Bayesian semiparametric Wiener system identification [⋆]

Fredrik Lindsten [a], Thomas B. Schön [a], Michael I. Jordan [b]

[a] *Division of Automatic Control, Linköping University, Linköping, Sweden.*

[b] *Departments of EECS and Statistics, University of California, Berkeley, USA.*

Abstract

We present a novel method for Wiener system identification. The method relies on a semiparametric, i.e. a mixed parametric/nonparametric, model of a Wiener system. We use a state-space model for the linear dynamical system and a nonparametric Gaussian process model for the static nonlinearity. We avoid making strong assumptions, such as monotonicity, on the nonlinear mapping. Stochastic disturbances, entering both as measurement noise and as process noise, are handled in a systematic manner. The nonparametric nature of the Gaussian process allows us to handle a wide range of nonlinearities without making problem-specific parameterizations. We also consider sparsity-promoting priors, based on generalized hyperbolic distributions, to automatically infer the order of the underlying dynamical system. We derive an inference algorithm based on an efficient particle Markov chain Monte Carlo method, referred to as particle Gibbs with ancestor sampling. The method is profiled on two challenging identification problems with good results. Blind Wiener system identification is handled as a special case.

Key words: System identification, Wiener, block-oriented models, Gaussian process, semiparametric, particle filter, ancestor sampling, particle Markov chain Monte Carlo

1 Introduction

Block-oriented systems are a useful and general class of nonlinear dynamical systems. These systems consist of interconnected linear dynamics and static nonlinearities. The most well-known members of this family are the Hammerstein (static nonlinearity followed by a linear dynamical system) and the Wiener (linear dynamical system followed by a static nonlinearity) systems, introduced by [22] and [54], respectively. In this work, we are concerned with identification of the latter class. Based on observed inputs $u_{1:T} \triangleq \{u_t\}_{t=1}^{T}$ and outputs $y_{1:T}$, we wish to infer the linear dynamical system $\mathcal{G}$ and the static nonlinearity h of the Wiener system depicted in Figure 1.

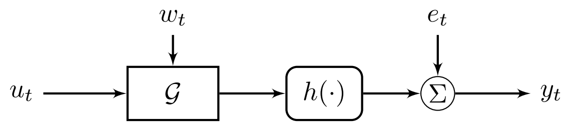

Fig. 1. A Wiener system, consisting of a linear dynamical system $\mathcal{G}$ followed by a static nonlinearity $h(\cdot)$. The system noise w_t and the measurement noise e_t are both unobserved.

Wiener systems have attracted significant attention in the system identification community, as is evident from the vast literature on the topic. See e.g. [57,17,35,19,39,18,24,51] and the references therein. However, the approach presented here differs from the existing literature on several accounts.

We consider a semiparametric (i.e., a mixed parametric/nonparametric) model of a Wiener system, in which a parametric state-space model is used for the linear block $\mathcal{G}$ and a nonparametric model is used for the nonlinear block $h(\cdot)$. Let $\theta = \{\mathcal{G}, h(\cdot)\}$ denote the unknowns of the system, i.e. θ contains both the parameters of $\mathcal{G}$ and the nonparametric representation of $h(\cdot)$ (the precise definition of θ will be made clear in Section 3). We take a Bayesian approach, modelling the parameters as random variables and the nonparametric function $h(\cdot)$ as a stochastic process. In particular, we use a Gaussian process (GP) model for $h(\cdot)$. We then provide a method for computing $p(\theta \mid y_{1:T})$, the posterior probability density

[⋆] This work was supported by: the project Calibrating Nonlinear Dynamical Models (Contract number: 621-2010-5876) funded by the Swedish Research Council and CADICS, a Linnaeus Center also funded by the Swedish Research Council. A preliminary version of this paper was presented at the 16th IFAC Symposium on System Identification, Brussels, 2012.

Email addresses: `lindsten@isy.liu.se` (Fredrik Lindsten), `schon@isy.liu.se` (Thomas B. Schön), `jordan@eecs.berkeley.edu` (Michael I. Jordan).

function (PDF) of θ given the measurements $y_{1:T}$ (and, implicitly, the inputs $u_{1:T}$). To the best of our knowledge, this is the first time the posterior PDF $p(\theta \mid y_{1:T})$ is computed for the Wiener identification problem.

In this probabilistic framework, we can handle stochastic disturbances in a systematic manner. Most notably, we are able to deal with process noise entering internally to the linear dynamical system, which can be critical in obtaining an accurate model [57]. The inclusion of such process noise in the model significantly complicates the estimation problem, and is therefore often neglected in the existing literature. Furthermore, due to the nonparametric nature of the GP, the proposed method is flexible. It can be used for a wide range of nonlinear mappings, without making any problem-specific parameterizations. We do not impose strong assumptions such as invertibility or monotonicity of the nonlinearity.

The posterior PDF $p(\theta \mid y_{1:T})$ does not allow for a closed-form solution. To cope with this, we make use of a Markov Chain Monte Carlo (MCMC) method (see e.g. [41] for a general introduction) to compute an approximation of $p(\theta \mid y_{1:T})$. More specifically, we employ the recently proposed particle MCMC (PMCMC) framework [2]. The basic idea underlying PMCMC is to use a particle filter (PF) as a component of an MCMC sampler. This is done in a manner such that, for any fixed (and finite) number of particles, no systematic error is introduced. Here, we use a state-of-the-art PMCMC method denoted particle Gibbs with ancestor sampling (PG-AS), which has been found to be efficient even when using few particles in the underlying PF [26].

Finally, we note that the proposed method can also be applied in the absence of any (measured) exogenous input u_t. This problem, refereed to as blind Wiener system identification, has received considerable attention on its own [56,49,4,1], and it can be treated as a special case of the proposed method. We have published a preliminary version of the current work (specifically targeting the blind identification problem) in [28].

2 A Bayesian semiparametric model

We consider a semiparametric model of a Wiener system. The linear dynamical system is modeled using a (parametric) state-space representation, and a nonparametric GP model is used for the static nonlinearity. The model can be described in state-space form as

$$x_{t+1} = Ax_t + Bu_t + w_t, \qquad w_t \sim \mathcal{N}(0, Q), \tag{1a}$$
$$z_t = Cx_t, \tag{1b}$$
$$y_t = h(z_t) + e_t, \qquad e_t \sim \mathcal{N}(0, r). \tag{1c}$$

Here, $x_t \in \mathbb{R}^{n_x}$ is the state of the dynamical system, $w_t \in \mathbb{R}^{n_x}$ is the process noise, $u_t \in \mathbb{R}^{n_u}$ is the input signal, $z_t \in \mathbb{R}$ is the output from the linear block and $y_t \in \mathbb{R}$ is the output from the static nonlinearity $h(z_t)$ with measurement noise e_t added. For clarity, we will write $\mathsf{X} \triangleq \mathbb{R}^{n_x}$ and $\mathsf{Y} \triangleq \mathbb{R}$ for the state-space and measurement space, respectively. For simplicity, we have restricted our attention to multiple-input single-output systems, since we then only have to consider one-dimensional GPs. However, the proposed method can be extended to multiple outputs via the use of a multidimensional GP.

The linear system is assumed to be observable. Hence, we can, without loss of generality, fix the matrix C according to $C = (1 \ 0 \ \cdots \ 0)$. Let $\Gamma = [A \ B]$. Then, the unknown quantities of the model are the system parameters Γ, Q and r as well as the nonlinear mapping $h(\cdot)$. We take a Bayesian approach and model the parameters as random variables. In the two subsequent sections we describe two different models that will be employed for the linear dynamics and in Section 2.3 the GP model for the nonlinearity is introduced.

2.1 Alt. I – Conjugate priors

If the order of the dynamical system n_x is assumed to be known, we can place conjugate priors on the matrices describing the linear dynamics. Conjugate priors are commonly used in Bayesian statistics, since they result in closed-form expressions for the posterior distributions. A conjugate prior for the linear Gaussian model (1a) is the matrix normal, inverse Wishart (MNIW) distribution. Hence, we place an MNIW prior on the pair $\{\Gamma, Q\}$, $p(\Gamma, Q) = p(\Gamma \mid Q)p(Q)$ where,

$$p(\Gamma \mid Q) = \mathcal{MN}(\Gamma; M, Q, L), \tag{2a}$$
$$p(Q) = \mathcal{IW}(Q; n_0, S_0). \tag{2b}$$

Here $\mathcal{MN}(\Gamma; M, V, L)$ is a matrix normal density [1] with mean matrix M and left and right covariances L^{-1} and V, respectively; $\mathcal{IW}(\Sigma; n, S)$ is an inverse Wishart (IW) density with n degrees of freedom and scale matrix S. As pointed out above, the MNIW prior is a standard choice for a linear Gaussian model as in (1a) (see e.g. [50]). Furthermore, for suitably chosen hyperparameters (i.e. M, L, n_0 and S_0), the effects of this prior on the posterior density will be minor. For a discussion on how to choose the hyperparameters, see Appendix A. Similarly, we put a conjugate IW prior on r (the univariate IW distribution is also known as inverse Gamma), according to,

$$p(r) = \mathcal{IW}(r; m_0, R_0). \tag{3}$$

2.2 Alt. II – Sparsity-promoting prior

It is also possible to do automatic order selection via the use of an over-parameterized model, which is then

[1] If $\Gamma \sim \mathcal{MN}(M, V, L)$, then $\text{vec}(\Gamma) \sim \mathcal{N}(\text{vec}(M), L^{-1} \otimes V)$.

regulated by some sparsity-promoting mechanism. For optimization-based methods, it is common to use ℓ_1-regularization to control sparsity. This gives rise to well known methodologies such as the lasso [44] and compressed sensing [11]. In the Bayesian setting, the analogue of sparsity regularization is to use sparsity-promoting priors. As an example, lasso can be interpreted as the Bayes posterior mode under the Laplace prior [34].

