Consider the discrete time, stable, linear state space system

$$x_{t+1} = A x_t + B u_t, \quad (18a)$$

$$y_t = C x_t + D u_t, \quad (18b)$$

with $A \in \mathbb{R}^{n_x \times n_x}$, $B \in \mathbb{R}^{n_x \times n_u}$, $C \in \mathbb{R}^{n_y \times n_x}$, $D \in \mathbb{R}^{n_y \times n_u}$, and the matrices $P, Q \in \mathbb{R}^{n_x \times n_x}$ which are the positive definite solutions to the corresponding Lyapunov equations

$$APA^\top - P + BB^\top = 0, \quad (19a)$$

$$A^\top QA - Q + C^\top C = 0. \quad (19b)$$

The values $\sigma_i = \sqrt{\lambda_i(PQ)}$, $i \in \{1, \dots, n_x\}$, are called the *Hankel singular values* of the system (18).

Since the Hankel singular values are closely related to the order of a model, they have been widely employed in model order reduction (M. Green & Limebeer, 2012). A Hankel singular value close to zero indicates that the order of the model can be reduced by one with almost negligible effect on the model properties (e.g., its bandwidth, step response, DC gain, etc.). Thus, the specification of a prior over the Hankel singular values of G_θ reflects prior knowledge (or belief) about the order of the system (2).

To evaluate the prior over the Hankel singular values associated with G_θ , we first write G_θ in the state space form (18), and then we compute the Hankel singular values according to Definition 1. In this manner, a prior pdf over the Hankel singular values of G_θ implicitly accounts for its complexity.

4.2 Improper priors

At this point we need to emphasize that a prior distribution over the Hankel singular values of G_θ is not necessarily normalizable (or proper). A prior pdf p_Θ is said to be *normalizable* (or *proper*) if and only if

$$\int_{\Theta} p_{\Theta}(\theta) d\theta < \infty. \quad (20)$$

If expression (20) does not hold, then p_Θ is said to be an *improper* prior.

In case the prior is proper but not normalized, the MH sampler can still be used, since the computation of the acceptance probability (16) only requires the knowledge of p_Θ up to a normalization constant. On the other hand, if the prior p_Θ is improper, the computation of (5) is still possible provided that the normalization constant (6) is finite (Robert, 2007, Section 1.5). However, the samples obtained from the MH sampler can no longer be interpreted as samples from the posterior distribution (5), since p_Θ is not a pdf anymore.

As noted above, the correctness of the MH sampler is guaranteed as long as the normalization constant (6) is finite. However, the boundedness of (6) is difficult to verify in complex settings (Robert, 2007, p. 30), which is the case for this work as we impose priors over the Hankel singular values of G_θ .

To guarantee the correctness of the MH sampler, we impose the following

Assumption 2: *The set Θ can be written as*

$$\Theta = \Theta_G \times \Theta_H \times \Theta_e, \quad (21)$$

where $\theta = (\theta_G^\top, \theta_H^\top, \theta_e^\top)^\top \in \Theta$ is such that the model (4) is given by $G_\theta = G_{\theta_G}$, $H_\theta = H_{\theta_H}$ and $e_t^\theta = e_t^{\theta_e}$ for all t .

Assumption 3: *For the set Θ satisfying Assumption 2, the set Θ_G is bounded.*

Assumption 2 requires that the rational transfer functions G_θ , H_θ and the white noise process $\{e_t^\theta\}$ in (4) are independently parameterized. In addition, the boundedness of the set Θ_G required by Assumption 3 does not introduce considerable conservatism to the model structure (4). Indeed, for reasonable parameterizations of (G_θ, H_θ) , from Assumption 1 it follows that the parameters describing the denominator of G_θ and H_θ lie in a bounded set. Hence, Assumption 3 only implies that, for reasonable parameterizations of G_θ , the parameters describing the numerator of G_θ also lie in a bounded set. This is a natural assumption as the model structure (4) describes systems satisfying the bounded input-bounded output property.

To guarantee the boundedness of the normalization constant (6), we also impose the following requirement over p_Θ :

Assumption 4: For the set Θ satisfying Assumption 2, the prior p_Θ is given for all $\theta \in \Theta$ by

$$p_\Theta(\theta) = p_{\Theta_G}(\theta_G) p_{\Theta_H}(\theta_H) p_{\Theta_e}(\theta_e), \quad (22)$$

where p_{Θ_G} , p_{Θ_H} and p_{Θ_e} are proper priors defined over Θ_G , Θ_H and Θ_e , respectively. In addition, p_{Θ_G} is bounded on Θ_G .

We note that Assumption 4 considers priors over Θ that are possibly not normalized.

In our setting, the prior p_{Θ_G} is specified over the Hankel singular values associated with G_{θ_G} . If we denote by p_Σ a proper and bounded prior imposed over the Hankel singular values $(\sigma_1, \dots, \sigma_{n_{\mathcal{M}}})$ of models with $\theta_G \in \Theta_G$, then a prior p_{Θ_G} is given by

$$p_{\Theta_G}(\theta_G) = \frac{p_\Sigma(\sigma_1(\theta_G), \dots, \sigma_{n_{\mathcal{M}}}(\theta_G))}{K_{\Theta_G}}, \quad (23)$$

where $\sigma_i(\theta_G)$ denotes the i -th Hankel singular value associated with G_{θ_G} ($i \in \{1, \dots, n_{\mathcal{M}}\}$), and

$$K_{\Theta_G} := \int_{\Theta_G} p_\Sigma(\sigma_1(\theta_G), \dots, \sigma_{n_{\mathcal{M}}}(\theta_G)) d\theta_G, \quad (24)$$

is the normalization constant.

Remark 2: A closer inspection to p_{Θ_G} in (23) tells us that the prior imposed in this article results in a uniform distribution over all $\theta_G \in \Theta_G$ having the same Hankel singular values $\{\sigma_i(\theta_G)\}_{i=1}^{n_{\mathcal{M}}}$.

We note that (24) is finite as the set Θ_G is bounded. Indeed,

$$K_{\Theta_G} \leq \sup_{\theta_G \in \Theta_G} p_\Sigma(\sigma_1(\theta_G), \dots, \sigma_{n_{\mathcal{M}}}(\theta_G)) \int_{\Theta_G} d\theta < \infty. \quad (25)$$

Hence, the prior pdf (23) fulfills the requirements in Assumption 4.

Finally, if the set Θ satisfies Assumption 2 and p_Θ satisfies Assumption 4 with p_{Θ_G} given by (23), then

$$\int_{\Theta} p_\Theta(\theta) d\theta \leq \int_{\Theta_G} p_{\Theta_G}(\theta_G) d\theta_G \int_{\Theta_e} \int_{\Theta_H} p_{\Theta_H}(\theta_H) p_{\Theta_e}(\theta_e) d\theta_H d\theta_e < \infty. \quad (26)$$

Based on (26), we conclude that the normalization constant (6) is bounded for almost every value of $\mathbf{Z}_N$ (Robert, 2007, Problem 1.47.b). Therefore, the MH sampler draws samples that are approximately distributed according to (5), when p_{Θ_G} is a proper prior defined over the Hankel singular values of G_{θ_G} (according to (23)) and Assumptions 2-4 are satisfied.

