Nonparametric models for Hammerstein-Wiener and Wiener-Hammerstein system identification

Riccardo S. Risuleo*,** Håkan Hjalmarsson*

* ACCESS Linnaeus Centre, School of Electrical Engineering, KTH – Royal Institute of Technology, Stockholm, Sweden
(e-mail: {risuleo; hjalmars}@kth.se)

** Department of Information Technology, Uppsala University, Uppsala, Sweden
(e-mail: riccardo.risuleo@it.uu.se)

Abstract: We propose a framework for modeling structured nonlinear systems using nonparametric Gaussian processes. In particular, we introduce a two-layer stochastic model of latent interconnected Gaussian processes suitable for modeling Hammerstein-Wiener and Wiener-Hammerstein cascades. The posterior distribution of the latent processes is intractable because of the nonlinear interactions in the model; hence, we propose a Markov Chain Monte Carlo method consisting of a Gibbs sampler where each step is implemented using elliptical-slice sampling. We present the results on two example nonlinear systems showing that they can effectively be modeled and identified using the proposed nonparametric modeling approach.

Copyright © 2020 The Authors. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0)

Keywords: Nonlinear system identification, Bayesian methods, Nonparametric methods

1. INTRODUCTION

In the identification of nonlinear dynamical systems, block-oriented approaches are often used for their ability to capture complex nonlinear relationships without sacrificing mathematical tractability (Giri and Bai, 2010).

Among the block-oriented models, three-block cascades offer great model flexibility and have been used with success in many applications.

For instance, the Wiener-Hammerstein cascade consists of a static nonlinear function sandwiched between two linear dynamical systems that has been used in the modeling of skeletal muscles (Bai et al., 2009; Dewhirst et al., 2010), power electronics (Oliver et al., 2009), heat exchangers and superheaters (Haryanto and Hong, 2013), and in model-predictive control applications (Lawryczuk, 2016), among others.

The methods for Wiener-Hammerstein identification available in the literature follow four main directions: iterative nonlinear optimization methods (Marconato et al., 2012; Paduart et al., 2012; Tan et al., 2012), stochastic methods (Bershad et al., 2001; Pillonetto and Chiuso, 2009), frequency-domain methods (Westwick and Schoukens, 2012; Schoukens and Tiels, 2017), and two-stage methods (Van-Beyleen, 2014; Schoukens et al., 2014; Giordano et al., 2018).

The Hammerstein-Wiener cascade, on the contrary, is a three-block cascade consisting of a linear dynamical system sandwiched between two static nonlinear functions. These structures have been used to model radio frequency transmitters and power amplifiers (Taringou et al., 2010), the magnetosphere and ionosphere (Palanthandalam-Madapusi et al., 2005), the decomposition of stored crops (Nadimi et al., 2012), turbofan engines (Wang et al., 2017), and geothermal borefields (Atam et al., 2018). To estimate Hammerstein-Wiener models, various techniques have been proposed (see, for instance, Zhu, 2002; Hasiewicz and Mzyk, 2004; Ni et al., 2013).

Differently from the previously mentioned works, in this paper we formulate a unified nonparametric framework for modeling of three-block cascades. Modeling the impulse responses of the linear blocks and the characteristics of the nonlinear blocks using appropriate Gaussian-process models, we show that the Hammerstein-Wiener and Wiener-Hammerstein cascades can both be formulated as two-layer stochastic models such as the one presented in Fig. 1. The model consists of a set of a-priori independent latent variables observed through some nonlinear likelihood functions. We propose a sampling method to approximate the posterior distribution of the latent variables using a Markov-chain Monte Carlo algorithm. The method consists of a Gibbs sampler where each step is implemented using an elliptical-slice sampler targeting the full conditional distribution of one of the latent variables.