A general class of sparsity-promoting priors are the generalized hyperbolic (GH) distributions [5]. This class contains many distributions that have been successfully used to control sparsity in different settings; examples include the Laplace [34], normal inverse-Gaussian [8] and Student's t [46]. Automatic relevance determination (ARD) [31,33], which is a popular Bayesian approach to automatic order selection, is also based on a special type of GH prior. The class of GH distributions has recently been extended to a dynamic setting [7].

We use a hierarchical representation of the GH distribution. If $x \sim \mathcal{N}(0, \tau)$ with $\tau \sim \mathcal{GIG}(\nu, a, b)$, then it holds that x is distributed according to the (zero-mean) GH distribution with parameters ν, a and b. Here, $\mathcal{GIG}$ is the generalized inverse-Gaussian (GIG) distribution with density,

$$\frac{(a/b)^{\nu/2}}{2K_\nu(\sqrt{ab})}\tau^{\nu-1}\exp\left(-\frac{1}{2}\left(a\tau + b\tau^{-1}\right)\right), \tag{4}$$

where K_ν is a modified Bessel function of the second kind. The distribution is defined for $a \geq 0$, $b \geq 0$ and $\nu \in \mathbb{R}$. For $a = 0$ or $b = 0$, the normalization constant must be interpreted in a limiting sense.

To make use of sparsity-promoting GH priors for automatic order selection in state-space models, we use a multivariate generalization of the GH distribution. The prior is defined by placing independent, zero-mean Gaussian priors on the columns $\{\gamma_j\}_{j=1}^{n_x+n_u}$ of the matrix Γ,

$$p(\Gamma \mid \bar{\tau}) = \prod_{j=1}^{n_x+n_u} \mathcal{N}(\gamma_j; 0, \tau_j I_{n_x}), \tag{5}$$

where I_d is a $d \times d$ identity matrix and $\bar{\tau} = \{\tau_j\}_{j=1}^{n_x+n_u}$ are hyperparameters governing the variances of each of the columns. These are assigned independent GIG priors,

$$p(\tau_j) = \mathcal{GIG}(\tau_j; \nu, a, b), \tag{6}$$

for $j = 1, \ldots, n_x + n_u$. A similar construction has previously been used by [16], for the special case of ARD, to automatically identify the order of a state-space model.

The resulting marginal distributions of the columns of Γ will have distinct peaks at the origin. Hence, if there is not enough evidence for the jth state/input component to be non-zero, the corresponding variance parameter τ_j will decrease toward zero. This will in turn drive the jth column of Γ to zero. Contrary to the MNIW prior, which in general will lead to a full Γ-matrix, the GH prior will thus result in a Γ-matrix with a sparse column-pattern. As a result, the corresponding state components will be unobservable and they can be discarded from the model. We emphasize that these unobservable modes are inherent to the model and they should not be thought of as representing unobservable modes of the true system (which is assumed to be observable).

In summary, if a suitable model order is not known beforehand, the GH prior can thus be used for automatic order determination (as well as input selection). This is done by over-parameterizing the model and letting the GH prior switch irrelevant model components off.

The process noise and measurement noise variances are given the same IW priors as in the MNIW case, i.e. $p(Q)$ is defined according to (2b) and $p(r)$ according to (3).

2.3 Gaussian process prior

For the nonlinear mapping we develop a nonparametric model by placing a GP prior on h,

$$h(\cdot) \sim \mathcal{GP}(m(z), k_\eta(z, z')). \tag{7}$$

See [40] for a thorough introduction to GPs. The GP is governed by a mean function $m(z)$ and a covariance function (also referred to as a kernel) $k_\eta(z, z')$. We use a linear mean function $m(z) = z$, i.e. the prior is that no nonlinearity is present. However, any alternative mean function can be used if desired.

The covariance function can be taken as any positive definite kernel. Standard choices in the GP regression literature are the squared exponential kernel, the Matérn class of kernels and the rational quadratic kernel. See [40, Chapter 4] for further details and additional examples. The covariance function is (typically) parameterized by some hyperparameter η, determining for instance its amplitude and length-scale. The hyperparameter is inferred from data alongside the system parameters. To complete the model we place a prior $p(\eta)$ on the hyperparameter, depending on the choice of kernel.

Note that, due to the nonparametric nature of the GP, the proposed model is flexible and can describe a wide range of nonlinear mappings. We do not assume any specific form of h. However, since we are dealing with data affected by stochastic disturbances, we will in general favor smooth regression functions to avoid over-fitting. Still, as we shall see in Section 6, the proposed method can perform well even when the true nonlinearity is non-differentiable.

3 Inference via particle Gibbs sampling

Assume that we have observed a batch of input/output data. Let $\Pi \triangleq \{\Gamma, Q, r\}$ (for the MNIW prior), or $\Pi \triangleq \{\Gamma, Q, r, \bar{\tau}\}$ (for the GH prior) denote the system parameters. The task at hand is to identify the unknown quantities of the model, i.e. the parameters Π, the hyperparameter η and the nonlinear mapping $h(\cdot)$. Let us introduce the augmented parameter $\theta \triangleq \{\Pi, \eta, h(\cdot)\} \in \Theta \triangleq \mathcal{S} \times \mathbb{F}$, where $\mathcal{S}$ is a finite-dimensional space (containing Π and η) and $\mathbb{F}$ is an appropriate function space. Note that we use the term "parameter" to refer to θ, which also includes the nonparametric part of the model, h. We then seek the posterior density of θ given the observations $y_{1:T}$. More generally, we compute the joint posterior density of the parameter and the system states $x_{1:T}$, i.e.

$$p(\theta, x_{1:T} \mid y_{1:T}) = p(x_{1:T} \mid \theta, y_{1:T}) p(\theta \mid y_{1:T}). \tag{8}$$

The density $p(\theta \mid y_{1:T})$ is obtained by straightforward marginalization of (8). Here, and throughout this paper, conditioning on the inputs $u_{1:T}$ is implicit.

The posterior density (8) is analytically intractable and we shall make use of an MCMC sampler to address the inference problem. In Section 3.1 below we outline the solution offered by a standard Gibbs sampler and point out a fundamental problem with this approach. This problem is then solved by introducing the particle Gibbs sampler in Section 3.2.

3.1 Ideal Gibbs sampling

A Gibbs sampler is an MCMC method which targets some joint density by alternately sampling from its conditionals [41]. For the problem under study, we suggest to use a multi-stage Gibbs sampler, targeting (8) by iterating the following steps:

$$\text{Draw } \Pi^\star \mid h, x_{1:T}, y_{1:T}; \tag{9a}$$
$$\text{Draw } \eta^\star \mid \Pi^\star, x_{1:T}, y_{1:T}; \tag{9b}$$
$$\text{Draw } h^\star \mid \eta^\star, \Pi^\star, x_{1:T}, y_{1:T}; \tag{9c}$$
$$\text{Draw } x^\star_{1:T} \mid \theta^\star = \{\Pi^\star, \eta^\star, h^\star\}, y_{1:T}. \tag{9d}$$

These four steps represent the basic splitting of the variables used in the Gibbs sampler. For the conjugate MNIW prior, the posterior distribution for the system parameters in (9a) is available in closed form. For the GH prior, we need to divide step (9a) into further substeps. We return to this in Section 4.2. Note that the system parameters Π are conditionally independent of the hyperparameter η. Step (9b) is partially collapsed (we do not condition on h when sampling η). When possible, collapsing is beneficial since it allows larger updates of the involved variables and it will thus improve the mixing of the chain. For this reason, it is a standard procedure, frequently used in Gibbs sampling; see e.g. [14] and [29, Sec. 6.7].

Unfortunately, step (9d) of this Gibbs sweep is still problematic. Sampling from the exact posterior is not possible since the joint smoothing density $p(x_{1:T} \mid \theta, y_{1:T})$ is not available in closed form. In other words, the state inference problem is intractable, even if we fix the parameters of the model, due to the presence of the nonlinearity. Neither is it easy to construct a good proposal kernel for a Metropolis-Hastings (MH) sampler, due to the high dimension of $x_{1:T}$ (for large T). However, it is possible to address this problem by exploiting a powerful statistical inference tool, recently introduced in [2], known as particle MCMC (PMCMC).

3.2 Particle Gibbs sampling

A thorough treatment of PMCMC is well beyond the scope of this paper and we refer the interested reader to [2,3,37,26,53,27]. However, in this section we briefly introduce the particular PMCMC method that we have employed in this work. It is a version of the particle Gibbs (PG) sampler that we refer to as *PG with ancestor sampling* (PG-AS) [26]. It is worth emphasizing that from a practitioner's point of view, it is not necessary to understand all the technical details of PMCMC to be able to use it as a component in a composite identification procedure. Whenever we are faced with the problem of sampling from an intractable joint smoothing density, such as $p(x_{1:T} \mid \theta, y_{1:T})$, PMCMC can be used as a substitute for an exact sample, without introducing any systematic error.

The basic idea underlying PMCMC is to use a particle filter (PF) to construct a Markov kernel leaving the exact joint smoothing distribution invariant. This Markov kernel can then be used as a component of an MCMC sampler, e.g. the multi-stage Gibbs sampler given by (9). We thus seek a family of Markov kernels on X^T,

$$\{M_\theta : \theta \in \Theta\}, \tag{10}$$

such that, for each θ, $M_\theta(x_{1:T} \mid x'_{1:T})$ leaves the joint smoothing density $p(x_{1:T} \mid \theta, y_{1:T})$ invariant. In PG-AS, these kernels are constructed using a procedure referred to as a conditional particle filter with ancestor sampling (CPF-AS). Other options are available, e.g. to use the original CPF by [2] or the CPF with backward simulation [52,27]. However, we focus on CPF-AS since ancestor sampling has been found to considerably improve the mixing over the basic CPF, it can be implemented in a forward only recursion and its computational cost is linear in the number of particles.