The correctness of the MH sampler for the proposed setup is summarized by

Theorem 3: Consider the set Θ satisfying Assumptions 2-3 and a prior p_Θ satisfying Assumption 4, with p_{Θ_G} given by (23). Let $\{\theta^{(i)}\}_{i \geq 1}$ be the samples generated by the Markov chain in Algorithm 1, with $\pi(\theta) = p(y_{1:N}|u_{1:N}, \theta) p_\Theta(\theta)$, and the conditional pdf $f(\theta|\theta^*)$ whose support includes Θ . If

$$P \left\{ p(y_{1:N}|u_{1:N}, \theta^{(i)}) p_\Theta(\theta^{(i)}) f(\theta^*|\theta^{(i)}) \leq p(y_{1:N}|u_{1:N}, \theta^*) p_\Theta(\theta^*) f(\theta^{(i)}|\theta^*) \right\} < 1 \quad (27)$$

is satisfied for all $\theta^{(i)}$, with P being the probability measure associated with the Markov process

$\{\theta^{(i)}\}_{i \geq 1}$, then

$$\lim_{i \rightarrow \infty} \left\| \int_{\Theta} p^{(i)}(\cdot | \theta^{(1)}) \mu(\theta^{(1)}) d\theta^{(1)} - p(\cdot | \mathbf{Z}_N) \right\|_{\text{TV}} = 0 \quad (28)$$

for almost every value of $\mathbf{Z}_N$ and any initial pdf μ , where $p^{(i)}$ denotes the conditional pdf of $\theta^{(i)}$ given $\theta^{(1)}$, and

$$\|\beta\|_{\text{TV}} := \frac{1}{2} \int_{\Theta} |\beta(\theta)| d\theta, \quad (29)$$

with $\beta: \Theta \rightarrow \mathbb{R}$.

Proof. Follows from the existence of the posterior pdf $p(\cdot | \mathbf{Z}_N)$ according to Robert (2007, Problem 1.47.b) and the convergence of the MH sampler provided the existence of $p(\cdot | \mathbf{Z}_N)$ from Robert & Casella (2004, Corollary 7.5). $\square$

Remark 4: The expression (27) establishes a sufficient condition for guaranteeing that the Markov chain generated by the MH sampler is aperiodic. The statement in Theorem 3 can also be proven under less restrictive conditions than (27), we refer to Robert & Casella (2004, Section 7.3.2) for more details.

5. A random walk over the parameter set

To run the MH sampler over Θ , we need to define a random walk over Θ based on the proposal distribution f . One of the main difficulties in implementing such a random walk is to ensure the stability of the pair (G_θ, H_θ) and H_θ^{-1} for the value of θ being sampled. To address this point, we note that the real rational functions G_θ, H_θ can be written as

$$G_\theta(q) = K \prod_{i=1}^{n_{\mathcal{M}}/2} \frac{1 + b_1^i q^{-1} + b_2^i q^{-2}}{1 + \ell_1^i q^{-1} + \ell_2^i q^{-2}}, \quad (30a)$$

$$H_\theta(q) = \prod_{i=1}^{n_{\mathcal{M}}/2} \frac{1 + c_1^i q^{-1} + c_2^i q^{-2}}{1 + d_1^i q^{-1} + d_2^i q^{-2}}, \quad (30b)$$

if $n_{\mathcal{M}}$ is even, and

$$G_\theta(q) = K \frac{1 + b_0 q^{-1}}{1 + \ell_0 q^{-1}} \prod_{i=1}^{n_{\mathcal{M}}/2-1} \frac{1 + b_1^i q^{-1} + b_2^i q^{-2}}{1 + \ell_1^i q^{-1} + \ell_2^i q^{-2}}, \quad (31a)$$

$$H_\theta(q) = \frac{1 + c_0 q^{-1}}{1 + d_0 q^{-1}} \prod_{i=1}^{n_{\mathcal{M}}/2-1} \frac{1 + c_1^i q^{-1} + c_2^i q^{-2}}{1 + d_1^i q^{-1} + d_2^i q^{-2}}, \quad (31b)$$

if $n_{\mathcal{M}}$ is odd. In Equations (30)–(31) the parameter θ is a real valued function of the gain K and the polynomial coefficients. The requirements in Assumption 1 are satisfied by guaranteeing that all the poles of G_θ and all the poles and zeros of H_θ belong to $\mathbb{D}$. Assumption 1 is satisfied if and

only if (Åström & Wittenmark, 1997)

$$|\ell_0| < 1, \quad |c_0| < 1, \quad |d_0| < 1, \quad (32a)$$

$$|\ell_2^i| < 1, \quad |c_2^i| < 1, \quad |d_2^i| < 1, \text{ for all } i, \quad (32b)$$

$$|\ell_1^i| < 1 + \ell_2^i, \quad |c_1^i| < 1 + c_2^i, \quad |d_1^i| < 1 + d_2^i, \text{ for all } i. \quad (32c)$$

Thus, given the proposal distribution f , at step i of the MH sampler we generate θ^* from $f(\theta|\theta^{(i)})$ and check if the conditions (32) are satisfied. If they are not satisfied, we then sample a new parameter θ^* until the inequalities (32) are fulfilled. In this manner we guarantee that the samples θ^* generated from f belong to Θ . Note that this method can be seen as a rejection sampling procedure.

Remark 5: The approach presented here for implementing a random walk over the parameter set does not guarantee that the resulting Markov chain will possess a fast convergence rate to its stationary distribution. A fast convergence rate of the Markov chain implies that a small amount of samples have to be discarded at the beginning (to avoid the influence of the initial conditions in the samples), and that the stationary distribution of the chain can be approximated by a small amount of samples. Hence, the convergence rate of the resulting Markov chain affects the computational cost of approximating the stationary distribution by its samples.

One way to improve the convergence rate is to scale the covariance matrix of the random walk proposal by the covariance matrix of the posterior distribution (Dahlin & Schön, 2015). To this end, we first run the MH sampler to estimate by Monte Carlo the covariance matrix of the posterior distribution based on the generated samples, which is then used to build the random walk for the second run of the MH sampler. In the second run of the MH sampler, the covariance matrix of the random walk is the one employed in the first run, scaled by the estimated covariance matrix of the posterior distribution. In this manner, the samples from the second run are obtained from a Markov chain with better convergence rate than for the first run.

Another option to improve the convergence rate is to reparameterize the model (4) so that the resulting MH sampler is performed over an unconstrained space. To this end, the acceptance probability (16) needs to be modified to include the gradient of the transformation between parameters in Θ and parameters in the unconstrained space. Finally, it is also possible to consider an adapting scheme for improving the convergence properties of the MH sampler, where the covariance matrix of the proposal distribution is tuned to achieve a good acceptance ratio of the samples drawn from this distribution. For further details we refer to Dahlin & Schön (2015), where a more elaborate discussion on extensions for the MH sampler is presented.

Remark 6: As the MH sampler approximates the posterior distribution of the model parameters given $\mathbf{Z}_N$, the proposed method requires us to draw several samples from the designed Markov chain. Hence, the computational time associated with the MH sampler can be significantly larger than the one associated with other identification methods, such as PEM and subspace techniques. Nevertheless, both PEM and subspace methods generate a single estimate given $\mathbf{Z}_N$, and not a set of samples from a target distribution as done by the MH sampler.