Fig. 1. Bayesian network of the two-layer stochastic model. Empty nodes indicate latent variables; shaded nodes indicate observed variables; edges indicate conditional dependencies.
The rest of the paper is organized as follows: in Sec. 2, we propose the general two-layer stochastic model; in Sec. 3, we propose the sampling approach targeting the posterior distribution of the latent variables of the model; in Sec. 4, we show applications to Hammerstein-Wiener and Wiener-Hammerstein cascades with general noninvertible nonlinearities.

2. TWO-LAYER STOCHASTIC MODEL

We consider a stochastic model with \( p \) independent latent variables \( z_1, \ldots, z_p \). Each latent variable \( z_i \) represents an unknown quantity in the model and is the realization of a zero-mean Gaussian process observed at some known input locations \( x_1^i, \ldots, x_{n_i}^i \) (which may represent external excitation signals, reference values, or time, among others). Hence, from the Gaussian-process model, each latent variable has a multivariate normal distribution,

\[
    z_i \sim N(0, K_{z_i}),
\]

where the covariance matrix \( K_{z_i} \) is determined by the covariance function (the kernel) of the Gaussian process,

\[
    [K_{z_i}]_{j,k} = K_i(x_j^i, x_k^i; \zeta),
\]

where \( \zeta \) is a set of prior hyperparameters that determine the shape of the kernel. In this work, we assume that the hyperparameters are given, or determined with some external procedure (e.g., cross validation or marginal likelihood maximization). Approaches to efficiently estimate these hyperparameters from data are still object of research.

We suppose that we have available \( m \) independent vectors of observations \( y_1, \ldots, y_m \), that depend on the latent variables according to some joint likelihood function

\[
    p(Y | Z; \lambda) = p(y_1 | Z; \lambda_1) \cdots p(y_m | Z; \lambda_m),
\]

where \( Y = \{y_1, \ldots, y_m\} \), \( Z = \{z_1, \ldots, z_p\} \), and where \( \lambda = \{\lambda_1, \ldots, \lambda_m\} \) are hyperparameters describing the observation models. The Bayesian network of the two-layer stochastic model is presented in Fig. 1.

In Section 4, we show how two classical structures from block-oriented system identification, the Wiener-Hammerstein and the Hammerstein-Wiener cascades, can be modeled using such two-layer stochastic models.

The general estimation problem we consider is then to estimate the posterior distribution,

\[
    p(Z | Y),
\]

of the latent Gaussian processes given the observations \( Y \).

Note that, in the following, we will drop explicit dependencies on the kernel and the likelihood hyperparameters \( \lambda \) and \( \zeta \) for notational convenience.

3. MONTE CARLO APPROXIMATE INFERENCE

In the general case, the Gaussian priors of the latent variables (1) and the likelihood function (2) are not conjugate, so the posterior distribution (3) is intractable. Therefore, we make a Monte Carlo approximation of the posterior according to

\[
    p(Z | Y) \approx \frac{1}{M} \sum_{m=1}^{M} \delta(Z_1 - z_1^{(m)}) \cdots \delta(Z_p - z_p^{(m)}),
\]

where the particles \( \{z_i^{(m)}\}_{m=1}^{M} \) are drawn using a Markov chain sampler targeting the posterior distribution (3).

To create the Markov chain, we use a Gibbs sampling procedure: we start from an initialization \( z_1^{(0)}, \ldots, z_p^{(0)} \), and we iteratively sample each latent variable conditioned on the data and all the remaining latent variables,

\[
\begin{align*}
    z_1^{(k+1)} & \sim p(z_1 | Y, z_2^{(k)}, \ldots, z_p^{(k)}), \\
    z_2^{(k+1)} & \sim p(z_2 | Y, z_1^{(k+1)}, z_3, \ldots, z_p^{(k)}), \\
    & \vdots \\
    z_p^{(k+1)} & \sim p(z_p | Y, z_1^{(k+1)}, \ldots, z_{p-1}^{(k+1)})
\end{align*}
\]

Note that, from Fig. 1, there may be conditional independence properties that reduce the number of conditioning variables in (5). In particular, \( z_j \) is conditionally independent of all the measurements \( y_j \) whose likelihood function (2) does not contain \( z_j \).