CPF-AS is a sequential Monte Carlo sampler, similar to a standard PF, but with the important difference that one particle at each time step is specified *a priori.*

Let these particles be denoted as $x'_{1:T} = \{x'_1, \ldots, x'_T\}$. The method is most easily described as an auxiliary PF; see [13,21,36] for an introduction. As in a standard auxiliary PF, the sequence of joint smoothing densities $p(x_{1:t} \mid \theta, y_{1:t})$, for $t = 1, \ldots, T$, is approximated sequentially by collections of weighted particles. Let $\{x^i_{1:t-1}, w^i_{t-1}\}_{i=1}^N$ be a collection of weighted particles approximating $p(x_{1:t-1} \mid \theta, y_{1:t-1})$ by the empirical distribution,

$$\widehat{p}(x_{1:t-1} \mid \theta, y_{1:t-1}) \triangleq \sum_{i=1}^{N} w^i_{t-1} \delta_{x^i_{1:t-1}}(x_{1:t-1}). \qquad (11)$$

Here, $\delta_z(x)$ is a point mass located at z. To propagate this sample to time t, we introduce the auxiliary variables $\{a^i_t\}_{i=1}^N$, referred to as *ancestor indices*. The variable a^i_t is the index of the ancestor particle at time $t-1$, of particle x^i_t. Hence, x^i_t is generated by first sampling the ancestor index with $P(a^i_t = j) = w^j_{t-1}$. Then, x^i_t is drawn from some proposal kernel,

$$x^i_t \sim q(x_t \mid \theta, x^{a^i_t}_{t-1}, y_t). \qquad (12)$$

The particle trajectories are then augmented according to $x^i_{1:t} = \{x^{a^i_t}_{1:t-1}, x^i_t\}$. In the auxiliary PF formulation, the resampling step is implicit and corresponds to sampling the ancestor indices.

In a standard auxiliary PF, this procedure is repeated for each $i = 1, \ldots, N$, to generate N particles at time t. In CPF-AS, however, we condition on the event that x'_t is contained in the collection $\{x^i_t\}_{i=1}^N$. To accomplish this, we sample according to (12) only for $i = 1, \ldots, N-1$. The Nth particle is then set deterministically: $x^N_t = x'_t$.

To be able to construct the Nth particle trajectory, the conditioned particle has to be associated with an ancestor at time $t-1$. This is done by sampling a value for the corresponding ancestor index a^N_t conditionally on x'_t. From Bayes' theorem we have $p(x_{t-1} \mid \theta, x'_t, y_{1:t}) \propto p(x'_t \mid \theta, x_{t-1}) p(x_{t-1} \mid \theta, y_{1:t-1})$. By plugging (11) into this expression, we arrive at the approximation,

$$\widehat{p}(x_{t-1} \mid \theta, x'_t, y_{1:t}) = \sum_{i=1}^{N} w^i_{t-1|t} \delta_{x^i_{t-1}}(x_{t-1}) \qquad (13)$$

with $w^i_{t-1|t} \propto w^i_{t-1} p(x'_t \mid \theta, x^i_{t-1})$. To sample an ancestor particle for x'_t, we draw from this empirical distribution. That is, we sample a^N_t with $P(a^N_t = j) = w^j_{t-1|t}$.

Finally, all the particles, for $i = 1, \ldots, N$, are assigned importance weights, analogously to a standard auxiliary PF. The CPF-AS is summarized in Algorithm 1. The transition and observation densities used to compute the importance weights are, for the model (1), given by,

Algorithm 1. CPF-AS, conditioned on $x'_{1:T}$

```
1. Initialize:
  (a) Draw $x^i_1 \sim q(x_1 \mid \theta, y_1)$ for $i = 1, \ldots, N-1$.
  (b) Set $x^N_1 = x'_1$.
  (c) For $i = 1, \ldots, N$, set
        $w^i_1 \propto \frac{p(y_1 \mid \theta, x^i_1) p(x^i_1)}{q(x^i_1 \mid \theta, y_1)}$,
      where the weights are normalized to sum to 1.
2. For $t = 2, \ldots, T$ do:
  (a) Draw $a^i_t$ with $P(a^i_t = j) = w^j_{t-1}$ for $i = 1, \ldots, N-1$.
  (b) Draw $x^i_t \sim q(x_t \mid \theta, x^{a^i_t}_{t-1}, y_t)$ for $i = 1, \ldots, N-1$.
  (c) Draw $a^N_t$ with $P(a^N_t = j) \propto w^j_{t-1} p(x'_t \mid \theta, x^j_{t-1})$.
  (d) Set $x^N_t = x'_t$.
  (e) For $i = 1, \ldots, N$, set
        $w^i_t \propto \frac{p(y_t \mid \theta, x^i_t) p(x^i_t \mid \theta, x^{a^i_t}_{t-1})}{q(x^i_t \mid \theta, x^{a^i_t}_{t-1}, y_t)}$,
      where the weights are normalized to sum to 1.
```

$$p(x_{t+1} \mid \theta, x_t) = \mathcal{N}(x_{t+1}; Ax_t + Bu_t, Q), \qquad (14a)$$
$$p(y_t \mid \theta, x_t) = \mathcal{N}(y_t; h(Cx_t), r). \qquad (14b)$$

The conditioning on a prespecified collection of particles implies an invariance property of the CPF-AS, which is key to its applicability in an MCMC sampler. To state this more formally, we first make a standard assumption on the support of the proposal kernels used in the PF.

(A1) For any $\theta \in \Theta$ and $t = 1, \ldots, T$, $\mathcal{P}^\theta_t \subset \mathcal{Q}^\theta_t$ where,

$$\mathcal{P}^\theta_t = \{x_{1:t} : p(x_{1:t} \mid \theta, y_{1:t}) > 0\},$$
$$\mathcal{Q}^\theta_t = \{x_{1:t} : q(x_t \mid \theta, x_{t-1}, y_t) p(x_{1:t-1} \mid \theta, y_{1:t-1}) > 0\}.$$

The key property of CPF-AS can now be stated as follows.

Proposition 1 *Assume (A1). Then, for any $\theta \in \Theta$ and any $N \geq 2$, the procedure*

(i) Run Algorithm 1 conditionally on $x'_{1:T}$;
(ii) Sample $x^\star_{1:T}$ with $P(x^\star_{1:T} = x^i_{1:T}) = w^i_T$;

defines an irreducible and aperiodic Markov kernel M^N_θ on X^T, with invariant distribution $p(x_{1:T} \mid \theta, y_{1:T})$.

PROOF. The invariance property follows by the construction of the CPF-AS in [26], and the fact that the law of $x^\star_{1:T}$ is independent of permutations of the particle indices. This allows us to always place the conditioned particles at the Nth position. Irreducibility and aperiodicity follows from [2, Theorem 5].

Consequently, if $x'_{1:T} \sim p(x_{1:T} \mid \theta, y_{1:T})$ and we sample $x^\star_{1:T}$ according to the procedure given in Proposition 1, then, for any number of particles N, it holds that $x^\star_{1:T} \sim p(x_{1:T} \mid \theta, y_{1:T})$. For $N = 1$ we get, by construction, $x^\star_{1:T} = x'_{1:T}$, i.e. the trajectories are perfectly correlated (this is why we need $N \geq 2$ to get an irreducible kernel). As $N \to \infty$, on the other hand, the conditioning will have a negligible effect on the CPF-AS and $x^\star_{1:T}$ will be effectively independent of $x'_{1:T}$. Hence, the number of particles N will affect the mixing of the Markov kernel M^N_θ. The invariance property of the kernel holds for any N, but the larger we take N, the smaller the correlation will be between $x^\star_{1:T}$ and $x'_{1:T}$. However, it has been experienced in practice that the correlation drops off very quickly as N increases [26,27], and for many models a moderate N (e.g. in the range 5–20) is enough to obtain a rapidly mixing kernel.

4 Posterior parameter distributions

We now turn our attention to steps (9a)–(9c) of the Gibbs sampler. That is, we assume that a fixed state trajectory $x_{1:T}$ is given and consider the problem of sampling from the posterior parameter distributions. Conditioned on $x_{1:T}$, the variables $\{\Gamma, Q, \bar{\tau}\}$ are independent of $\{h(\cdot), \eta, r\}$. Furthermore, $\{\Gamma, Q, \bar{\tau}\}$ are conditionally independent of $y_{1:T}$. Hence, the densities of the conditional variables appearing in (9a)–(9c) can be written as

$$\begin{aligned} p(\Pi \mid h, x_{1:T}, y_{1:T}) &= p(\Gamma, Q, \bar{\tau} \mid x_{1:T}) \\ &\quad \times p(r \mid h, x_{1:T}, y_{1:T}), & \text{(15a)} \\ p(\eta \mid \Pi, x_{1:T}, y_{1:T}) &= p(\eta \mid r, x_{1:T}, y_{1:T}), & \text{(15b)} \\ p(h \mid \eta, \Pi, x_{1:T}, y_{1:T}) &= p(h \mid \eta, r, x_{1:T}, y_{1:T}). & \text{(15c)} \end{aligned}$$

For the MNIW prior, the variable $\bar{\tau}$ is not present. The factorization of the posterior in (15a) suggests that sampling from this distribution can be done in two decoupled steps. In the subsequent sections, we derive expressions for the PDFs appearing on the right hand sides of (15).