6. Numerical illustrations

6.1 System and algorithm setup

Consider the system

$$y_t = G_o(q) u_t + e_t, \quad (33)$$

where

$$G_o(q) = \frac{2q^{-2}}{1 - 1.4q^{-1} + 0.81q^{-2}}, \quad (34)$$

and $\{u_t\}, \{e_t\}$ are Gaussian white noise processes, with zero mean and unit variance. For simplicity, we assume that the system (33) is at rest for all $t \leq 0$. The system (33) is estimated based on $\mathbf{Z}_N$, with $N = 100$, and on the model

$$y_t = G_{\theta_G}(q) u_t + H_{\theta_H}(q) e_t^{\theta_e}, \quad (35)$$

where $\{e_t^{\theta_e}\}$ is Gaussian white noise, with zero mean and variance $\theta_e > 0$, which has to be estimated. In this example, the pair $(G_{\theta_G}, H_{\theta_H})$ can describe all real rational functions in q satisfying Assumption 1 and that are of order at most $n_{\mathcal{M}} = 10$. In particular, the parameter vectors $\theta_G \in \Theta_G \subset \mathbb{R}^{2n_{\mathcal{M}}+1}$, $\theta_H \in \Theta_H \subset \mathbb{R}^{2n_{\mathcal{M}}}$ contain the gain and the numerator and denominator polynomial coefficients of the rational functions G_{θ_G} and H_{θ_H} , i.e.,

$$G_{\theta_G}(q) = \theta_{G,0} \frac{1 + \sum_{i=1}^{n_{\mathcal{M}}} \theta_{G,i} q^{-i}}{1 + \sum_{i=1}^{n_{\mathcal{M}}} \theta_{G,i+n_{\mathcal{M}}} q^{-i}}, \quad (36a)$$

$$H_{\theta_H}(q) = \frac{1 + \sum_{i=1}^{n_{\mathcal{M}}} \theta_{H,i} q^{-i}}{1 + \sum_{i=1}^{n_{\mathcal{M}}} \theta_{H,i+n_{\mathcal{M}}} q^{-i}}, \quad (36b)$$

with $\theta_G = (\theta_{G,0}, \dots, \theta_{G,2n_{\mathcal{M}}})^\top$, $\theta_H = (\theta_{H,1}, \dots, \theta_{H,2n_{\mathcal{M}}})^\top$. Hence, the parameter vector is given by $\theta = (\theta_G^\top, \theta_H^\top, \theta_e)^\top \in \Theta \subset \mathbb{R}^{2(2n_{\mathcal{M}}+1)}$. In addition, we require that each entry of the parameter vector $(\theta_{G,0}, \dots, \theta_{G,n_{\mathcal{M}}})^\top$ satisfies $|\theta_{G,i}| \leq 10^3$ (hence Assumption 3 is satisfied).

To implement the MH sampler for this example, we define a proposal distribution over Θ following the discussion in Section 5, where the model parameters in (36) are obtained as functions of the parameters in (30). Based on the rational functions (30), we define f as the product of the proposals for each parameter in (30) and the proposal for θ_e , where

$$\theta_e^* \sim \mathcal{N}(\theta_e, 10^{-3}), \quad (37a)$$

$$K^* \sim \mathcal{N}(K, 10^{-4}), \quad (37b)$$

$$\ell_j^{i*} \sim \mathcal{N}(\ell_j^i, 10^{-4}), i \in \{1, \dots, n_{\mathcal{M}}/2\}, j \in \{1, 2\}, \quad (37c)$$

$$c_j^{i*} \sim \mathcal{N}(c_j^i, 10^{-4}), i \in \{1, \dots, n_{\mathcal{M}}/2\}, j \in \{1, 2\}, \quad (37d)$$

$$d_j^{i*} \sim \mathcal{N}(d_j^i, 10^{-4}), i \in \{1, \dots, n_{\mathcal{M}}/2\}, j \in \{1, 2\}. \quad (37e)$$

From the proposals (37) we note that $f(\theta|\theta^*) = f(\theta^*|\theta)$. Hence, the acceptance probability (16) for this example becomes

$$a(\theta^*, \theta) = \min \left(1, \frac{p(y_1|u_1, \theta^*) \prod_{t=2}^N p_e^{\theta^*}(\varepsilon_t(\theta^*)) p_\Theta(\theta^*)}{p(y_1|u_1, \theta) \prod_{t=2}^N p_e^\theta(\varepsilon_t(\theta)) p_\Theta(\theta)} \right). \quad (38)$$

For the prior distribution over Θ , we consider

$$p_\Theta(\theta) = p_{\Theta_G}(\theta_G) p_{\Theta_H}(\theta_H) p_{\Theta_e}(\theta_e), \quad (39)$$

Figure 1. Bode diagram of the true system G_o (black), the maximum a posteriori estimate $\hat{G}_{MAP}$ (red, continuous line), and the reduced model $\hat{G}_{MAP}^r$ (red, dashed line), obtained by balanced truncation of $\hat{G}_{MAP}$. The shaded area corresponds to the 98% confidence interval for the posterior distribution, centered around its mean value.

where

$$p_{\Theta_G}(\theta_G) = \prod_{i=1}^{n_M} p_{\sigma_i}(\sigma_i(\theta_G)), \quad (40)$$

with p_{σ_i} denoting the prior pdf associated with the i -th Hankel singular value of G_{θ_G} , for $i \in \{1, \dots, n_M\}$. In addition, $\{p_{\sigma_i}\}_{i=1}^{n_M}$, and p_{Θ_e} are defined as the pdf of a Gamma distribution with scale parameter 1 and shape parameter 1, and p_{Θ_H} is defined as the pdf of a uniform distribution over Θ_H .

To approximate the posterior (5) we draw samples with the MH sampler introduced in Algorithm 1, where we initialize the parameters at zero for $k = 1$. We generate 10^5 samples and discard the first $2 \cdot 10^4$ ones. For this setup, the acceptance ratio of the samples drawn in line 3 of Algorithm 1 is 27%, and the execution of one iteration of lines 2 to 7 in Algorithm 1 takes approximately 0.35 seconds on a stationary computer with an Intel Core i7 CPU of clock frequency 3.4 [GHz] and 16 [GB] of RAM memory equipped with Matlab 2015b, leading to a computational time of approximately 10 hours for completing the execution of Algorithm 1.

6.2 Model estimates

The samples obtained by running Algorithm 1 are employed to compute the parameters estimates (7) and (8), where the posterior distribution (5) is replaced by its approximation based on $\{\theta^{(i)}\}_{i=2 \cdot 10^4}^{10^5}$.

Figure 1 presents¹ the Bode diagram of $\hat{G}_{MAP} := G_{\hat{\theta}_N}$, where $\hat{\theta}_N$ is the maximum a posteriori estimate (7). In addition, Figure 1 also presents the balanced truncation of the estimate $\hat{G}_{MAP}$, denoted by $\hat{G}_{MAP}^r$, which is obtained by removing the states in the balanced state-space representation of $\hat{G}_{MAP}$ associated with the Hankel singular values with a value that is smaller than 10% of the largest Hankel singular value. The same results are presented in Figure 2 for the posterior mean estimate (8), where $\hat{G}_m$ and $\hat{G}_m^r$ denote the mean estimate and its balanced truncation,

¹The phase of the estimates in Figures 1-2 is shifted by 2π to ease the comparison.

Figure 2. Bode diagram of the true system G_o (black), the mean value estimate $\hat{G}_m$ (red, continuous line), and the reduced model $\hat{G}_m^r$ (red, dashed line), obtained by balanced truncation of $\hat{G}_m$. The shaded area corresponds to the 98% confidence interval.

Figure 3. Magnitude of the Bode diagram for the model error. Top: Maximum a posteriori estimate $\hat{G}_{MAP}$ and its reduced model $\hat{G}_{MAP}^r$. Bottom: Posterior mean estimate $\hat{G}_m$ and its reduced model $\hat{G}_m^r$.

respectively. From these figures we conclude that the estimates capture the dynamics of the true system (33) rather well. Moreover, we also note that the degradation in the estimates is larger in the frequency range where the magnitude of the response of G_o is at most 0 [dB], which is the value of the noise variance. This can also be seen in the confidence intervals in Figures 1 and 2, where the uncertainty is large when the magnitude of the response of G_o is less than or equal to the noise variance.