Iteratively running (5) for a large number of iterations and discarding the initial samples, we obtain a sequence of values that can be used in (4) to approximate the posterior distribution (3) and compute point estimates of interest and (Bayesian) credible intervals.

Using the Gibbs sampler, we can sample the joint posterior distribution of \( Z \) by sampling each variable \( z_i \) in turns. To sample \( z_i \), we notice that we are drawing from a conditional density that is proportional to the product of a nonlinear likelihood function and a Gaussian prior distribution,

\[
    p(z_i | Y, Z_{\setminus i}) \propto p(Y | Z_{\setminus i}) p(z_i)
\]

where \( Z_{\setminus i} = \{z_j | j \neq i\} \). Hence, to sample each step, we can use elliptical slice sampling (ESS).

As the name suggests, ESS is a modification of the standard slice sampler (Neal, 2003) specialized for drawing samples from a target distribution that is proportional to the product of a Gaussian distribution and a nonlinear likelihood function (Murray et al., 2010). To sample the latent variable, ESS starts from a sample \( z_i^{(k)} \) and draws a point \( \nu \) randomly from the prior \( p(z_i) \). Then, samples \( z_i' \) are proposed along an ellipse passing through \( z_i^{(k)} \) and \( \nu \) with an adaptively decreasing step size until the proposal has a likelihood that exceeds a threshold \( L \) determined by the starting point \( z_i^{(k)} \), in which case it is accepted as the next sample in the chain. The ESS procedure is presented in detail in Alg. 1 (adapted from Murray et al., 2010).

The whole Gibbs sampling algorithm targeting the joint posterior (3) is presented in Alg. 2.

4. APPLICATION TO THREE-BLOCK CASCADES

In this section, we present two examples of block-oriented nonlinear systems that can be modeled as two-layer stochastic models. To this end, we first introduce nonparametric Gaussian-process models on the unknowns; then, we manipulate the models to put them in the structure described by Fig. 1, in order to run Alg. 2 and approximate the posterior distribution of the latent variables.
Algorithm 1 Elliptical slice sampling updating the $i$th latent variable of a two-layer stochastic model.

1: procedure ESS($i, Z$)
2: $\nu \sim p(z_i)$, $u \sim U[0,1]$ $\triangleright$ $U$ is uniform
3: $L \leftarrow \log p(y | Z) + \log u$ $\triangleright$ Likelihood threshold
4: $\tau \sim U[0,2\pi]$ $\triangleright$ Step-size parameter
5: $\tau_{\min} \leftarrow \tau - 2\pi$, $\tau_{\max} \leftarrow \tau$
6: $Z'_i \leftarrow z_i \cos \tau + \nu \sin \tau$ $\triangleright$ Proposal on the ellipse
7: if $\log p(y | Z_i \cup \{Z'_i\}) < L$ then
8: if $\tau < 0$ then $\tau_{\min} \leftarrow \tau$ else $\tau_{\max} \leftarrow \tau$
9: $\tau \sim U[\tau_{\min}, \tau_{\max}]$
10: goto 7 $\triangleright$ Update proposal
11: end if
12: return $Z = Z_i \cup \{Z'_i\}$ $\triangleright$ Accept proposal
13: end procedure

Algorithm 2 Gibbs sampling algorithm targeting the joint posterior of the two-layer model returning $M$ samples after a burn-in of $B$ samples.

1: procedure 2LAYERGIBBS($M, B$
2: initialize $Z$ $\triangleright$ Arbitrary initialization
3: for $m = -B, \ldots, M$ do
4: for $i = 0, \ldots, p$ do $\triangleright$ For each latent variable
5: $Z \leftarrow ESS(i, Z)$ $\triangleright$ update variable
6: end for
7: if $m > 0$ then $\triangleright$ Burn-in finished
8: $Z^{(m)} \leftarrow Z$ $\triangleright$ Save samples
9: end if
10: end for
11: return $\{Z^{(m)}\}_{m=1}^M$
12: end procedure

4.1 Wiener-Hammerstein Cascades

Consider the Wiener-Hammerstein cascade in Fig. 2: it consists of a static nonlinear function $f(\cdot)$ sandwiched between two linear dynamical systems (represented by the impulse responses $g_1$ and $g_2$). The output is measured with additive Gaussian white noise.