4.1 MNIW prior – Posterior of Γ and Q

For the MNIW prior, the posterior density of $\{\Gamma, Q\}$ is available in closed form and is given as follows. Let,

$$X = \begin{bmatrix} x_2 & \dots & x_T \end{bmatrix}, \qquad W = \begin{bmatrix} w_1 & \dots & w_{T-1} \end{bmatrix},$$
$$\bar{X} = \begin{bmatrix} x_1 & \dots & x_{T-1} \\ u_1 & \dots & u_{T-1} \end{bmatrix}.$$

It follows from (1a) that $p(x_{1:T} \mid \Gamma, Q)$ can be described in terms of the relation

$$X = \Gamma \bar{X} + W. \tag{16}$$

The prior (2) is conjugate to this likelihood model and it follows [50] that the posterior parameter distribution is MNIW and given by

$$\begin{aligned} p(\Gamma, Q \mid x_{1:T}) = \mathcal{MN}(\Gamma; S_{X\bar{X}} S^{-1}_{\bar{X}\bar{X}}, Q, S_{\bar{X}\bar{X}}) \\ \times \mathcal{IW}(Q; T - 1 + n_0, S_{X|\bar{X}} + S_0), \end{aligned} \tag{17a}$$

with

$$\begin{aligned} S_{\bar{X}\bar{X}} &= \bar{X}\bar{X}^\mathsf{T} + L, & \text{(17b)} \\ S_{X\bar{X}} &= X\bar{X}^\mathsf{T} + ML, & \text{(17c)} \\ S_{XX} &= XX^\mathsf{T} + MLM^\mathsf{T}, & \text{(17d)} \\ S_{X|\bar{X}} &= S_{XX} - S_{X\bar{X}} S^{-1}_{\bar{X}\bar{X}} S^\mathsf{T}_{X\bar{X}}. & \text{(17e)} \end{aligned}$$

4.2 GH prior – Posterior of Γ, Q and $\bar{\tau}$

If we instead use the GH prior for the system matrix Γ, there is no closed-form expression for the posterior density of $\{\Gamma, Q, \bar{\tau}\}$. To get around this, we split the sampling of these variables (in step (9a)) into sub-steps according to,

$$\begin{aligned} \Gamma^\star &\sim p(\Gamma \mid Q, \bar{\tau}, x_{1:T}), & \text{(18a)} \\ Q^\star &\sim p(Q \mid \Gamma^\star, x_{1:T}), & \text{(18b)} \\ \bar{\tau}^\star &\sim p(\bar{\tau} \mid \Gamma^\star). & \text{(18c)} \end{aligned}$$

To find the posterior of Γ, we note that (1a) can be written

$$x_{t+1} = \begin{bmatrix} \bar{x}_{t,1} I_{n_x} & \dots & \bar{x}_{t,n_x+n_u} I_{n_x} \end{bmatrix} \mathrm{vec}(\Gamma) + w_t, \tag{19}$$

where $\bar{x}_t = [x_t^\mathsf{T} \; u_t^\mathsf{T}]^\mathsf{T}$ and $\mathrm{vec}(\cdot)$ is the vectorization operator, which stacks the columns of a matrix into a vector. Hence, we may write (16) as,

$$\mathrm{vec}(X) = \underbrace{(\bar{X}^\mathsf{T} \otimes I_{n_x})}_{\triangleq \Psi} \mathrm{vec}(\Gamma) + \mathrm{vec}(W), \tag{20}$$

where $\otimes$ is the Kronecker product. Together with the prior (5) this yields the posterior of Γ as,

$$p(\Gamma \mid Q, \bar{\tau}, x_{1:T}) = \mathcal{N}(\mathrm{vec}(\Gamma); \mu_\Gamma, \Sigma_\Gamma), \tag{21a}$$

with

$$\begin{aligned} \mu_\Gamma &= \Sigma_\Gamma (\bar{X} \otimes Q^{-1}) \mathrm{vec}(X), & \text{(21b)} \\ \Sigma_\Gamma &= \left(\mathrm{diag}(\bar{\tau})^{-1} \otimes I_{n_x} + (\bar{X} \otimes Q^{-1}) \Psi \right)^{-1}, & \text{(21c)} \end{aligned}$$

where $\mathrm{diag}(v)$ is a diagonal matrix with the elements of the vector v on the diagonal.

For the posterior of Q, the IW prior (2b) is conjugate to the likelihood defined by (16) (now with Γ considered fixed). Hence, the posterior is given by an IW distribution according to

$$p(Q \mid \Gamma, x_{1:T}) = \mathcal{IW}(Q; T-1+n_0, S_{\mathrm{GH}}+S_0), \quad (22)$$

with $S_{\mathrm{GH}} = (X - \Gamma\bar{X})(X - \Gamma\bar{X})^{\mathsf{T}}$ (cf. with (17a)).

Finally, for the variance parameters of the GH prior, the (independent) Gaussian likelihoods given by (5) are conjugate to the GIG priors (6). We have,

$$\begin{aligned} &p(\tau_j \mid \gamma_j) \propto p(\gamma_j \mid \tau_j)p(\tau_j) \\ &\propto \tau_j^{-\frac{n_x}{2}} \exp\left(-\frac{1}{2\tau_j}\gamma_j^{\mathsf{T}}\gamma_j\right) \tau_j^{\nu-1} \exp\left(-\frac{1}{2}\left(a\tau + b\tau_j^{-1}\right)\right). \end{aligned} \quad (23)$$

It follows that,

$$p(\bar{\tau} \mid \Gamma) = \prod_{j=1}^{n_x+n_u} \mathcal{GIG}\left(\tau_j; \nu - \frac{n_x}{2}, a, b + \gamma_j^{\mathsf{T}}\gamma_j\right), \quad (24)$$

where we recall that $\{\gamma_j\}_{j=1}^{n_x+n_u}$ are the columns of the matrix Γ.

4.3 Posterior of r

For fixed $x_{1:T}$ and $h(\cdot)$, let $\mathbf{h} = (h(Cx_1) \;\cdots\; h(Cx_T))^{\mathsf{T}}$ and $\mathbf{y} = (y_1 \;\cdots\; y_T)^{\mathsf{T}}$ be the vectors of function outputs and observations, respectively. Furthermore, let $\mathbf{e} = (e_1 \;\cdots\; e_T)^{\mathsf{T}}$. It then follows from (1c) that the likelihood $p(y_{1:T} \mid r, h, x_{1:T})$ can be described in terms of the relation $\mathbf{y} = \mathbf{h} + \mathbf{e}$. The prior $p(r \mid h, x_{1:T}) = p(r)$ given in (3) is conjugate to this likelihood model and it follows that the posterior parameter distribution is IW and given by,

$$p(r \mid h, x_{1:T}, y_{1:T}) = \mathcal{IW}(r; T + m_0, S_r + R_0), \quad (25)$$

with $S_r = (\mathbf{y} - \mathbf{h})^{\mathsf{T}}(\mathbf{y} - \mathbf{h})$.

4.4 Posterior of h(·)

The GP prior (7) is conjugate to the likelihood model given by (1c). Hence, the posterior distribution of $h(\cdot)$ given r, $x_{1:T}$ and $y_{1:T}$ is a GP. Sampling from this posterior distribution thus involves drawing a sample path from the posterior stochastic process. When it comes to implementing a Gibbs sampler containing such a GP posterior, a problem that we need to address is how to represent this sample path.

Here, we present two alternative approaches. The first, and most proper, solution is to sample from the GP whenever an evaluation of the function h is needed in the algorithm. This will be done for N query points for each time $t = 1, \ldots, T$, where N is the number of particles used in the PG-AS sampler (see Section 3.2). The second alternative is a simpler approach, namely to evaluate the GP on a fixed grid of points. This is done once for each iteration of the MCMC sampler. When evaluating the function h in the PF, we do a linear interpolation between the grid points. This approximate solution is the approach that we have employed in the numerical examples presented in Section 6.

In either approach, let $\mathbf{z}_\star = (z^{(1)} \;\ldots\; z^{(M)})^{\mathsf{T}}$ be the points for which we wish to evaluate the GP (these can either be random points generated in the PF or fixed grid points). Furthermore, let $\mathbf{h}_\star = (h(z^{(1)}) \;\ldots\; h(z^{(M)}))^{\mathsf{T}}$. It then follows (see [40, Section 2.2]) that the posterior distribution of $\mathbf{h}_\star$ is given by

$$p(\mathbf{h}_\star \mid \eta, r, x_{1:T}, y_{1:T}) = \mathcal{N}\left(\mathbf{h}_\star; \mu_\star, \Sigma_\star\right), \quad (26a)$$

where

$$\mu_\star = \mathbf{m}_\star + P_\star^{\mathsf{T}}(P + rI_T)^{-1}(\mathbf{y} - \mathbf{m}), \quad (26b)$$
$$\Sigma_\star = P_{\star\star} - P_\star^{\mathsf{T}}(P + rI_T)^{-1}P_\star. \quad (26c)$$

Here, we have introduced the notation

$$\mathbf{m}_\star = \left(m(z^{(1)}) \;\cdots\; m(z^{(M)})\right)^{\mathsf{T}}, \quad (27a)$$
$$\mathbf{m} = \left(m(z_1) \;\cdots\; m(z_T)\right)^{\mathsf{T}}, \quad (27b)$$

and the matrices P, $P_\star$ and $P_{\star\star}$ are given by,

$$[P]_{ij} = k_\eta(z_i, z_j), \qquad i, j = 1, \ldots, T, \quad (27c)$$
$$[P_\star]_{ij} = k_\eta(z_i, z^{(j)}), \qquad i = 1, \ldots, T,\; j = 1, \ldots, M, \quad (27d)$$
$$[P_{\star\star}]_{ij} = k_\eta(z^{(i)}, z^{(j)}), \quad i, j = 1, \ldots, M. \quad (27e)$$

Using the expressions above, we can generate a sample of $\mathbf{h}_\star$ from the posterior distribution (26).