To complement the discussion, Figure 3 shows the magnitude of the frequency response of the model error for the maximum a posteriori estimate $\hat{G}_{MAP}$ and the posterior mean $\hat{G}_m$, together with their balanced truncations $\hat{G}_{MAP}^r$ and $\hat{G}_m^r$. From this figure we observe that the magnitude of the model error is frequency dependent, and that its peak occurs close to the resonance peak of

the magnitude of G_o .

The models obtained by balanced truncations of the estimated functions $\hat{G}_{\text{MAP}}^r$, and $\hat{G}_m^r$ are

$$\hat{G}_{\text{MAP}}^r(q) = \frac{-0.29(1 - 3.26q^{-1})(1 + 1.99q^{-1})}{1 - 1.39q^{-1} + 0.80q^{-2}}, \quad (41)$$

$$\hat{G}_m^r(q) = \frac{-0.35(1 - 2.96q^{-1})(1 + 1.73q^{-1})}{1 - 1.39q^{-1} + 0.80q^{-2}}. \quad (42)$$

From these expressions we see that the order of the reduced models coincides with the order of the system (33). In addition, the relative degree of G_o is estimated by the finite nonminimum phase zeros in (41) and (42). Indeed, the relative degree of a discrete time linear system is associated with time delays, which can be interpreted as nonminimum phase zeros at infinity. Hence, the models (41) and (42) try to capture the time delays (and thus the relative degree of G_o) by finite nonminimum phase zeros.

We can also compute the model fit for the estimated functions (41) and (42), by defining

$$\mathcal{F}(G_1, G_2) := 100 \cdot \left(1 - \frac{\|G_1 - G_2\|_2}{\|G_2\|_2}\right) [\%], \quad (43)$$

where G_1 , and G_2 are rational functions in q , and $\|\cdot\|_2$ is computed using (1). Hence, for the reduced models we obtain $\mathcal{F}(\hat{G}_{\text{MAP}}^r, G_o) = 93.86\%$ for the MAP estimate, and $\mathcal{F}(\hat{G}_m^r, G_o) = 91.98\%$ for the mean estimate. Therefore, the estimated models describe the main characteristics in the frequency response of G_o .

To analyze the effect of imposing priors over the Hankel singular values, we run an MH sampler targeting the posterior distribution of θ when p_{Θ_G} is defined as a normal distribution with zero mean and identity covariance matrix for the parameter vector $\{\theta_{G,i}\}_{i=0}^{2n_M}$. However, due to convergence issues, this time the MH sampler runs over $5 \cdot 10^5$ iterations, and we discard the first 10^5 samples. In this case, the acceptance ratio of the samples drawn in line 3 of Algorithm 1 is 18%. As before, we compute the balanced truncations of the new estimated functions $\hat{G}_{\text{MAP}}^r$, and $\hat{G}_m^r$, which are given by

$$\hat{G}_{\text{MAP}}^{r,N}(q) = \frac{3.2 \cdot 10^{-2}(1 - 15.73q^{-1} + 74.8q^{-2})}{1 - 1.39q^{-1} + 0.80q^{-2}}, \quad (44)$$

$$\hat{G}_m^{r,N}(q) = \frac{4.5 \cdot 10^{-2}(1 - 14.44q^{-1} + 54.79q^{-2})}{1 - 1.39q^{-1} + 0.80q^{-2}}. \quad (45)$$

The model fit for the new estimated models are $\mathcal{F}(\hat{G}_{\text{MAP}}^{r,N}, G_o) = 87.27\%$, and $\mathcal{F}(\hat{G}_m^{r,N}, G_o) = 83.49\%$. Hence, adding priors over the Hankel singular values in this example helps to improve the model estimates by capturing the main properties of the underlying system.

Before concluding this section, we present in Figure 4 the Bode diagram of the maximum a posteriori estimate ($\hat{H}_{\text{MAP}}$) and the posterior mean estimate ($\hat{H}_m$) of the noise filter, where the true noise filter is $H_o = 1$. Since we do not penalize the model complexity for these estimates, we expect that the models are not necessarily of low complexity. This can be seen in Figure 4, where the frequency response cannot be directly associated to a constant filter. Nevertheless, we emphasize that the present contribution tries to improve the model estimates for G_o by imposing priors on the model complexity of G_{θ_G} , while the estimates of H_o do not account for such complexity.

Figure 4. Bode diagram of the noise model derived by the maximum a posteriori estimate $\hat{H}_{MAP}$ and the posterior mean estimate $\hat{H}_m$. The true noise filter is $H_o = 1$.

6.3 Comparison with standard methods

We can further compare the results obtained from the MH sampler with those computed by standard identification methods. To this end, the same data set $\mathbf{Z}_N$ used by the MH sampler is employed to estimate a model based on PEM and subspace identification. Both methods are implemented in Matlab using the commands `pem` for PEM and `n4sid` for subspace identification. For PEM, a Box-Jenkins model structure is chosen, where both G_θ and H_θ are rational functions in q of order 10. For subspace identification, a linear state space model in innovations form of order 10 is chosen as model structure. The estimated models from PEM and subspace identification are reduced by the same balanced truncation procedure considered for the models obtained from the MH sampler.

The Bode diagrams of the estimated models for G_o using PEM and subspace identification are presented in Figure 5 and Figure 6, respectively². The models estimated by both identification methods are close in magnitude to the frequency response of G_o . In addition, they present a degradation in the model fit for the frequency region where the magnitude of G_o is close to the noise variance, which is also observed in the estimates derived from the MH sampler. We can complement this discussion with Figure 7, where the magnitude of the Bode diagram for the model error is shown for the PEM and subspace estimates.

Once the balanced truncation is performed over the models obtained by PEM ($\hat{G}_{PEM}$) and subspace identification ($\hat{G}_{SS}$), the reduced order models are³

$$\hat{G}_{PEM}^r(q) = G_{PEM}^1(q) G_{PEM}^2(q) G_{PEM}^3(q), \quad (46)$$

$$\hat{G}_{SS}^r(q) = \frac{-0.08 q^{-1}(1 - 25 q^{-1})}{1 - 1.41 q^{-1} + 0.80 q^{-2}}, \quad (47)$$

where $\hat{G}_{PEM}^r$ and $\hat{G}_{SS}^r$ denote the reduced models for PEM and subspace identification respectively,

²To ease the presentation, the phase of the estimates in Figures 5 and 6 is shifted by 6π and 2π respectively.

³The PEM estimate is decomposed into the product of three rational functions to ease the presentation.

Figure 5. Bode diagram of the true system G_o (yellow, triangle line), the PEM estimate $\hat{G}_{PEM}$ (red, continuous line), and the reduced PEM model $\hat{G}_{PEM}^r$ (blue, dashed line), obtained by balanced truncation of $\hat{G}_{PEM}$.

and

$$G_{PEM}^1(q) = \frac{-0.03(1 - 12.2q^{-1})(1 + 4.35q^{-1})(1 + 1.89q^{-1} + 0.93q^{-2})}{(1 + 1.95q^{-1} + 0.99q^{-2})(1 - 1.39q^{-1} + 0.80q^{-2})}, \quad (48)$$

$$G_{PEM}^2(q) = \frac{1 + 1.26q^{-1} + 1.23q^{-2}}{1 + 1.09q^{-1} + 0.94q^{-2}}, \quad (49)$$

$$G_{PEM}^3(q) = \frac{1 - 0.62q^{-1} + 1.05q^{-2}}{1 - 0.60q^{-1} + 0.93q^{-2}}. \quad (50)$$

From the expression for $\hat{G}_{PEM}^r$ we observe that the reduced model for the PEM estimate is of order 8, and hence it overestimates the order of G_o . On the other hand, the reduced order model $\hat{G}_{SS}^r$ results in a second order rational function. Thus, the reduced order model derived from the subspace estimate coincides with the order of the rational function G_o . The model fit for the reduced order models are $\mathcal{F}(\hat{G}_{PEM}^r, G_o) = 81.3\%$ for PEM and $\mathcal{F}(\hat{G}_{SS}^r, G_o) = 90.4\%$ for subspace identification. Thus, even though the subspace estimate $\hat{G}_{SS}^r$ captures the true order of the system, its performance is still slightly below the one achieved by estimates generated from the MH sampler when priors over the Hankel singular values of G_θ are imposed.