$u \rightarrow g_1 \rightarrow w \rightarrow f(\cdot) \rightarrow g_2 \rightarrow y \rightarrow e$

Fig. 2. The Wiener-Hammerstein cascade.

We suppose that we have collected $N$ measurements of the output in a vector of samples $y$. Similarly, we have collected the values of the input in the vector $u$. Then, the $N$ samples of the internal signal $w$ can be represented by the vector

$w = G_i u$,

where $G$ is the $N \times N$ lower-triangular Toeplitz matrix of the impulse response samples that represents the convolution operated by $g_1$:

$[G_i]_{i,j} = \begin{cases} g_1[i - j + 1] & \text{if } 0 < i - j + 1 \leq N, \\ 0 & \text{otherwise}. \end{cases}$

(6)

The whole system can be represented, in vector form, by

$y = G_2 f(G_1 u) + e$,

where $G_2$, analogously to (6), is the Toeplitz matrix that represents the convolutions operated by $g_2$ and where the function $f(\cdot)$ acts on vectors elementwise. We suppose that the vector of noise samples $e$ contains Gaussian white noise with variance $\sigma^2$.

We model the impulse responses of the linear systems as independent zero-mean Gaussian processes. This means that the vectors of impulse responses are multivariate Gaussian vectors with

$g_1 \sim N(0, K_{g_1}), \quad g_2 \sim N(0, K_{g_2})$.

(7)

where the covariance matrices are determined by an appropriately chosen kernel function—for instance the first order stable spline kernel (Pillonetto et al., 2014):

$[K_{g_i}]_{i,j} = \alpha_1 \beta_{\max(i,j)}$,

(8)

where $\alpha_1 > 0$ is a scaling parameter that determines the overall amplitude of $g_1$ and $\beta_1 \in (0, 1)$ is a shaping parameter that determines the overall exponential decay of $g_1$. We use a similar model for $K_{g_2}$.

In addition, we model the static nonlinearity using a zero-mean Gaussian process

$\varphi(\cdot) \sim GP\{0, H(\cdot, \cdot)\}$,

(9)

with a covariance function $H(\cdot, \cdot)$ suitable for functional estimation—for instance, the squared exponential kernel

$H(x_i, x_j) = \eta \exp \left\{ -\frac{1}{\rho} (x_i - x_j)^2 \right\}$,

(10)

where the length-scale parameter $\rho$ determines the overall smoothness of the functions and $\eta$ is a scaling parameter. Note, however, that the method is general and in no way limited to the presented kernel functions.

From the Gaussian-process model (9), we have that the intermediate variable $f = \varphi(G_1 u)$ is a multivariate Gaussian vector with marginal distribution given by

$f \mid g_1 \sim N(0, H)$,

(11)

where the elements of the marginal covariance matrix $H$ are given by the covariance function (10) evaluated in the entries of $G_1 u$. Note that this covariance matrix is a function of the random variable $g_1$; however, we keep this dependency implicit for notational convenience.

Conditioned on the intermediate variable $f$ and on the impulse response of the second system in the cascade, the output samples have a joint multivariate Gaussian distribution given by the independent noise samples:

$y \mid f, g_2 \sim N(G_2 f, \sigma^2 I)$.

The complete nonparametric model of the Wiener-Hammerstein cascade is presented in Fig. 3 (left). It is important to note that the Gaussian process model (9) for the static nonlinearity is independent of the random variables $g_1$ and $g_2$; however, in (11), we are computing the marginal distribution of $\varphi(\cdot)$ evaluated in $G_1 u$, hence, the distribution of $f$ depends on the impulse response $g_1$ (see Fig. 3, middle).