It should be noted that the computational complexity of evaluating and sampling from a posterior GP is cubic in the number of query points as well as in the number of data points, i.e. of order $O(M^3 + T^3)$. Hence, the cost of sampling from the GP can be prohibitive when T is large. However, there exist several methods in the literature, dedicated to enabling GP regression for large datasets, e.g. based on low-rank approximations; see [40, Chapter 8] and the references therein. In this work we have not resorted to such techniques.

Algorithm 2. Wiener system identification using PG-AS

```
1. Initialize:
  (a) Set $A[0] = M, Q[0] = S_0, r[0] = R_0$ and $\mathbf{h}_\star[0] = \mathbf{z}_\star$.
  (b) Set $x_{1:T}[0]$ and (for GH prior) $\bar{\tau}[0]$ arbitrarily.
2. For $k \geq 1$, iterate:
  (a) Run Algorithm 3 (for MNIW prior) or Algorithm 4
      (for GH prior) to sample $\Pi[k]$.
  (b) Sample $\eta[k]$ given $r[k]$, $x_{1:T}[k-1]$ and $y_{1:T}$ using
      an MH step as described in Section 4.5.
  (c) Sample $\mathbf{h}_\star[k] \sim p(\mathbf{h}_\star \mid \eta[k], r[k], x_{1:T}[k-1], y_{1:T})$
      according to (26).
  (d) Set $\theta[k] = \{\Pi[k], \eta[k], \mathbf{h}_\star[k]\}$.
  (e) Run Algorithm 1, targeting $p(x_{1:T} \mid \theta[k], y_{1:T})$, con-
      ditioned on $x_{1:T}[k-1]$.
  (f) Sample $J$ with $P(J = i) = w_T^i$. Set $x_{1:T}[k] = x_{1:T}^J$.
```

4.5 Posterior of η

Finally, we need to sample from the posterior of the hyperparameters of the GP kernel in (9b). Due to the often intricate dependence of the covariance kernel on η, it is in general not possible to find a closed-form expression for this posterior. Instead, we apply an MH accept/reject step to sample η. That is, we sample a value from some proposal kernel $\eta' \sim \upsilon(\eta' \mid \eta)$ (in this work we use a Gaussian random walk). The proposed sample is then accepted with probability

$$1 \wedge \frac{p(y_{1:T} \mid \eta', r, x_{1:T})}{p(y_{1:T} \mid \eta, r, x_{1:T})} \frac{p(\eta')}{p(\eta)} \frac{\upsilon(\eta \mid \eta')}{\upsilon(\eta' \mid \eta)}, \tag{28}$$

otherwise the previous value is kept. Let P_η be given as in (27c), but where we now emphasize the dependence on η in the notation. We then have

$$p(y_{1:T} \mid \eta, r, x_{1:T}) = \mathcal{N}(\mathbf{y}; \mathbf{m}, P_\eta + rI_T). \tag{29}$$

To compute the acceptance probability in (28) in a numerically robust way, we make a Cholesky factorization of the covariance $P_\eta + rI_T = R_\eta^\top R_\eta$. By straightforward manipulations it then follows that

$$\frac{p(y_{1:T} \mid \eta', r, x_{1:T})}{p(y_{1:T} \mid \eta, r, x_{1:T})} = \frac{\det(R_\eta)}{\det(R_{\eta'})} \mathrm{e}^{\left(\frac{1}{2} s_\eta^\top s_\eta - \frac{1}{2} s_{\eta'}^\top s_{\eta'}\right)}, \tag{30}$$

where $s_\eta = R_\eta^{-T}(\mathbf{y} - \mathbf{m})$, which can be computed efficiently due to the triangularity of R_η.

5 Convergence analysis

The proposed identification procedure is summarized in Algorithms 2–4. In this section we study the convergence properties of the method. Let π_T be the target distribution for the MCMC sampler, i.e. the distribution with density $p(\theta, x_{1:T} \mid y_{1:T})$. First, we provide a result stating that, for a fixed data record of length T, the empirical distribution of the generated Markov chain approaches π_T as the number of iterations $k \to \infty$. Second, we study the consistency of the Bayes estimator as the number of data points tend to infinity. We assess that the Bayes estimator is almost surely (a.s.) consistent w.r.t. the prior, for any identifiable functional.

Algorithm 3. Sampling system parameters for MNIW prior

```
1. Using (17) and (25), respectiveley,
  (a) Sample $\{\Gamma[k], Q[k]\} \sim p(\Gamma, Q \mid x_{1:T}[k-1])$.
  (b) Sample $r[k] \sim p(r \mid \mathbf{h}_\star[k-1], x_{1:T}[k-1], y_{1:T})$.
2. Return $\Pi[k] = \{\Gamma[k], Q[k], r[k]\}$.
```

Algorithm 4. Sampling system parameters for GH prior

```
1. Using (21), (22), (24) and (25), respectiveley,
  (a) Sample $\Gamma[k] \sim p(\Gamma \mid Q[k-1], \bar{\tau}[k-1], x_{1:T}[k-1])$.
  (b) Sample $Q[k] \sim p(Q \mid \Gamma[k], x_{1:T}[k-1])$.
  (c) Sample $\bar{\tau}[k] \sim p(\bar{\tau} \mid \Gamma[k])$.
  (d) Sample $r[k] \sim p(r \mid \mathbf{h}_\star[k-1], x_{1:T}[k-1], y_{1:T})$.
2. Return $\Pi[k] = \{\Gamma[k], Q[k], r[k], \bar{\tau}[k]\}$.
```

5.1 Convergence of the Markov chain

Due to the invariance property of Proposition 1, the PG-AS sampler can be treated as a regular MCMC sampler, and standard convergence analysis applies (see e.g. [45,41,32]). We start by analyzing the *ideal Gibbs* sampler, defined by running Algorithm 2 but replacing Steps 2(e)–(f) by a draw from the exact joint smoothing density; $x_{1:T}[k] \sim p(x_{1:T} \mid \theta[k], y_{1:T})$. This procedure cannot be implemented in practice, but it is useful to consider it as an intermediate step in the analysis of the PG-AS sampler.

Lemma 1 *The ideal Gibbs sampler has invariant distribution π_T.*

PROOF. The ideal Gibbs sampler is a cyclic MCMC sampler according to (9). Steps (9a), (9c) and (9d) are implemented as standard Gibbs steps. Step (9b) is implemented as an MH step. This hybrid scheme does not invalidate the MCMC kernel [45]. Collapsing is done over h in step (9b). Since this is done prior to sampling h in step (9c), the Gibbs sampler is properly collapsed [14]. Consequently, each step of the sampler leaves π_T invariant. ■

We thus know that π_T is a possible equilibrium distribution of the ideal Gibbs sampler. To assess convergence, i.e. to show that a Markov chain generated by the sampler indeed will approach π_T, we need to show that it is irreducible and aperiodic. For simplicity, we make the following assumption on the proposal kernel that is used in the inherent MH step for the hyperparameter η.

(A2) $\upsilon(\eta \mid \eta') > 0$ for any (η, η') with $p(\eta) > 0$.

The assumption is trivially satisfied if the proposal kernel is taken as, for instance, a Gaussian random walk.

Lemma 2 *Assume (A2). Then, the ideal Gibbs sampler is irreducible and aperiodic.*

PROOF. For a probability density $p(x)$, let $\operatorname{supp}[p(x)] = \{x : p(x) > 0\}$ be its support. The target density can be expressed as

$$p(\theta, x_{1:T} \mid y_{1:T}) \propto p(y_{1:T} \mid \theta, x_{1:T}) p(x_{1:T} \mid \theta) p(\theta).$$

The first two factors are Gaussian and thus everywhere positive, i.e.

$$\operatorname{supp}[p(\theta, x_{1:T} \mid y_{1:T})] = \operatorname{supp}[p(\theta)] \times \mathsf{X}^T.$$

From the definition of the model in Section 2,

$$\operatorname{supp}[p(\theta)] = \operatorname{supp}[p(\Pi)] \times \operatorname{supp}[p(\eta)] \times \operatorname{supp}[p(h)].$$

That is, by construction, the support of the target density is the Cartesian product of the supports of the marginals. Hence, the target density satisfies a positivity condition [41, Definition 9.4]. Let K be the Markovian transition kernel of the ideal Gibbs sampler. From the positivity condition and assumption (A2) it follows that $K(A \mid \{\theta', x'_{1:T}\}) > 0$ for any A with $\pi_T(A) > 0$. That is, any set with positive posterior probability is accessible in one step. Irreducibility and aperiodicity of the ideal Gibbs sampler follows. ■

Due to Proposition 1, the properties of the ideal Gibbs sampler carries over directly to the PG-AS sampler. We can thus provide the following convergence result for the proposed identification algorithm.

Theorem 1 *Assume (A1) and (A2). For any $N \geq 2$, the PG-AS sampler defined by Algorithm 2 generates a sequence $\{\theta[k], x_{1:T}[k]\}$ whose distribution $\mathcal{L}^N(\{\theta[k], x_{1:T}[k]\} \in \cdot)$ satisfies*

$$\|\mathcal{L}^N(\{\theta[k], x_{1:T}[k]\} \in \cdot) - \pi_T\|_{TV} \to 0$$

as $k \to \infty$ for almost all (π_T) starting points, where $\|\cdot\|_{TV}$ is the total variation norm.