6.4 Hankel singular values

To illustrate the effect of imposing priors over the Hankel singular values of G_θ , we present in Figure 8 the Hankel singular values (in logarithmic scale) for the estimated models $\hat{G}_{MAP}$ and $\hat{G}_m$ obtained from the MH sampler with prior distribution (39)–(40) and the estimated models $\hat{G}_{PEM}$ and $\hat{G}_{SS}$. Note that $\sigma_i(G_o) = 0$ for all $i > 2$, which are not depicted in Figure 8. From

Figure 6. Bode diagram of the true system G_o (yellow, triangle line), the subspace estimate $\hat{G}_{SS}$ (red, continuous line), and the reduced subspace model $\hat{G}_{SS}^r$ (blue, dashed line), obtained by balanced truncation of $\hat{G}_{SS}$.

this plot we see that all the estimated models behave similarly for the first and second Hankel singular values. In the case of $\hat{G}_{PEM}$, we see that the remaining Hankel singular values cannot be completely discarded, as they indicate that the associated states have a noticeable effect in the resulting model. The situation improves for $\hat{G}_{SS}$, where we see that the states associated with the first and second Hankel singular values can almost completely explain the resulting model. The behavior observed for the Hankel singular values of $\hat{G}_{SS}$ is sharpened for $\hat{G}_{MAP}$ and $\hat{G}_m$, where the evidence for a reduced model order is clearer than for $\hat{G}_{SS}$. Interestingly, the same behavior is observed for the Hankel singular values of the posterior estimates with Gaussian prior ($\hat{G}_{MAP}^N$ and $\hat{G}_m^N$), where the Hankel singular values decay faster than those associated with the Bayesian models with priors over the Hankel singular values. However, this does not necessarily mean that the estimates $\hat{G}_{MAP}^N$ and $\hat{G}_m^N$ attain a better model fit than $\hat{G}_{MAP}$ and $\hat{G}_m$. In fact, we have previously seen that a Gaussian prior over the model parameters can degrade the model fit when compared to the Bayesian estimates based on model order priors. Hence, imposing priors over the Hankel singular values of G_θ helps to account for the model complexity of G_θ , while attaining a good model fit.

As a final comparison, we consider the fit to the estimation data using the index

$$\mathcal{D}(\theta) := 100 \cdot \left(1 - \frac{\sqrt{\sum_{t=1}^N \varepsilon_t(\theta)^2}}{\sqrt{\sum_{t=1}^N y_t^2}} \right) [\%], \quad (51)$$

where $\varepsilon_t(\theta)$ corresponds to the one step ahead prediction error given θ , defined in (15). Table 1 presents the fit index to the data for the PEM estimate and its balanced truncation (θ_{PEM} and θ_{PEM}^r , respectively), the subspace estimate and its balanced truncation (θ_{SS} and θ_{SS}^r , respectively),

Figure 7. Magnitude of the Bode diagram for the model error. Top: PEM estimate $\hat{G}_{PEM}$ and its reduced model $\hat{G}_{PEM}^r$. Bottom: Subspace estimate $\hat{G}_{SS}$ and its reduced model $\hat{G}_{SS}^r$.

Figure 8. Hankel singular values for the estimated models and the true system G_o (in logarithmic scale).

Estimate	θ_{PEM}	θ_{PEM}^r	θ_{SS}	θ_{SS}^r	θ_{MAP}	θ_{MAP}^r	θ_m	θ_m^r
$\mathcal{D}(\cdot)$	78.2%	78.5%	78.7%	78.7%	77.3%	78.4%	77.7%	78.1%

the maximum a posteriori estimate and its balanced truncation (θ_{MAP} and θ_{MAP}^r respectively), and the posterior mean and its balanced truncation (θ_m and θ_m^r respectively). Hence, the indices indicate that imposing an appropriate prior over the Hankel singular values drives the estimates towards models with low complexity, without affecting negatively the fit to the output data.

6.5 Posterior distributions

The samples obtained from the MH sampler can be used to compute posterior distributions of nontrivial functions of the parameter vector. For example, we can use the samples $\{\theta^{(i)}\}_{i=2 \cdot 10^4}^{10^5}$ to

Figure 9. Empirical pdf of the phase margin.

approximate the posterior distribution of the phase margin associated with G_{θ_G} . Figure 9 presents the empirical pdf obtained for the phase margin using the samples of the posterior pdf $p(\theta|\mathbf{Z}_N)$, when the prior p_{Θ_G} is defined over the Hankel singular values of the transfer function G_{θ_G} . We note in Figure 9 that the phase margin is nonpositive with very high probability, which implies that the resulting closed loop with negative feedback and unit controller will be unstable with very high probability.

6.6 Choice of the proposal distribution

To analyze the effect of the proposal distribution in the performance of Algorithm 1, we consider the same setup described in Section 6.1, but three cases for the random walk over the parameter set:

- Case 1:

$$\theta_e^* \sim \mathcal{N}(\theta_e, 10^{-2}), \quad (52a)$$

$$K^* \sim \mathcal{N}(K, 10^{-3}), \quad (52b)$$

$$\ell_j^{i*} \sim \mathcal{N}(\ell_j^i, 10^{-3}), i \in \{1, \dots, n_{\mathcal{M}}/2\}, j \in \{1, 2\}, \quad (52c)$$

$$c_j^{i*} \sim \mathcal{N}(c_j^i, 10^{-3}), i \in \{1, \dots, n_{\mathcal{M}}/2\}, j \in \{1, 2\}, \quad (52d)$$

$$d_j^{i*} \sim \mathcal{N}(d_j^i, 10^{-3}), i \in \{1, \dots, n_{\mathcal{M}}/2\}, j \in \{1, 2\}. \quad (52e)$$

- Case 2: The random walk in Equation (37).
- Case 3:

$$\theta_e^* \sim \mathcal{N}(\theta_e, 10^{-4}), \quad (53a)$$

$$K^* \sim \mathcal{N}(K, 10^{-5}), \quad (53b)$$

$$\ell_j^{i*} \sim \mathcal{N}(\ell_j^i, 10^{-5}), i \in \{1, \dots, n_{\mathcal{M}}/2\}, j \in \{1, 2\}, \quad (53c)$$

$$c_j^{i*} \sim \mathcal{N}(c_j^i, 10^{-5}), i \in \{1, \dots, n_{\mathcal{M}}/2\}, j \in \{1, 2\}, \quad (53d)$$

$$d_j^{i*} \sim \mathcal{N}(d_j^i, 10^{-5}), i \in \{1, \dots, n_{\mathcal{M}}/2\}, j \in \{1, 2\}. \quad (53e)$$

Figure 10. Evolution of the logarithm of the unnormalized posterior distribution $p(y_{1:N}|u_{1:N}, \theta^{(i)}) p_{\Theta}(\theta^{(i)})$ as function of the iteration i .