To formulate the nonparametric Wiener-Hammerstein model as a two-layer model, we marginalize out $f$ according to

$p(y \mid g_1, g_2) = \int p(y \mid f, g_2) p(f \mid g_1) df$.

This marginalization is given by:

$y \mid g_1, g_2 \sim N(G_2 f, \sigma^2 I)$.

(12)
We suppose that we have collected a Gibbs sampling algorithm targeting the Gaussian vector with marginal distribution given by:

\[ \mathbf{g}_1 \sim \mathcal{N}(0, \mathbf{K}_{\mathbf{g}_1}), \]

\[ \mathbf{g}_2 \sim \mathcal{N}(0, \mathbf{K}_{\mathbf{g}_2}), \]

\[ \mathbf{y} \mid \mathbf{g}_1, \mathbf{g}_2 \sim \mathcal{N}(0, \mathbf{G}_2 \mathbf{H} \mathbf{G}_2^T + \sigma^2 \mathbf{I}). \]

To evaluate the proposed approach, we consider a Wiener-Hammerstein cascade with:

\[ G_1(q) = \kappa_1 \frac{1 - 0.5q^{-1} + 0.6q^{-2}}{1 - 0.3q^{-1} - 0.45q^{-2} + 0.175q^{-3}}, \]

\[ G_2(q) = \kappa_2 \frac{1}{1 - 0.8q^{-1}}, \]

where \( \kappa_1 \) and \( \kappa_2 \) are such that the blocks have unit gain. The static nonlinearity is given by the smooth and noninvertible function:

\[ f(x) = \frac{\sin(4\pi x)}{4\pi x}. \]

We generated \( N = 300 \) samples of the output of the system in response to a white-noise input, uniform in the interval \([-1,1]\), and we corrupted the output measurements with Gaussian white noise with variance equal to 10% of the variance of the noiseless output.

We considered a two-layer model, where the impulse responses are modeled with stable-spline kernels such as (8) and the static nonlinearity is modeled using a squared exponential kernel (10).

Using the procedure presented in Alg. 2, we estimated posterior distribution of \( \mathbf{g}_1 \) and \( \mathbf{g}_2 \) using \( M = 1000 \) samples, collected after burn-in of \( B = 100 \) samples. The approximate wallclock time needed to estimate one system using the method is 2 minutes.

The hyperparameters of the stable-spline kernels were set using a validation set of 150 samples (50% of initial the training set); \( \alpha_i \) were chosen among 6 values logarithmically spaced between \( 10^{-3} \) and \( 10^{2} \); \( \beta_i \) were chosen among 6 values uniformly spaced between 0.2 and 0.8. As the model is not identifiable, the hyperparameters of the static nonlinearity were fixed arbitrarily to \( \eta = 10 \) and \( \rho = 5 \). Different models were estimated using 50% of the data as training and then evaluated using the prediction error on the remaining 50% of the data. Finally, the hyperparameters that minimized prediction error were used to sample the posterior using the whole dataset.

### 4.2 Hammerstein-Wiener cascades

Consider the Hammerstein-Wiener cascade in Fig. 5; as we did for the Wiener-Hammerstein case, we represent the linear block with the vector of impulse response samples \( \mathbf{g} \). The output of the cascade is subject to an additive noise \( \mathbf{e} \).

\[ \mathbf{u} \xrightarrow{f_1(\cdot)} \mathbf{g} \xrightarrow{w} \mathbf{f}_2(\cdot) \xrightarrow{\oplus} \mathbf{y} \]

If we suppose that we have collected \( N \) measurements of the input in a vector of samples \( \mathbf{u} \) then the samples of the internal signal \( \mathbf{w} \) can be represented by the vector

\[ \mathbf{w} = \mathbf{G}_{f_1(\mathbf{u})}, \]

where \( f_1(\cdot) \) is evaluated elementwise and where \( \mathbf{G} \), analogously to (6), is the \( N \times N \) lower-triangular Toeplitz matrix that represents the convolution operated by the linear system. Similarly, the \( N \) samples of the output can be represented, in vectorized form, as

\[ \mathbf{y} = \mathbf{f}_2(\mathbf{w}) + \mathbf{e}, \]

where \( \mathbf{e} \) is a vector of samples of Gaussian white measurement noise with variance \( \sigma^2 \).