PROOF. From Proposition 1 and Lemma 1, π_T is an invariant distribution of the PG-AS sampler. Due to the fact that the family of Markov kernels defined in Proposition 1 are irreducible and aperiodic, the results of Lemma 2 carry over to the PG-AS sampler. Convergence in total variation follows from [45, Theorem 1].

5.2 Consistency of the Bayes estimator

We now turn to consistency of estimators constructed from the posterior distribution $p(\theta \mid y_{1:T})$. We adapt the general result of Doob [12] to our setting (see also [25,42]). This classical result gives almost sure consistency w.r.t. the prior. More recent developments have improved upon the classical results, avoiding the exceptional null set on which consistency may fail; see e.g. [10,6]. Other forms of posterior convergence have also been studied extensively; see e.g. [9] for a Bernstein-von Mises theorem in the semiparametric setting and [47] for contraction rates under Gaussian process priors.

Let $\mathbb{Y} = \mathsf{Y} \times \mathsf{Y} \times \dots$ be the infinite product space, $\mathfrak{Y}$ be the smallest σ-algebra generated by all open subsets of $\mathbb{Y}$ and let $\mathrm{y} = (y_1, y_2, \dots) \in \mathbb{Y}$ be the infinite sequence of observations. Let λ be the prior distribution of θ (i.e. the distribution with density $p(\theta)$) and let $\{P_\theta : \theta \in \Theta\}$ be a family of distributions governing the law of y for each θ. We use the following identifiability criterion.

Condition 1 *Let $(Z, \mathfrak{Z})$ be a measurable space. The mapping $\zeta : \Theta \mapsto Z$ is said to satisfy Condition 1 if it is measurable, $\mathcal{L}^1$-integrable and if there exists a $\mathfrak{Y}/\mathfrak{Z}$-measurable function $f : \mathbb{Y} \mapsto Z$ with $f(\mathrm{y}) = \zeta(\theta)$ a.s. (P_θ) for any $\theta \in \Theta$.*

Theorem 2 *Let ζ satisfy Condition 1, then the Bayes estimator of $\zeta(\theta)$ is strongly consistent a.s. (λ), i.e.*

$$P_\theta\left(\lim_{t\to\infty} \beta_t = \zeta(\theta)\right) = 1, \qquad a.s.\ (\lambda),$$

with $\beta_t = \int \zeta(\theta) p(\theta \mid y_{1:t}) d\theta$.

PROOF. The proof follows [12,42]. Take $\Omega = \Theta \times \mathbb{Y}$ and let $\mu = P_\theta \times \lambda$ be the kernel product measure. Let $\gamma(\theta, \mathrm{y}) = \zeta(\theta)$ and $\beta_t(\theta, \mathrm{y}) = \beta_t(\mathrm{y}) = \mathrm{E}[\gamma \mid y_{1:t}]$. By the tower property of conditional expectation,

$$\mathrm{E}[\beta_t \mid y_{1:t-1}] = \beta_{t-1}.$$

Hence, $\{\beta_t\}$ is a martingale sequence and the martingale convergence theorem (see e.g. [55, Theorem 14.2]) implies that $\beta_t \to E[\gamma \mid \mathrm{y}]$ a.s. (μ). By Condition 1, γ is equivalent to a $\mathfrak{Y}$-measurable function a.s. (μ) and it follows that $E[\gamma \mid \mathrm{y}] = \gamma$ a.s. (μ). Hence,

$$\begin{aligned} 1 &= \mu\left(\{(\theta, \mathrm{y}) : \beta_t(\theta, \mathrm{y}) \to \gamma(\theta, \mathrm{y})\}\right) \\ &= \int P_\theta\left(\{\mathrm{y} : \beta_t(\mathrm{y}) \to \zeta(\theta)\}\right) \lambda(d\theta). \end{aligned}$$ ■

Remark 1 Condition 1 is an identifiability condition. Since the Wiener system is inherently unidentifiable (i.e. different θ can give rise to the same input-output relation), we focus on a class of identifiable functionals ζ. If

ζ is such that, given an infinite amount of data, it can be uniquely determined (i.e. it is $\mathfrak{Y}$-measurable), then the Bayes estimator of ζ is strongly consistent a.s. (λ). The identity function $\zeta(\theta) = \theta$ is not contained in this class for a Wiener system, but that is not necessarily an issue. Indeed, for instance, the one-step predictor is generally identifiable, even when θ is not. Hence, if the model is to be used for making predictions, the predictor offered by the Bayes estimator is consistent a.s. (λ).

6 Numerical illustrations

In this section we apply the proposed method to identify two synthetic Wiener systems. In both cases, a bootstrap CPF-AS with $N = 15$ particles is used in the PG-AS sampler. We use a Matérn kernel for the GP, as recommended by [43],

$$k(z, z') = \alpha \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\Delta z}{\ell} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\Delta z}{\ell} \right), \quad (31)$$

with $\Delta z = |z - z'|$. In [40], it is recommended to set $\nu = 3/2$ or $\nu = 5/2$. Here, we use the latter, which gives a slightly smoother prior. The hyperparameters α and ℓ govern the amplitude and the length-scale of the kernel, respectively. We set $\eta \triangleq \{\log \alpha, \log \ell\}$ and place an improper flat prior on the hyperparameter $p(\eta) \propto 1$.

We compare the proposed algorithm with two standard methods from the literature: the semiparametric average derivative method (ADM) [19] and a fully parametric prediction-error method (PEM) [30]. ADM estimates a parametric FIR model of the linear block. We set the order p of the FIR model based on the true impulse responses, so that any coefficient above p is smaller than 0.01 times the first coefficient. The FIR model is then transformed into an LTI model of the same order as the true system, using a balanced reduction. The nonlinear block is given by a nonparametric Nadaraya-Watson estimate. For PEM, we use an output-error model for the linear block, with the same order as the true system. The nonlinearity is parameterized differently in the two examples (see below).

We evaluate the estimates using $\mathcal{H}_2$ and ℓ_2 errors. Let $\widehat{G}$ and $\widehat{h}$ be the estimates of the transfer function G and the nonlinearity h, respectively, for one of the methods. The aforementioned errors are then given by

$$\mathcal{H}_2 : \Big(\frac{1}{2\pi} \int_{-\pi}^{\pi} |G(i\omega) - \widehat{G}(i\omega)|^2 d\omega \Big)^{1/2}, \quad (32a)$$

$$\ell_2 : \Big(\frac{1}{z_+ - z_-} \int_{z_-}^{z_+} |h(z) - \widehat{h}(z)|^2 dz \Big)^{1/2}. \quad (32b)$$

In the latter expression, we integrate over a finite interval $[z_-, z_+]$ since the nonparametric estimates can only be computed over the range of the data.

6.1 6th-order system with saturation

Consider a 6th-order linear dynamical system according to (1) where the system matrices (A, B, C) conform with the transfer function

$$G(q) = \frac{c_1 q^{-1} + \cdots + c_6 q^{-6}}{1 + a_1 q^{-1} + \cdots + a_6 q^{-6}}, \quad (33)$$

with $\bar{a} = (a_1, \ldots, a_6)$, $\bar{c} = (c_1, \ldots, c_6)$ and

$$\bar{a} = (-2.67,\, 2.96,\, -2.01,\, 0.914,\, -0.181,\, -0.0102),$$
$$\bar{c} = (-0.467,\, 1.12,\, -0.925,\, 0.308,\, -0.0364,\, 0.00110).$$

The system is excited by a known input signal u_t, which is taken as a realization of a white Gaussian noise with variance 1. The process noise and measurement noise (co)variances are given by $Q = 0.5^2 I_6$ and $R = 0.1^2$, respectively. The nonlinear mapping h is given by a saturation,

$$h(z) = \begin{cases} 1 & \text{if } z \geq 0.5, \\ 2z & \text{if } -0.5 \leq z < 0.5, \\ -1 & \text{if } z < -0.5. \end{cases} \quad (34)$$

We generate $T = 1\,000$ samples from the system and apply the proposed method (Algorithm 2) for $20\,000$ MCMC iterations [2] (out of which $10\,000$ iterations are considered as burnin). The model order is fixed to the true value $n_x = 6$ and we thus use the MNIW prior. The hyperparameters are set as described in Appendix A.

We compare Algorithm 2 with ADM and PEM. For ADM, the order of the FIR model is set as described above, which gives $p = 13$. For PEM, we use a 6th order output-error model for the linear block. For the nonlinearity, we exploit the knowledge that h is a saturation and parameterize the function accordingly.

The results are given in Figures 2 and 3; the former showing the Bode diagram of the linear system and the latter the static nonlinearity. For comparison, to account for the inherent unidentifiability of the system, the estimates from all methods are rescaled so that the linear systems have the same $\mathcal{H}_2$-norms. The shaded areas illustrate the 99 % Bayesian credibility regions, computed from the posterior PDFs. In the legends of the figure we also report the $\mathcal{H}_2$ and the ℓ_2 errors, respectively, for each method.

All methods capture the main resonance peak of $\mathcal{G}$, but are less accurate at low frequencies (likely due to a lack

[2] For simplicity, we run the chain for a fixed number of iterations, chosen based on visual inspection of the trace plots. In practice, some convergence diagnostic could be used instead, e.g. the Raftery-Lewis test [38].