We still consider the normal distribution as the random walk for the parameters since its support covers the entire parameter set Θ , which ensures the correctness of Algorithm 1 (cf. Robert & Casella (2004)). For Case 3, the MH sampler cannot progress in the random walk, as the unnormalized posterior pdf $p(y_{1:N}|u_{1:N}, \theta) p_{\Theta}(\theta)$ is stuck at the initial vector. Hence, the variance of the random walk in Case 3 is not sufficiently large to guarantee that the random walk does not get locked in the initial point of the Markov chain or in local optima.

For Cases 1 and 2, we run the MH sampler and analyze the behavior of the product $p(y_{1:N}|u_{1:N}, \theta^{(i)}) p_{\Theta}(\theta^{(i)})$ in terms of the iteration i . Figure 10 shows the evolution of the logarithm of the unnormalized posterior distribution $p(y_{1:N}|u_{1:N}, \theta^{(i)}) p_{\Theta}(\theta^{(i)})$ as a function of the iteration i . In this figure we can note that the samples generated by the random walk in Case 1 are often locked in a specific parameter, as the value of $\log(p(y_{1:N}|u_{1:N}, \theta^{(i)}) p_{\Theta}(\theta^{(i)}))$ remains constant for several iterations. This implies that the resulting Markov chain has poor mixing properties, which means that more samples are needed for approximating the target distribution.

Concerning Case 2, we see from Figure 10 that the mixing properties are better than the ones associated with the Markov chain for Case 1. This can be seen by noticing that the random walk reaches higher values of $\log(p(y_{1:N}|u_{1:N}, \theta^{(i)}) p_{\Theta}(\theta^{(i)}))$ faster.

In conclusion, we see that the choice of the variance associated with the random walk is crucial for obtaining appropriate mixing conditions in the resulting Markov chain. On the one hand, a small variance does not allow to explore the parameter set, as the MH sampler can get stuck in local optima. On the other hand, a large variance affects negatively the mixing properties of the Markov chain. Therefore, the variance of the random walk needs to be specified by establishing a trade-off between the mixing properties of the associated Markov chain and the possibility of sticking to a local optimum.

6.7 Choice of the prior distribution

We continue our analysis of the method by inspecting the effect of the choice of the prior distribution on the Hankel singular values of the estimated models. To this end, we consider the same setup described in Section 6.1, and we modify the prior pdfs $\{p_{\sigma_i}\}_{i=1}^{n_M}$ defined for the Hankel singular values. In this case, we consider each p_{σ_i} to be the pdf of a Gamma distribution with shape

Figure 11. Hankel singular values for the estimated models for different priors (in logarithmic scale).

parameter 2 and scale parameter 1. Note that the pdf associated with this distribution attains its maximum for some strictly positive value. Hence, we expect that the estimates generated in this case will favour models that are more complex than those analyzed in the previous parts of this example.

Figure 11 shows the Hankel singular values (in logarithmic scale) for the maximum a posteriori and the posterior mean estimates, obtained with a prior pdf associated with a Gamma distribution with shape parameter 1 and scale parameter 1 (labeled as Gamma(1,1) in the figure), and Gamma distribution with shape parameter 2 and scale parameter 1 (labeled as Gamma(2,1) in the figure). From the figure we observe that the choice of the prior pdf significantly affects the Hankel singular values. In particular, the Hankel singular values $\{\sigma_i\}_{i=3}^{10}$ associated with Gamma(2,1) are significantly greater than the ones obtained with the prior pdf Gamma(1,1), implying that the model complexity is greater for the former case than for the latter one. In terms of model fit for the prior Gamma(2,1), we have $\mathcal{F}(\hat{G}_{MAP}, G_o) = 78.4\%$ and $\mathcal{F}(\hat{G}_{MAP}^r, G_o) = 80.2\%$ for the maximum a posteriori estimate and its balanced truncation, and $\mathcal{F}(\hat{G}_m, G_o) = 81.1\%$ and $\mathcal{F}(\hat{G}_m^r, G_o) = 84.3\%$ for the posterior mean estimate and its balanced truncation. By comparing the model fits attained with the prior pdf Gamma(2,1) with respect to those achieved with Gamma(1,1), we see that the fit to the true system is reduced when the prior pdf Gamma(2,1) is used. Hence, it is important to impose a prior which is aligned with the expected effect on the model complexity of the posterior estimates. How to choose the prior distribution to attain a prescribed model complexity is a very important question to be addressed. However, this goes beyond the scope of this article and it will be addressed in a future work.

6.8 Monte Carlo study

As a final analysis, we perform a Monte Carlo study on the proposed estimation technique. To this end, we run the MH sampler over 102 data sets available in the standard library **s2d2** introduced by Chen, Ohlsson, & Ljung (2012). The library **s2d2** is attractive for our purposes, as it contains data sets with a restricted number of samples ($N = 375$), which are generated from slow systems of high order ($n_M = 30$) with low signal to noise ratio. Hence, the estimation of dynamical models based on the data set **s2d2** constitutes a challenging problem.

Figure 12. Comparison of the model fit for the PEM estimate $\hat{G}_{\text{PEM}}$ and the maximum a posteriori $\hat{G}_{\text{MAP}}$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.

We consider the same setup presented in Section 6.1 (including the prior distribution in Equations (39)–(40)), with the only differences that the model order is here set to $n_{\mathcal{M}} = 30$ (as it corresponds to the order of the models generating the data in `s2d2`), and the random walk for θ_e^* is now given by $\theta_e^* \sim \mathcal{N}(\theta_e, 10^{-4})$. For this setup, the sample mean of the acceptance ratio of the samples drawn in line 3 of Algorithm 1 is 10% with a standard deviation of 4%. In addition, the computational time required for one iteration of lines 2 to 7 in Algorithm 1 now increases to approximately 1 second on a stationary computer with an Intel Core i7 CPU of clock frequency 3.4 [GHz] and 16 [GB] of RAM memory equipped with Matlab 2015b, leading to approximately 28 hours to execute Algorithm 1 for one data set. Nevertheless, we mention that it is possible to parallelize the computations by setting simultaneous simulations with different data sets.

To compare the results obtained with the MH sampler, we estimate models using also PEM and subspace methods. For PEM, we estimate a Box-Jenkins model of order 30 for both G_θ and H_θ , and we initialize the search over the parameter set at zero. The initialization at zero is introduced to fairly compare the performance of the estimates obtained by the MH sampler with respect to PEM, and to assess the effect of the initial point in the numerical simulations. In the case of subspace identification, a state-space model of order 30 is estimated.

Figures 12 and 13 present the comparison of the model fits for the maximum a posteriori estimate $\hat{G}_{\text{MAP}}$ with respect to the PEM estimate $\hat{G}_{\text{PEM}}$, and with respect to the subspace estimate $\hat{G}_{\text{SS}}$. The same plots are presented for the posterior mean estimate $\hat{G}_m$ in Figures 14–15. In those figures, the blue circles indicate the cases where the models obtained with PEM and subspace methods delivered unstable systems (within the numerical tolerance of Matlab). In those cases, we set the model fit arbitrarily to -100% . From these figures, we see that the maximum a posteriori in general achieves a better model fit than the PEM and subspace models. This can be seen by noticing that most of the samples are above the line that indicates equal fit for the methods. The model fit degrades for the posterior mean estimate when compared to the maximum a posteriori. This is an expected result, since the posterior mean estimate combines models attaining a value in the posterior pdf less or equal than the one attained by the maximum a posteriori estimate, hence a degradation in the model estimate can be expected.