We model the impulse response of the linear system \( g \) with a zero-mean Gaussian process, obtaining the following multivariate Gaussian distribution for the vector \( \mathbf{g} \):
where the covariance matrix is determined by the covariance function of the Gaussian process—for instance, the stable-spline kernel (8).

Similarly to what we did in the Wiener-Hammerstein case, we use zero-mean Gaussian-process models for the static nonlinearities,

\[ \varphi_1(\cdot) \sim \mathcal{GP}(0, H_1(\cdot, \cdot)), \quad \varphi_2(\cdot) \sim \mathcal{GP}(0, H_2(\cdot, \cdot)), \]

for appropriate covariance functions \( H_1(\cdot, \cdot) \) and \( H_2(\cdot, \cdot) \)—for example, the squared exponential kernel (10).

Considering the intermediate random variables \( f_1 = \varphi_1(u) \) and \( f_2 = \varphi_2(Gf_1) \). We can write the marginal distributions

\[ f_1 \sim \mathcal{N}(0, H_1), \quad f_2 | f_1, g \sim \mathcal{N}(0, H_2), \] (14)

where the elements of \( H_1 \) are given the covariance function evaluated in the entries of \( u \), and the elements of \( H_2 \) are given by the covariance function evaluated in the entries of \( Gf_1 \). Note that the covariance matrix \( H_2 \) has a dependency on \( f_1 \) and \( g \) which we keep implicit for notational convenience.

Conditioned on \( f_2 \), the output has a multivariate normal distribution given by

\[ y | f_2 \sim \mathcal{N}(f_2, \sigma^2 I). \]

This complete nonparametric model of the Hammerstein-Wiener cascade is presented in Fig. 6 (left). Note again that the Gaussian process models are \textit{a priori} independent and the dependency of \( f_2 \) and \( g \) and \( f_1 \) is introduced by marginalization (see Fig. 6, middle).

To formulate the two-layer model of the Hammerstein-Wiener cascade, we can marginalize out \( f_2 \) according to

\[ p(y | f_1, g) = \int p(y | f_2) p(f_2 | f_1, g) df_2. \]

Similarly to the Wiener-Hammerstein case, also this marginalization has a closed-form solution:

\[ y | f_1, g \sim \mathcal{N}(0, H_2 + \sigma^2 I). \] (15)

Collecting (13), (14), and (15), we see that the Hammerstein-Wiener structure here discussed can be modeled using a two-layer stochastic representation (see Fig. 6, right):
the latent variables. We have validated the approach on two example nonlinear systems showing that the proposed approach can be used to effectively identify nonparametric models of three-block cascades from data.

While the two layer stochastic approach seems very powerful and has many applications, there are limitations. In particular, the elliptical-slice sampling step involved in the algorithm can become expensive in large models or when there are many measurements: as shown in Algorithm 2, we need to sample every latent variable using a rejection scheme which involves the computation of the likelihood function. If the likelihood function is expensive to compute, this may lead to a large overall computational burden—for instance, in the examples considered here, the complexity of the likelihood function is cubic in the number of data. Accelerating the sampling procedure (possibly using variational approximations) is left as future work. In addition, there is no obvious way to tune the hyperparameters. In the simulation examples, we have used a validation set to select the hyperparameters. However, this is slow (for reference, the estimation is Sec. 4.1 took about 45 hours) and can possibly lead to overfitting. Extensions of the method attempting to estimate the hyperparameters from the marginal likelihood function are currently under study.

REFERENCES