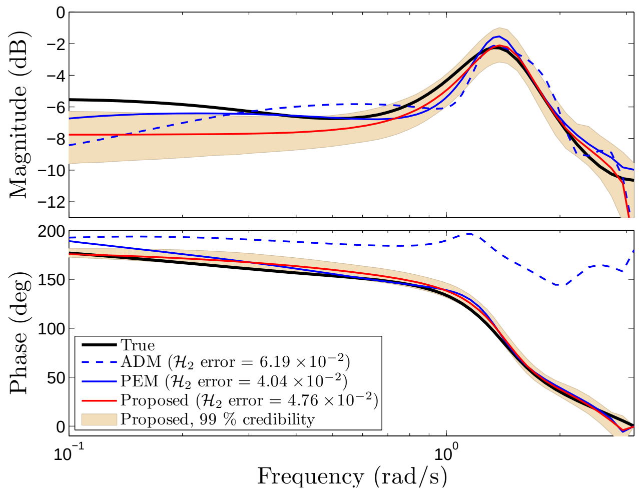

Fig. 2. Bode diagram of the 6th-order linear system and estimates for ADM, PEM and the proposed method. The red line is the posterior mean of the Bode diagram and the shaded area is the 99 % Bayesian credibility interval.

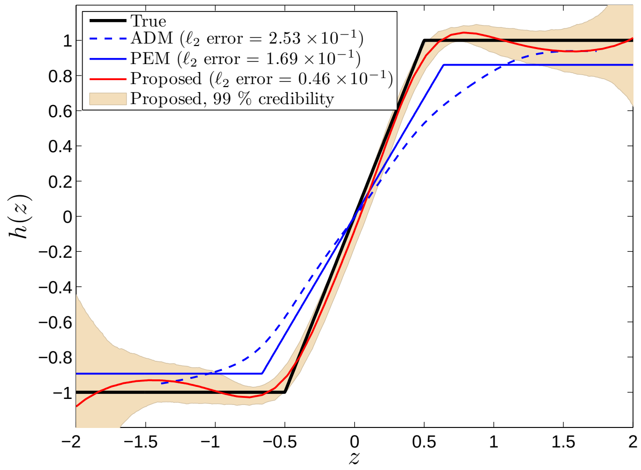

Fig. 3. Nonlinear mapping (saturation) and estimates for ADM, PEM and the proposed method. The red line is the posterior mean of the nonlinearity and the shaded area is the 99 % Bayesian credibility interval. The integration interval for the ℓ_2 error is $[z_-, z_+] = [-1.2, 1.2]$.

of excitation). ADM results in a larger error than the other two methods, especially noticeable in the phase and in the estimate of the nonlinearity. Also for PEM, there is a quite large error in the estimate of h, despite the fact that PEM uses a parametric function of the correct form. A possible reason for this is that PEM does not take the process noise into account, which results in biased estimates. The proposed method provides an accurate nonparametric estimate of the nonlinearity, despite the fact that h is non-differentiable and the GP is a smoothness prior. The uncertainty about the nonlinearity gets larger for $|z| \gtrsim 1.5$, reflecting the fact that there are few samples (≈ 2 %) in these regions available in the process underlying the observed data.

6.2 4th-order system with non-monotone nonlinearity

To highlight the flexibility of the GP model we consider a model with a non-monotonic function h, shown in Figure 5. For this example, we use a 4th-order linear dynamical system with transfer function,

$$G(q) = \frac{c_1 q^{-1} + \cdots + c_4 q^{-4}}{1 + a_1 q^{-1} + \cdots + a_4 q^{-4}}, \tag{35}$$

where $\bar{a} = (a_1, \ldots, a_4)$, $\bar{c} = (c_1, \ldots, c_4)$ and

$$\bar{a} = (1,\ 0.1,\ -0.49,\ 0.01),$$
$$\bar{c} = (0.368,\ 0.888,\ 0.524,\ 0.555).$$

The process noise and measurement noise (co)variances are given by $Q = 0.25^2 I_4$ and $R = 0.1^2$, respectively. We excite the system by a white Gaussian input signal with variance 0.5^2 and generate $T = 1\,000$ measurements. We apply the proposed identification method for 20 000 MCMC iterations (again, discarding 10 000 iterations as burnin). However, we now assume that the model order is unknown and that we wish to infer it alongside the parameters. Therefore, we employ the GH sparseness prior by over-parameterizing the model and assuming a model order of $n_x = 10$. We use the specific choice $a = 0$ in (6). For this choice, the GIG distribution reduces to an inverse-Gamma distribution, which means that the GH prior corresponds to the so called automatic relevance determination (ARD) prior [31,33].

Again, we compare the method with ADM and PEM. The order of the FIR model in ADM is set based on the true impulse response, as described above, resulting in $p = 97$. The FIR model is then reduced to an LTI system with the same order as the true system, $n_x = 4$. PEM uses a 4th order output-error model for the linear block and a piecewise affine model (with 10 segments) for the nonlinearity.

Figure 4 shows the Bode diagram of the linear system and Figure 5 shows the static nonlinearity. The non-monotonicity of h gives rise to an ambiguity of the value of z_t for a given observation y_t. Basically, for any observation y_t in the range $[-0.3, 0.3]$ there are three possible values for z_t which describe the observation equally well statically. Despite this, the proposed method accurately captures the function h, whereas both ADM and PEM fail in this respect. The linear system is also accurately estimated. Interestingly, PEM also finds a good model of $\mathcal{G}$, despite the poor estimate of h.

To analyze the effect of the GH prior and the ability to automatically determine the model order, we provide box plots of the GH precisions τ_j^{-1} for $j = 1, \ldots, n_x + n_u$ over the 10 000 MCMC iterations (taken after burnin). These are given in Figure 6. Recall from (5) that a large

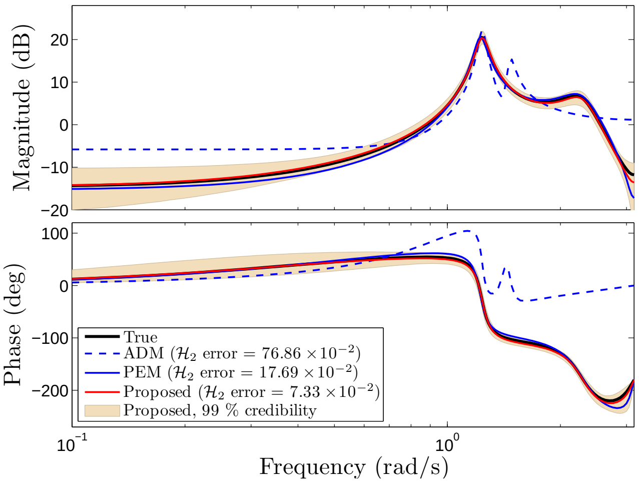

Fig. 4. Bode diagram of the 4th-order linear system and estimates for ADM, PEM and the proposed method. The red line is the posterior mean of the Bode diagram and the shaded area is the 99 % Bayesian credibility interval.

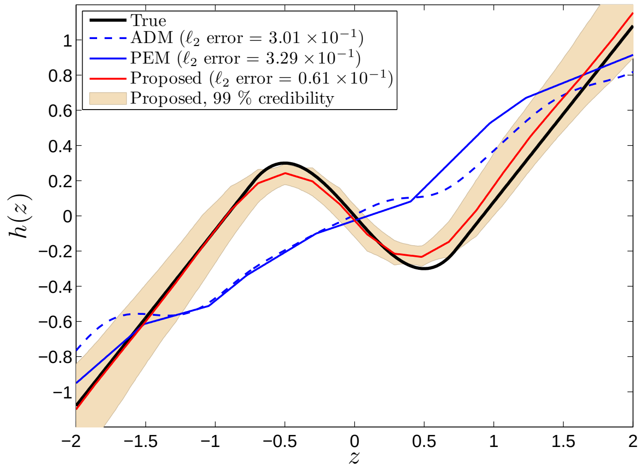

Fig. 5. Nonlinear mapping (non-monotonic) and estimates for ADM, PEM and the proposed method. The red line is the posterior mean of the nonlinearity and the shaded area is the 99 % Bayesian credibility interval. The integration interval for the ℓ_2 error is $[z_-, z_+] = [-2, 2]$.

value of τ_j^{-1} implies that the jth column of Γ is pushed to zero and that the corresponding state component in effect is switched off. It is clear that the effective model order is indeed 4, as 6 of the precision parameters take on much larger values than the remaining ones. Note that the last column of Γ, i.e. for $j = 11$, corresponds to the input signal u_t. These results indicate that sparsity-promoting priors (such as ARD) can be useful for automatic order determination of state-space models. Note, however, that there is no guarantee that the correct model order is found and further evaluation is needed in order to assess the accuracy and the robustness of this approach.

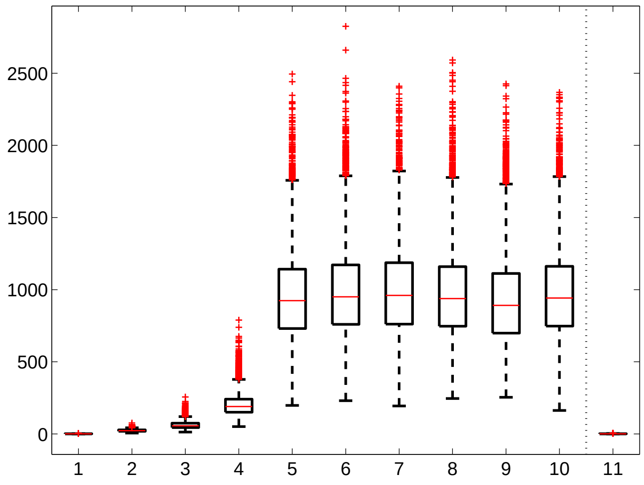

Fig. 6. ARD precision parameters τ_j^{-1} for $j = 1, \ldots, 11$. The rightmost box plot corresponds to the input signal.