We can also analyze the model fit attained by the noise model for the different identification methods. This is shown in Figures 16–19, where the model fit for the Bayesian estimates $\hat{H}_{\text{MAP}}$ and $\hat{H}_m$ are compared to the model fit for the PEM and subspace models $\hat{H}_{\text{PEM}}$ and $\hat{H}_{\text{SS}}$. The

Figure 13. Comparison of the model fit for the subspace estimate $\hat{G}_{SS}$ and the maximum a posteriori $\hat{G}_{MAP}$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.

Figure 14. Comparison of the model fit for the PEM estimate $\hat{G}_{PEM}$ and the posterior mean $\hat{G}_m$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.

blue circles denote the case when $\hat{H}_{PEM}$ and $\hat{H}_{SS}$ deliver unstable models (within the numerical tolerance of Matlab). In those cases, we set the model fit to -100% . The behavior of the model fit for the noise model is comparable to the one attained for the estimates of G_o . However, we observe that the outliers of the model fit for the Bayesian estimates are significantly poor. This can be explained due to the fact that the noise filter H_o is commonly more difficult to estimate than G_o , and that there is no prior information encoded in the noise model to drive the posterior distribution to a prescribed region of the model set.

Figures 20–23 show the fit index $\mathcal{D}$ to the training data for the estimated models, where the fit index $\mathcal{D}$ is given by (51). The blue circles denote the case when either PEM or subspace deliver unstable models (within the numerical tolerance of Matlab). In those cases, we set the fit index to -100% . From these plots we observe that, in general, imposing a prior on the Hankel singular

Figure 15. Comparison of the model fit for the subspace estimate $\hat{G}_{SS}$ and the posterior mean $\hat{G}_m$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.

Figure 16. Comparison of the model fit for the PEM estimate $\hat{H}_{PEM}$ and the maximum a posteriori $\hat{H}_{MAP}$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.

values of G_o does not degrade significantly the fit index to the estimation data. Hence, the Bayesian estimates improve the model fit while preserving a reasonable fit to the data when compared to PEM and subspace estimates.

Finally, Figures 24–27 present the model order of the balanced truncation of the estimates after removing the states associated with Hankel singular values that are less than 10% of the largest one. The blue circles denote the case when either PEM or subspace deliver unstable models (within the numerical tolerance of Matlab). In those cases, we set the model order to 30. From these figures, we see that the Bayesian estimates tend to produce low order models when compared to PEM and subspace. The results are more noticeable in Figures 24 and 26, where the model order for the reduced PEM estimates are significantly greater (in general) than the Bayesian estimates.

In conclusion, the method proposed in this article provides an attractive alternative to account for

Figure 17. Comparison of the model fit for the subspace estimate $\hat{H}_{SS}$ and the maximum a posteriori $\hat{H}_{MAP}$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.

Figure 18. Comparison of the model fit for the PEM estimate $\hat{H}_{PEM}$ and the posterior mean $\hat{H}_m$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.

the model complexity in a Bayesian setting, while achieving good accuracy in the model estimates.

7. Conclusions

In this work a Metropolis Hastings sampler for Bayesian identification of SISO linear systems has been presented. By using the Metropolis Hastings sampler, we allow the choice of more general priors over the model parameters than those considered by the Gibbs sampler. This property has been exploited by specifying priors over the Hankel singular values of the model, which are closely connected with the order of the system.

Future work on the subject will be focused on further analysis of the approximation of the pos-

Figure 19. Comparison of the model fit for the subspace estimate $\hat{H}_{SS}$ and the posterior mean $\hat{H}_m$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.

Figure 20. Comparison of the model fit on the estimation data for the PEM estimate θ_{PEM} and the maximum a posteriori θ_{MAP} for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.

terior distribution, extending the proposed method to multiple input-multiple output LTI models, and analyzing the effect of hyperparameters on the prior distribution.

References

- Andrieu, C., De Freitas, N., & Doucet, A. (1999, June). Sequential MCMC for Bayesian model selection. In *Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics* (pp. 130–134). Caesarea, Israel.
- Åström, K., & Wittenmark, B. (1997). *Computer controlled systems. Theory and design, 3rd ed.* Prentice-Hall.
- Cappé, O., Robert, C. P., & Rydén, T. (2003). Reversible jump, birth-and-death and more general continuous

Figure 21. Comparison of the model fit on the estimation data for the subspace estimate θ_{SS} and the maximum a posteriori θ_{MAP} for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.

Figure 22. Comparison of the model fit on the estimation data for the PEM estimate θ_{PEM} and the posterior mean θ_m for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.

- time Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society. Series B*, 65(3), 679–700.
- Carter, C. K., & Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81(3), 541–553.
- Chen, T., Ohlsson, H., & Ljung, L. (2012). On the estimation of transfer functions, regularizations and Gaussian processes—Revisited. *Automatica*, 48(8), 1525–1535.
- Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging* (Vol. 330). Cambridge University Press.
- Dahlin, J., Lindsten, F., Schön, T. B., & Wills, A. (2012, July). Hierarchical Bayesian ARX models for robust inference. In *Proceedings of the 16th IFAC Symposium on System Identification*. Brussels, Belgium.
- Dahlin, J., & Schön, T. B. (2015). Getting started with particle Metropolis-Hastings for inference in nonlinear dynamical models. *arXiv preprint arXiv:1511.01707*.
- Fox, E., Sudderth, E. B., Jordan, M. I., & Willsky, A. S. (2011). Bayesian nonparametric inference of

Figure 23. Comparison of the model fit on the estimation data for the subspace estimate θ_{SS} and the posterior mean θ_m for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.

Figure 24. Comparison of the model order of the reduced models for PEM $\hat{G}_{PEM}^r$ and the maximum a posteriori $\hat{G}_{MAP}^r$ over 102 data sets. The red crosses and blue circles represent the model order for the different models, while the black line presents the curve where the model order for both models are equal.

- switching dynamic linear models. *IEEE Transactions on Signal Processing*, 59(4), 1569–1585.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6), 721–741.
- Green, M., & Limebeer, D. (2012). *Linear robust control*. Dover.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4), 711–732.
- Hastie, D. I., & Green, P. J. (2012). Model choice using reversible jump Markov chain Monte Carlo. *Statistica Neerlandica*, 66(3), 309–338.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97–109.
- Hjalmarsson, H., & Gustafsson, F. (1995). Composite modeling of transfer functions. *IEEE Transactions on Automatic Control*, 40(5), 820–832.

Figure 25. Comparison of the model order of the reduced models for subspace $\hat{G}_{SS}^r$ and the maximum a posteriori $\hat{G}_{MAP}^r$ for 102 data sets. The red crosses and blue circles represent the model order for the different models, while the black line presents the curve where the model order for both models are equal.

Figure 26. Comparison of the model order of the reduced models for PEM $\hat{G}_{PEM}^r$ and the posterior mean $\hat{G}_m^r$ for 102 data sets. The red crosses and blue circles represent the model order for the different models, while the black line presents the curve where the model order for both models are equal.

Katayama, T. (2006). *Subspace methods for system identification*. Springer.

Lindsten, F., Jordan, M. I., & Schön, T. B. (2014). Particle Gibbs with ancestor sampling. *Journal of Machine Learning Research*, 15(1), 2145–2184.

Ljung, L. (1999). *System Identification. Theory for the User*, 2nd ed. Prentice-Hall.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6), 1087–1092.

Ninness, B., & Henriksen, S. (2010). Bayesian system identification via Markov chain Monte Carlo techniques. *Automatica*, 46(1), 40–51.

Peterka, V. (1981). Bayesian system identification. *Automatica*, 17(1), 41–53.

Prando, G., Chiuso, A., & Pillonetto, G. (2014, December). Bayesian and regularization approaches to multivariable linear system identification: the role of rank penalties. In *Proceedings of the 53rd Conference on Decision and Control (CDC)* (pp. 1482–1487). Los Angeles, USA.

Figure 27. Comparison of the model order of the reduced models for subspace $\hat{G}_{SS}^r$ and the posterior mean $\hat{G}_m^r$ for 102 data sets. The red crosses and blue circles represent the model order for the different models, while the black line presents the curve where the model order for both models are equal.

Priestley, M. B. (1981). *Spectral analysis and time series*. Academic Press.

Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer.

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods*. Springer.

Schön, T. B., & Lindsten, F. (2016). *Learning of dynamical systems - Particle filters and Markov chain methods*. (forthcoming)

Schön, T. B., Wills, A., & Ninness, B. (2011). System identification of nonlinear state-space models. *Automatica*, 47(1), 39–49.

Söderström, T., & Stoica, P. (1981). Comparison of some instrumental variable methods—consistency and accuracy aspects. *Automatica*, 17(1), 101–115.

Söderström, T., & Stoica, P. (1983). *Instrumental variable methods for system identification* (Vol. 57). Springer.

Söderström, T., & Stoica, P. (1988). *System identification*. Prentice-Hall.

Viberg, M. (1995). Subspace-based methods for the identification of linear time-invariant systems. *Automatica*, 31(12), 1835–1851.

Wills, A., Schön, T. B., Lindsten, F., & Ninness, B. (2012, July). Estimation of linear systems using a Gibbs sampler. In *Proceedings of the 16th IFAC Symposium on System Identification*. Brussels, Belgium.

List of figures

- Figure 1. Bode diagram of the true system G_o (black), the maximum a posteriori estimate $\hat{G}_{MAP}$ (red, continuous line), and the reduced model $\hat{G}_{MAP}^r$ (red, dashed line), obtained by balanced truncation of $\hat{G}_{MAP}$. The shaded area corresponds to the 98% confidence interval for the posterior distribution, centered at its mean value.
- Figure 2. Bode diagram of the true system G_o (black), the mean value estimate $\hat{G}_m$ (red, continuous line), and the reduced model $\hat{G}_m^r$ (red, dashed line), obtained by balanced truncation of $\hat{G}_m$. The shaded area corresponds to the 98% confidence interval.
- Figure 3. Magnitude of the Bode diagram for the model error. Top: Maximum a posteriori estimate $\hat{G}_{MAP}$ and its reduced model $\hat{G}_{MAP}^r$. Bottom: Posterior mean estimate $\hat{G}_m$ and its reduced model $\hat{G}_m^r$.

- Figure 4. Bode diagram of the noise model derived by the maximum a posteriori estimate $\hat{H}_{\text{MAP}}$ and the posterior mean estimate $\hat{H}_m$. The true noise filter is $H_o = 1$.
- Figure 5. Bode diagram of the true system G_o (yellow, triangle line), the PEM estimate $\hat{G}_{\text{PEM}}$ (red, continuous line), and the reduced PEM model $\hat{G}_{\text{PEM}}^r$ (blue, dashed line), obtained by balanced truncation of $\hat{G}_{\text{PEM}}$.
- Figure 6. Bode diagram of the true system G_o (yellow, triangle line), the subspace estimate $\hat{G}_{\text{SS}}$ (red, continuous line), and the reduced subspace model $\hat{G}_{\text{SS}}^r$ (blue, dashed line), obtained by balanced truncation of $\hat{G}_{\text{SS}}$.
- Figure 7. Magnitude of the Bode diagram for the model error. Top: PEM estimate $\hat{G}_{\text{PEM}}$ and its reduced model $\hat{G}_{\text{PEM}}^r$. Bottom: Subspace estimate $\hat{G}_{\text{SS}}$ and its reduced model $\hat{G}_{\text{SS}}^r$.
- Figure 8. Hankel singular values for the estimated models and the true system G_o (in logarithmic scale).
- Figure 9. Empirical pdf of the phase margin.
- Figure 10. Evolution of the logarithm of the unnormalized posterior distribution $p(y_{1:N}|u_{1:N}, \theta^{(i)}) p_{\Theta}(\theta^{(i)})$ as function of the iteration i , Subsection 6.6.
- Figure 11. Hankel singular values for the estimated models for different priors (in logarithmic scale).
- Figure 12. Comparison of the model fit for the PEM estimate $\hat{G}_{\text{PEM}}$ and the maximum a posteriori $\hat{G}_{\text{MAP}}$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.
- Figure 13. Comparison of the model fit for the subspace estimate $\hat{G}_{\text{SS}}$ and the maximum a posteriori $\hat{G}_{\text{MAP}}$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.
- Figure 14. Comparison of the model fit for the PEM estimate $\hat{G}_{\text{PEM}}$ and the posterior mean $\hat{G}_m$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.
- Figure 15. Comparison of the model fit for the subspace estimate $\hat{G}_{\text{SS}}$ and the posterior mean $\hat{G}_m$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.
- Figure 16. Comparison of the model fit for the PEM estimate $\hat{H}_{\text{PEM}}$ and the maximum a posteriori $\hat{H}_{\text{MAP}}$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.
- Figure 17. Comparison of the model fit for the subspace estimate $\hat{H}_{\text{SS}}$ and the maximum a posteriori $\hat{H}_{\text{MAP}}$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.
- Figure 18. Comparison of the model fit for the PEM estimate $\hat{H}_{\text{PEM}}$ and the posterior mean $\hat{H}_m$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.
- Figure 19. Comparison of the model fit for the subspace estimate $\hat{H}_{\text{SS}}$ and the posterior mean $\hat{H}_m$ for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.
- Figure 20. Comparison of the model fit on the estimation data for the PEM estimate θ_{PEM} and the maximum a posteriori θ_{MAP} for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.
- Figure 21. Comparison of the model fit on the estimation data for the subspace estimate θ_{SS} and the maximum a posteriori θ_{MAP} for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.
- Figure 22. Comparison of the model fit on the estimation data for the PEM estimate θ_{PEM}

and the posterior mean θ_m for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.

- Figure 23. Comparison of the model fit on the estimation data for the subspace estimate θ_{SS} and the posterior mean θ_m for 102 data sets. The red crosses and blue circles represent the fit for the data sets, while the black line presents the curve where the fits of both methods are equal.
- Figure 24. Comparison of the model order of the reduced models for PEM $\hat{G}_{PEM}^r$ and the maximum a posteriori $\hat{G}_{MAP}^r$ over 102 data sets. The red crosses and blue circles represent the model order for the different models, while the black line presents the curve where the model order for both models are equal.
- Figure 25. Comparison of the model order of the reduced models for subspace $\hat{G}_{SS}^r$ and the maximum a posteriori $\hat{G}_{MAP}^r$ for 102 data sets. The red crosses and blue circles represent the model order for the different models, while the black line presents the curve where the model order for both models are equal.
- Figure 26. Comparison of the model order of the reduced models for PEM $\hat{G}_{PEM}^r$ and the posterior mean $\hat{G}_m^r$ for 102 data sets. The red crosses and blue circles represent the model order for the different models, while the black line presents the curve where the model order for both models are equal.
- Figure 27. Comparison of the model order of the reduced models for subspace $\hat{G}_{SS}^r$ and the posterior mean $\hat{G}_m^r$ for 102 data sets. The red crosses and blue circles represent the model order for the different models, while the black line presents the curve where the model order for both models are equal.