6.3 Discussion

Compared to ADM and PEM, the new algorithm resulted in more accurate estimates, in particular of the nonlinearity $h(\cdot)$. We believe that ADM and PEM both suffer from the facts that: *(i)* the data are affected by process noise; *(ii)* only $T = 1\,000$ samples were used in the simulations. Our results suggest that the new method handles these difficulties better than the alternatives. It should be noted, however, that both ADM and PEM are considerably faster than the proposed method in terms of computation. For $T \gg 1\,000$, the computational complexity of the proposed method may be prohibitive. Hence, we believe that the proposed method is of particular interest when data are scarce and/or noisy. However, it is also worth exploring parallel and distributed implementations of our algorithms; note in particular that particle filtering lends itself naturally to distribution across particles.

7 Conclusions and future work

We have presented a Bayesian semiparametric method for Wiener system identification, using a state-space representation of the linear dynamical system $\mathcal{G}$ and a GP model for the static nonlinearity $h(\cdot)$. We considered two alternative priors for $\mathcal{G}$; first, a conjugate prior which is applicable when the model order is fixed (which is the case if, for instance, the order is found by cross-validation); second, a sparsity-promoting prior which can be used to automatically determine the model order. This is done by over-parameterizing the model and switching unnecessary state components off.

The new algorithm was profiled on two examples with good results. Compared to existing methods, we believe that the algorithm is of particular interest when data are scarce and/or noisy. Indeed, a concern with the proposed method is that it does not scale well with the number of measurements T, since the computational complexity of evaluating the posterior GP is cubic in T. However, this is a fairly well-studied problem in the GP literature and existing approaches can be used to mitigate this issue. Alternatively, a different type of nonparametric regression function can be used, e.g. based on the Dirichlet process mixture of generalized linear models [23].

We have found that sparsity-promoting priors can be useful for automatic order selection in state-space models. However, further evaluation is needed to determine the performance and the robustness of this approach. There are also alternative ways to do automatic order selection. For instance, reversible jump MCMC [20] can be used to infer parameters in spaces of varying dimensions. We could thus use a reversible jump sampler to include the model order n_x as a parameter of the model, and update the sizes of the system matrices accordingly when the value of this parameter is changed. This requires a way to incorporate reversible jump moves in PMCMC, which is a topic for future work.

A different line of future work is to leave the class of Wiener systems and use PMCMC for fully nonparametric identification of general nonlinear dynamical systems. This can, for instance, be done by modeling both the dynamical equation and the measurement equation of a state-space model as Gaussian processes.

A Choosing the hyperparameters

We use an approach known as empirical Bayes, in which the observed data are used to set the hyperparameters of the priors. For the MNIW prior, the following heuristic is used. First, we run a subspace identification algorithm on the input/output data (see e.g. [48]). The resulting model is transformed into observer canonical form. We set the mean M of the MN prior (2a) to the resulting $[A\ B]$-matrix. The covariance L^{-1} is set to identity. This choice allows for a considerable variability around the mean. For the IW priors (2b) and (3) we use the same heuristic as [15, p. 156–160], based on the empirical covariance of the observations $y_{1:T}$. For the ARD prior, we instead need to set the hyperparameters ν and b for the inverse-Gamma prior governing the variance vector $\bar{\tau}$. We follow [16] and set $\nu = n_x$ and $b = \nu \times 10^{-3}$. This choice fixes the prior mean to 10^{-3}, encouraging a sparse solution, and aims to provide a prior which is equally informative for different choices of n_x.

Acknowledgements

We thank the anonymous reviewers and the associate editor for their constructive comments which helped to significantly improve the quality of the manuscript.

References

[1] K. Abed-Meraim, W. Qiu, and Y. Hua. Blind system identification. *Proceedings of the IEEE*, 85(8):1310–1322, 1997.

[2] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 72(3):269–342, 2010.

[3] C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

[4] E-W Bai. A blind approach to Hammerstein-Wiener model identification. *Automatica*, 38(6):967–979, 2002.

[5] O. E. Barndorff-Nielsen and N. Shephard. Non-Gaussian OrnsteinUhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B*, 63(2):167–241, 2001.

[6] A. Barron, M. J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, 1999.

[7] F. Caron, L. Bornn, and A. Doucet. Sparsity-promoting Bayesian dynamic linear models. Technical Report Research Report 7895, INRIA, 2012.

[8] F. Caron and A. Doucet. Sparse Bayesian nonparametric regression. In *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

[9] I. Castillo. A semiparametric Bernstein-von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields*, 152:53–99, 2012.

[10] T. Choi and M. J. Schervish. On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98(10):1969–1987, 2007.

[11] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[12] J. L. Doob. Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications, Colloques Internationaux du Centre National de la Recherche Scientifique 13*, pages 23–27. CNRS, 1949.

[13] A. Doucet and A. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In D. Crisan and B. Rozovsky, editors, *The Oxford Handbook of Nonlinear Filtering*. Oxford University Press, 2011.

[14] D. A. Van Dyk and T. Park. Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103(482):790–796, 2008.

[15] E. B. Fox. *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. PhD thesis, Massachusetts Institute of Technology, 2009.

[16] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Bayesian nonparametric inference of switching dynamic linear models. *IEEE Transactions on Signal Processing*, 59(4):1569–1585, 2011.

[17] F. Giri and E-W. Bai, editors. *Block-oriented Nonlinear System Identification*, volume 404 of *Lecture notes in control and information sciences*. Springer, 2010.

[18] W. Greblicki. Nonparametric approach to Wiener system identification. *IEEE Transactions on Circuits and Systems – I: Fundamental Theory and Applications*, 44(6):538–545, 1997.

[19] W. Greblicki and M. Pawlak. *Nonparametric System Identification*. Cambridge University Press, 2008.

[20] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[21] F. Gustafsson. Particle filter theory and practice with positioning applications. *IEEE Aerospace and Electronic Systems Magazine*, 25(7):53–82, 2010.

[22] A. Hammerstein. Nichtlineare integralgleichungen nebst anwendungen. *Acta Mathematica*, 54(1):117–176, 1930.

[23] L. Hannah, D. M. Blei, and W. Powell. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12:1923–1953, 2011.

[24] A. D. Kalafatis, L. Wang, and W. R. Cluett. Identification of Wiener-type nonlinear systems in a noisy environment. *International Journal of Control*, 66(6):923–941, 1997.

[25] A. Lijoi, I. Prünster, and S. G. Walker. Extending Doobs consistency theorem to nonparametric densities. *Bernoulli*, 10(4):651–663, 2004.

[26] F. Lindsten, M. I. Jordan, and T. B. Schön. Ancestor sampling for particle Gibbs. In *Proceedings of the 2012 Conference on Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, USA, December 2012.

[27] F. Lindsten and T. B. Schön. On the use of backward simulation in the particle Gibbs sampler. In *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012.

[28] F. Lindsten, T. B. Schön, and M. I. Jordan. A semiparametric Bayesian approach to Wiener system identification. In *Proceedings of the 16th IFAC Symposium on System Identification*, Brussels, Belgium, July 2012.

[29] J. S. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer, 2001.

[30] L. Ljung. *System identification, Theory for the user.* System sciences series. Prentice Hall, Upper Saddle River, NJ, USA, second edition, 1999.

[31] D. J. C. MacKay. Bayesian non-linear modelling for the prediction competition. *ASHRAE Transactions*, 100(2):1053–1062, 1994.

[32] S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Cambridge University Press, 2nd edition, 2009.

[33] R. M. Neal. *Bayesian Learning for Neural Networks.* Springer, 1996.

[34] T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[35] G. Pillonetto and A. Chiuso. Gaussian processes for Wiener-Hammerstein system identification. In *Proceedings of the 15th IFAC Symposium on System Identification*, Saint-Malo, France, July 2009.

[36] M. K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599, 1999.

[37] M. K. Pitt, R. S. Silva, P. Giordani, and R. Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171:134–151, 2012.

[38] A. E. Raftery and S. Lewis. How many iterations in the Gibbs sampler? In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 763–773. Oxford University Press, 1992.

[39] R. Raich, G. T. Zhou, and M. Viberg. Subspace based approaches for Wiener system identification. *IEEE Transactions on Automatic Control*, 50(10):1629–1634, 2005.

[40] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* MIT Press, 2006.

[41] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer, 2004.

[42] L. Schwartz. On Bayes procedures. *Z. Wahrscheinlichkeitstheorie*, 4:10–26, 1965.

[43] M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer, 1999.

[44] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

[45] L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994.

[46] M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, 1:211–244, 2001.

[47] A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

[48] P. Van Overschee and B. De Moor. *Subspace Identification for Linear Systems: Theory, Implementation, Applications.* Kluwer Academic Publishers, 1996.

[49] L. Vanbeylen, R. Pintelon, and J. Schoukens. Blind maximum likelihood identification of Wiener systems. *IEEE Transactions on Signal Processing*, 57(8):3017–3029, 2009.

[50] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models.* Springer, New York, 1997.

[51] D. Westwick and M. Verhaegen. Identifying MIMO Wiener systems using subspace model identification methods. *Signal Processing*, 52(2):235–258, 1996.

[52] N. Whiteley. Discussion on Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B, 72(3), p 306–307, 2010.

[53] N. Whiteley, C. Andrieu, and A. Doucet. Efficient Bayesian inference for switching state-space models using discrete particle Markov chain Monte Carlo methods. Technical report, Bristol Statistics Research Report 10:04, 2010.

[54] N. Wiener. *Nonlinear Problems in Random Theory.* The MIT Press, Cambridge, MA, USA, 1966.

[55] D. Williams. *Probability with Martingales.* Cambridge University Press, 1991.

[56] A. Wills, T. B. Schön, L. Ljung, and B. Ninness. Blind identification of Wiener models. In *Proceedings of the 18th IFAC World Congress*, Milan, Italy, August 2011.

[57] A. Wills, T. B. Schön, L. Ljung, and B. Ninness. Identification of Hammerstein–Wiener models. *Automatica*, 49(1):70–81, 2013.